C.4

# TECHNIQUES D'ENQUÊTE

STATISTICS CANADA STATISTIQUE CANADA
JAN 12 1994
LIBRARY
BIBLIOTHÈQUE

Catalogue 12-001

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

DÉCEMBRE 1993

•

VOLUME 19

•

NUMÉRO 2

Statistique Canada Statistics Canada

Canada

# TECHNIQUES D'ENQUÊTE

## UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

DÉCEMBRE 1993 • VOLUME 19 • NUMÉRO 2

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, des Sciences
et de la Technologie, 1993

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

Décembre 1993

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

N° 12-001 au catalogue

ISSN 0714-0045

Ottawa

10101567 19
BIBLIOTHEQUE STATISTIQUE CANADA
STATISTICS CANADA LIBRARY

Statistique Canada Statistics Canada

Canada

603 Ca

# TECHNIQUES D'ENQUÊTE

## Une revue de Statistique Canada

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

**COMITÉ DE DIRECTION**

**Président** G.J. Brackstone

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

**COMITÉ DE RÉDACTION**

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

**Rédacteurs associés**

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

**Rédacteurs adjoints**

N. Laniel, P. Lavallée, L. Mach et H. Mantel, *Statistique Canada*

---

## POLITIQUE DE RÉDACTION

La revue 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

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

### Abonnement

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

# TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 19, numéro 2, décembre 1993

## TABLE DES MATIÈRES

Dans ce numéro ..... 135

P.P. BIEMER et D. ATKINSON
Estimation de l'erreur systématique de mesure par la prédiction modéliste ..... 137

J.B. ARMSTRONG et J.E. MAYDA
Estimation modéliste des taux d'erreur liés au couplage d'enregistrements ..... 147

D. PFEFFERMANN et S.R. BLEUER
Modélisation conjointe robuste de séries de données sur l'activité pour de petites régions ..... 159

I.U.H. MIAN et N. LANIEL
Estimation d'un modèle d'étalonnage à biais multiplicatif constant par la méthode du maximum de vraisemblance et application ..... 175

J.-C. DEVILLE
Plans de sondage à deux degrés optimaux pour des estimateurs de ratios: application au contrôle de qualité du recensement français de 1990 ..... 183

R.J. CASADY et R. VALLIANT
Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale ..... 193

S. BANDYOPADHYAY et A.K. ADHIKARI
Échantillonnage dans des bases imparfaites contenant un nombre inconnu d'enregistrements répétés ..... 205

F.A. ROESCH JR., E.J.GREEN et C.T. SCOTT
Un nouveau concept pour l'échantillonnage des forêts ..... 211

G. KALTON et C.F. CITRO
Enquêtes par panel: ajout d'une quatrième dimension ..... 217

Remerciements ..... 229

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, p. 135
Statistique Canada



# Dans ce numéro

Ce numéro de *Techniques d'enquête* contient des articles portant sur différents sujets. Dans le premier article, Biemer et Atkinson présentent une méthode générale pour construire et évaluer des estimateurs obtenus à l'aide de la prédiction modéliste de l'erreur systématique de mesure pour un plan de sondage à deux phases stratifié avec échantillonnage aléatoire simple dans chaque phase. À des fins d'évaluation, ils ont étendu la méthode bootstrap de Bickel et Freedman à l'échantillonnage à deux phases. L'exemple utilisé pour illustrer la méthode proposée montre qu'il est possible d'améliorer l'estimateur classique de la différence nette et de réaliser ainsi des économies au niveau du coût des enquêtes de réinterview.

À l'origine, l'article de Armstrong et Mayda devait paraître dans la section spéciale sur le *couplage d'enregistrements et l'appariement statistique*. Les auteurs examinent l'estimation fondée sur un modèle des taux d'erreur de classification liés au couplage d'enregistrements. La classe de modèles étudiés permet la non-indépendance du statut d'appariement de différents champs d'appariement dans une paire d'enregistrements. On élabore des méthodes d'estimation et l'on en compare certaines qui portent sur l'estimation du taux d'erreur à l'aide de données synthétiques et réelles.

Pfeffermann et Bleuer étudient l'estimation pour de petites régions à l'aide de données provenant d'une enquête par panel avec renouvellement dans le temps. Leur méthode est fondée sur un modèle d'espace d'états pour les valeurs de la population dans le temps et de modèles autorégressifs distincts pour les séries d'erreurs de l'enquête provenant de chaque panel. Afin d'obtenir une certaine robustesse, on impose aux estimateurs pour de petites régions la contrainte additionnelle que leur somme doit être égale à celle des estimateurs directs de l'enquête dans de plus grandes régions définies à l'avance. La méthode est présentée à l'aide de données de l'Enquête (canadienne) sur la population active pour les provinces de l'Atlantique.

Mian et Laniel traitent de deux procédures itératives pour trouver les estimations d'un modèle d'étalonnage non linéaire par la méthode du maximum de vraisemblance qui semble approprié pour des séries chronologiques économiques provenant de grosses enquêtes par sondage. Des expressions en forme analytique fermée pour les variances et les covariances asymptotiques des séries étalonnées et des valeurs ajustées sont aussi fournies. Les méthodes sont illustrées à l'aide de données sur le commerce de détail au Canada.

Deville utilise des modèles de superpopulation pour anticiper, avant l'enquête, les variances d'estimation de ratios. À l'aide de modèles simples et réalistes, il produit des expressions plus ou moins complexes qu'il parvient à optimiser. Il traite du problème de l'estimation de la fréquence des erreurs dans l'ensemble des formules recueillies lors du contrôle de la qualité du recensement français.

Casady et Valliant utilisent des techniques asymptotiques pour étudier la stratification a posteriori à partir d'un point de vue conditionnel fondé sur un plan d'enquête. Les auteurs calculent, pour de grands échantillons, le biais et les erreurs quadratiques moyennes de l'estimateur de stratification a posteriori courant, de l'estimateur de Horvitz-Thompson, d'un estimateur par quotient et d'un nouvel estimateur de régression de stratification a posteriori. La théorie élaborée fait l'objet de tests empiriques portant sur des populations réelles et sur une population artificielle. Le problème du biais imputable aux bases de sondage défectueuses est aussi étudié.

Bandyopadhyay et Adhikari examinent l'estimation fondée sur des bases de sondage où certaines unités apparaissent plus d'une fois, chaque fois avec une identité différente. On compare les erreurs quadratiques moyennes des estimateurs fondés sur des bases parfaite et imparfaite. On étudie l'estimation d'un ratio, d'une moyenne et d'un total de la population quand aucune information supplémentaire n'est disponible sur la base de sondage.

Roesch, Green et Scott présentent un concept généralisé pour toutes les méthodes couramment utilisées d'échantillonnage des forêts. Selon ce concept, la forêt est perçue comme une image bidimensionnelle découpée en pièces comme un casse-tête, les pièces étant définies par les probabilités de sélection individuelles des arbres de la forêt.

L'article de Kalton et Citro est une version révisée du discours-programme prononcé dans le cadre du Symposium '92 de Statistique Canada portant sur les enquêtes longitudinales. L'article examine comment différents plans d'enquête visant à recueillir des données dans le temps permettent d'atteindre divers objectifs analytiques. L'auteur se concentre ensuite sur les enquêtes par panel et traite des décisions qui doivent être prises au moment de la conception de ces enquêtes.

Le rédacteur en chef

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 137–146
Statistique Canada



# Estimation de l'erreur systématique de mesure par la prédiction modéliste

PAUL P. BIEMER et DALE ATKINSON[1]

RÉSUMÉ

Les méthodes qui servent à estimer le biais de réponse dans les enquêtes requièrent des mesures répétées "non biaisées" pour à tout le moins un sous-échantillon d'observations. L'estimateur habituel du biais de réponse est la différence entre la moyenne des observations originales et la moyenne des observations non biaisées. Dans cet article, nous étudions divers estimateurs du biais de réponse tirés de la prédiction modéliste. Nous supposons comme plan de sondage un échantillonnage à deux phases stratifié, avec échantillonnage aléatoire simple dans chaque phase. Nous supposons que la caractéristique $y$ est observée pour chaque unité échantillonnée dans la phase 1, tandis que la valeur vraie de la caractéristique, $\mu$, est observée pour chaque unité du sous-échantillon prélevée dans la phase 2. Nous supposons en outre qu'une variable auxiliaire $x$ est connue pour chaque unité de l'échantillon de la phase 1 et que le chiffre de population de $x$ est connu. On suppose un certain nombre de modèles qui mettent en relation $y$, $\mu$ et $x$; de ces modèles découlent divers estimateurs de $E(y - \mu)$, le biais de réponse. Les estimateurs sont calculés à l'aide d'une méthode d'auto-amorçage destinée à l'estimation de la variance, du biais et de l'erreur quadratique moyenne. La méthode que nous utilisons est en fait la méthode de Bickel-Freedman à une phase, étendue à un plan à deux phases stratifié. À des fins d'illustration, nous appliquons la méthode étudiée à des données du programme de réinterview du National Agricultural Statistics Service. Nous montrons par ces données que l'estimateur fondé sur un modèle de Särndal, Swensson et Wretman (1991) est supérieur à l'estimateur de différence habituel, ce qui prouve qu'il est possible d'améliorer les estimateurs classiques au moyen de la prédiction modéliste.

MOTS CLÉS: Réinterview; mesures répétées; erreur de réponse; auto-amorçage (bootstrap).

## 1. INTRODUCTION

Dans les ouvrages sur les enquêtes statistiques, on reconnaît que la valeur observée des caractéristiques étudiées peut différer sensiblement de la valeur réelle de ces caractéristiques lorsque l'enquête par sondage fait intervenir des répondants. Il existe un grand nombre de cas où des erreurs de mesure sont observées dans des enquêtes. Par exemple, on peut vérifier l'exactitude d'une réponse fournie en comparant cette donnée à la valeur réelle (ou, du moins, à une valeur plus exacte) de la caractéristique étudiée, contenue dans des dossiers administratifs ou des documents juridiques. Une autre méthode de vérification consiste à réinterviewer des répondants et à produire des rapports révisés. Dans une réinterview, on communique à nouveau avec un répondant pour lui faire subir une seconde interview qui vise à mesurer les mêmes caractéristiques que dans la première interview. Au lieu de reprendre simplement les questions de l'interview initiale, on peut approfondir certains questions dans le but d'obtenir des réponses plus exactes, ou on peut demander au répondant de consulter un document dans lequel figurent des valeurs possibles pour les caractéristiques étudiées. Dans le cas de certaines enquêtes de réinterview, l'intervieweur tente de résoudre, avec l'aide du répondant, les différences de réponse entre la première et la seconde interview jusqu'à ce qu'il n'y ait plus de doute sur l'exactitude de la réponse. Forsman et Schreiner (1991) donnent un aperçu des ouvrages qui portent sur ce type de réinterviews. Parmi les autres moyens de vérifier l'exactitude des réponses d'enquête, notons: a) la comparaison des statistiques d'enquêtes (c.-à-d., moyennes, totaux, proportions, *etc.*) avec des statistiques provenant de sources externes, plus précises; b) l'utilisation de plans d'expérience dans le but d'estimer l'incidence des intervieweurs et des autres membres du personnel affectés aux enquêtes sur les estimations; et c) la vérification de la cohérence interne des résultats d'une enquête.

Les recherches actuelles portent plus spécialement sur les estimateurs de l'erreur systématique de mesure, évaluée par rapport à des données recueillies dans des études de réévaluation – par ex., études de vérification des enregistrements et réinterviews – qui ont pour objectif de faire connaître la valeur vraie de la caractéristique à un coût (par observation) probablement beaucoup plus élevé que dans l'enquête initiale.

À cause du coût normalement élevé des réinterviews, on recueille à nouveau des données uniquement pour une petite partie de l'échantillon initial. Tandis que la taille de l'échantillon peut être suffisante pour estimer l'erreur systématique au niveau national ou régional, elle peut ne

[1] Paul P. Biemer, chercheur principal, Center for Survey Research, Research Triangle Institute, Research Triangle Park (NC), É.-U. 27709; Dale Atkinson, statisticien superviseur, National Agricultural Statistics Service, 3251 Old Lee Highway, salle 305, Fairfax (VA), É.-U. 22030.

pas l'être pour l'estimation de l'erreur rattachée à de faibles sous-populations ou à des caractéristiques peu courantes. Dans cet article, nous nous proposons de considérer des estimateurs du biais dans les réponses qui sont plus efficaces que les estimateurs classiques. Le fondement de notre approche peut être décrit dans les termes suivants.

Dans une étude de réévaluation typique, on tire un sous-échantillon aléatoire des participants à l'enquête et on détermine par un moyen quelconque la valeur vraie des caractéristiques à l'étude. Désignons par $n_1$ le nombre de participants à la première enquête et par $n_2$ le nombre d'unités qui constituent le sous-échantillon ou l'échantillon d'évaluation. L'estimateur habituel du biais dans les réponses est le taux d'écart net, qui est calculé pour les $n_2$ répondants de l'échantillon d'évaluation au moyen de la formule

$$\text{TÉN} = \bar{y}_2 - \bar{\mu}_2, \tag{1.1}$$

où $\bar{y}_2$ est la moyenne d'échantillon des réponses observées et $\bar{\mu}_2$, la moyenne d'échantillon des valeurs vraies. L'inconvénient du TÉN est qu'il fait abstraction de toute information sur les $n_1 - n_2$ unités de l'échantillon initial qui ne participent pas à l'étude de réévaluation. Il ne renferme pas non plus d'information sur des variables auxiliaires, $x$, laquelle pourrait être combinée à l'information sur $y$ et $\mu$ tirée de l'enquête pour produire un estimateur du biais dans les réponses plus précis.

Comme nous avons ici un plan d'échantillonnage à deux phases stratifié et les données correspondantes $(y, \mu, x)$, notre objectif est de déterminer le "meilleur" estimateur de l'erreur systématique de mesure, étant donné ces observations. Il s'agit essentiellement de définir un modèle pour la valeur vraie, $\mu_i$, qui soit une fonction des valeurs observées $y_i$, $i = 1, \ldots, n_1$, et de toute information supplémentaire, $x$, qui peut exister sur la population. Ce modèle sert ensuite à prédire $\mu_i$ pour toutes les unités de la population pour lesquelles cette valeur est inconnue. On peut par la suite utiliser ces prévisions pour calculer des estimations de la moyenne, du total ou de la proportion de population vrais. On peut donc déterminer des estimateurs du biais dans les réponses pour ces paramètres à partir de l'enquête principale. Comme la méthode produit une équation prédictive pour $\mu_i$ qui est une fonction des observations, on peut calculer des estimateurs du biais dans les réponses pour des domaines à taille d'échantillon faible. Dans de telles circonstances, on peut ajouter des variables géographiques et des variables individuelles dans l'équation prédictive pour $\mu_i$ (par ex., caractéristiques démographiques, type d'unité, taille de l'unité, caractéristiques géographiques, et ainsi de suite).

Dans la section qui suit, nous exposons les principes d'estimation et d'évaluation qui sous-tendent la méthode prédictive appliquée à l'estimation du biais dans les réponses. Suivant un plan d'échantillonnage aléatoire stratifié, nous produisons des estimateurs de moyennes et de totaux ainsi que les variances et les erreurs quadratiques moyennes correspondantes. Nous présentons aussi les résultats de l'application de la méthode à des données du National Agricultural Statistics Service (NASS).

## 2. MÉTHODES D'ESTIMATION ET D'ÉVALUATION

### 2.1 Modèle de l'erreur de mesure

Plus précisément, nous allons étudier le cas de l'échantillonnage aléatoire simple sans remise (ÉASSR) dans une population unique. Plus loin, nous étendrons notre raisonnement au cas de l'échantillonnage aléatoire stratifié, ce qui est assez simple.

Désignons par $U = \{1, 2, \ldots, N\}$ l'ensemble des indices pour la population et désignons par $S_1 = \{1, 2, \ldots, n_1\}$ sans perte de généralité, l'ensemble des indices pour l'ÉASSR de première phase formé de $n_1$ unités de $U$.

Pour $y_i$, $i \in S_1$, supposons le modèle

$$y_i = \gamma_0 + \gamma\mu_i + \epsilon_i, \tag{2.1}$$

où $\mu_i$ est la valeur vraie de la caractéristique étudiée, $\gamma_0$ et $\gamma$ sont des constantes et $\epsilon_i$ est un terme d'erreur indépendant d'espérance nulle et de variance conditionnelle $\sigma^2_{\epsilon i}$.

Puisque notre analyse porte précisément sur le biais rattaché aux valeurs $y_i$, considérons l'espérance de $y_i$. Soit $\mathrm{E}(y_i \mid i)$ l'espérance conditionnelle de $y_i$ pour la distribution des $\epsilon_i$, l'unité $i$ étant fixe, et soit $\mathrm{E}(y_i) = \mathrm{E}_i[\mathrm{E}(y_i \mid i)]$ l'espérance de $\mathrm{E}(y_i \mid i)$ pour la distribution d'échantillonnage. Alors, pour une unité donnée $i$,

$$\mathrm{E}(y_i \mid i) = \gamma_0 + \gamma\mu_i \tag{2.2}$$

et, par conséquent, l'espérance inconditionnelle est

$$\mathrm{E}(y_i) = \gamma_0 + \gamma\bar{M}, \tag{2.3}$$

où $\bar{M} = \sum_{i=1}^{N} \mu_i/N$. L'erreur systématique de mesure s'écrit donc

$$\bar{B} = \mathrm{E}(y_i - \mu_i) = \gamma_0 + (\gamma - 1)\bar{M}. \tag{2.4}$$

Le paramètre $\gamma_0$ est un terme d'erreur systématique constant qui ne dépend pas de la grandeur de $\bar{M}$. Notons que cette définition de $\gamma_0$ est conforme à la définition de l'erreur systématique de mesure qui découle normalement du modèle élémentaire

$$y_i = \mu_i + \epsilon_i, \tag{2.5}$$

avec $\epsilon_i \sim (\gamma_0, \sigma^2_{\epsilon i})$. (Voir, par exemple, Biemer et Stokes 1991.)

Considérons l'estimation de $\bar{B}$. Supposons qu'un sous-échantillon de taille $n_2$ est prélevé parmi les $n_1$ unités de l'échantillon initial et que la valeur vraie $\mu_i$ est déterminée pour ces $n_2$ unités. On peut établir la valeur vraie soit par une réinterview, une vérification d'enregistrements, une observation faite par l'intervieweur ou un autre moyen. Désignons par $S_2 \subseteq S_1$ cet échantillon de seconde phase. L'estimateur habituel de l'erreur systématique de mesure est le TÉN défini en (1.1). Si l'hypothèse selon laquelle "la valeur vraie, $\mu_i$, est observée dans la phase 2 pour tous $i \in S_2$" est

satisfaite, alors le TÉN est un estimateur non biaisé de $\bar{B}$. On peut montrer en outre que la variance du TÉN est

$$\mathrm{E}\left\{\left(1-\frac{n_2}{n_1}\right)\frac{s_\mu^2}{n_2}\left(1-\frac{s_{\mu y}^2}{s_y^2 s_\mu^2}\right)+\left(1-\frac{n_2}{n_1}\right)\frac{s_y^2}{n_2}(1-r)^2\right\}, \quad (2.6)$$

où $s_\mu^2 = \sum_{j \in S_2} (\mu_j - \bar{\mu}_2)^2/(n_2 - 1)$, avec des définitions analogues pour $s_y^2$ et $s_{\mu y}$, et $r = s_{\mu y}/s_y^2$.

Le TÉN peut être sous-optimal dans un certain nombre de situations qui se reproduisent périodiquement. Pour nous en convaincre, considérons des estimateurs de $\bar{B}$ de la forme suivante

$$\bar{b}_{ga} = \bar{y}_g - \bar{\mu}_{Ra}, \quad (2.7)$$

où $\bar{y}_g = \sum_{j \in s_g} y_j/n_g$, $g = 1, 2$,

$$\bar{\mu}_{Ra} = \bar{\mu}_2 + a(\bar{y}_1 - \bar{y}_2) \quad (2.8)$$

et $\bar{\mu}_2 = \sum_{j \in s_2} \mu_j/n_2$, pour $a$, une constante, étant donné le sous-échantillon $S_1$. On peut montrer que la valeur de $a$ qui minimise $\mathrm{Var}(\bar{b}_{ga})$ est

$$a = r \quad \text{pour} \quad g = 1,$$

ou (2.9)

$$a = r - 1 \quad \text{pour} \quad g = 2.$$

Par conséquent, pour $g = 1$ ou 2, la valeur "optimale" de $\bar{b}_{ga}$ est

$$\bar{b}_{\mathrm{opt}} = \bar{y}_1 - [\bar{\mu}_2 + r(\bar{y}_1 - \bar{y}_2)], \quad (2.10)$$

qui diffère de TÉN par le terme $(r - 1)(\bar{y}_1 - \bar{y}_2)$. Comme, en règle générale, $\bar{y}_1 \neq \bar{y}_2$, TÉN est optimal seulement si $r = 1$. On peut montrer que ce cas correspond à celui où $\gamma_1$ est égal à 1 dans l'équation (2.1).

Dans cet article, nous allons étudier d'autres estimateurs que TÉN qui renferment de l'information sur $y$ pour les unités de l'ensemble $S_1 \sim S_2$ ainsi que de l'information sur des variables auxiliaires $x$. Afin d'illustrer les concepts, nous allons nous limiter au départ aux modèles linéaires "sans ordonnée à l'origine", c'est-à-dire aux modèles pour lesquels $\gamma_0 = 0$ dans l'équation (2.1). Cette importante classe de modèles comprend l'estimateur par différence de même que les estimateurs par quotient.

### 2.2 Méthodes d'estimation fondées sur la prédiction modéliste

La littérature statistique traite abondamment, en ce qui regarde le sondage de populations finies, les méthodes d'estimation de paramètres de population fondées sur la prédiction modéliste. Cochran (1977) et d'autres auteurs ont démontré les fondements de modèle de l'estimateur par quotient, d'usage très courant. Il existe aussi beaucoup d'ouvrages dans lesquels on pose le choix entre l'utilisation de poids qui découlent d'hypothèses explicites de modèle dans l'estimation pour enquêtes complexes et la suppression des poids d'échantillonnage. Les adeptes de l'estimation dite "à modèles" déconseillent l'utilisation de poids dans l'estimation de paramètres (voir, par exemple, Royall et Herson 1973 et Royall et Cumberland 1981). Ils prétendent qu'en ce qui concerne le sondage de populations finies, les probabilités d'échantillonnage – qu'elles soient égales ou inégales – importent peu une fois que l'échantillon est formé. Les critères de fiabilité utilisés pour les échantillons fondés sur un modèle découlent des hypothèses de distribution du modèle plutôt que des distributions d'échantillonnage. Si on choisit un modèle adéquat pour décrire la relation entre la variable de réponse et d'autres variables étudiées dans l'enquête, on peut obtenir des estimateurs "non biaisés selon le modèle" de paramètres de la population qui sont plus fiables que des estimateurs comprenant des poids.

En contrepartie, il y a les adeptes de l'échantillonnage fondé sur un plan. Au lieu de reconnaître les hypothèses fondées sur un modèle, ce groupe de personnes supposent qu'un estimateur tiré d'une enquête compte comme une réalisation parmi une vaste population de réalisations possibles et que chaque réalisation dépend de l'échantillon formé. La distribution des valeurs de l'estimateur est appelée "distribution d'échantillonnage de l'estimateur" lorsqu'on tient compte de tous les échantillons qui peuvent être constitués à l'aide du plan d'échantillonnage. Les critères qui servent à évaluer les estimateurs selon cette approche portent alors sur les caractéristiques des distributions d'échantillonnage des estimateurs. Dans ce contexte, la pondération des estimateurs est indispensable pour obtenir des estimateurs sans biais, si on a recours à l'échantillonnage avec probabilités inégales.

Bien que les estimateurs de $\bar{B}$ considérés ici représentent les trois catégories d'estimateurs, le but de cet article n'est pas nécessairement de comparer les estimateurs de chaque catégorie, c'est-à-dire estimateurs fondés sur un plan, estimateurs dépendants d'un modèle et estimateurs fondés sur un modèle. Nous cherchons plutôt, en premier lieu, à élaborer une méthode systématique pour évaluer des estimateurs pour un plan d'échantillonnage à deux phases donné. Nous posons le problème en ces termes: étant donné un plan d'échantillonnage à deux phases et des estimateurs de $B = N\bar{B}$ désignés par $\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_p$, comment l'analyste détermine-t-il l'estimateur à erreur quadratique moyenne minimum? Comme deuxième objectif, nous nous proposons de définir un certain nombre d'estimateurs et d'en faire l'évaluation à l'aide d'une méthode systématique. À titre d'exemple, nous nous servons de données de l'enquête agricole réalisée en décembre 1990 par le National Agricultural Statistics Service.

### 2.3 Estimateurs considérés dans l'étude

Si nous étendons la notation définie plus haut aux plans à deux phases stratifiés, désignons par $N_h$ la taille de la strate $h$, pour $h = 1, \ldots, L$. Un échantillon à deux

phases est prélevé dans chaque strate au moyen d'un échantillonnage aléatoire simple à chaque phase. Désignons par $n_{1h}$ et $n_{2h} \leq n_{1h}$ la taille des échantillons des phases 1 et 2 respectivement dans la strate $h$. Soient $S_{1h}$ et $S_{2h} \subseteq S_{1h}$ les ensembles d'éléments pour les échantillons des phases 1 et 2 respectivement dans la strate $h$. Supposons que les données suivantes ont été observées ou qu'elles sont connues d'une autre manière:

variables de résultat: $y_i \ \forall \ i \in S_{1h}$

valeurs vraies: $\mu_i \ \forall \ i \in S_{2h}$

variables auxiliaires: $x_i \ \forall \ i \in S_{1h}$.

Supposons en outre que $X_h = \sum_{i \in U_h} x_i$ est connu, pour $h = 1, \ldots, L$, où $U_h$ est l'ensemble d'éléments pour la strate $h$.

### 2.3.1 Estimateurs pondérés de $M$ et $B$

Pour des raisons de commodité, nous allons considérer l'estimation du biais de l'estimateur d'un total de population désigné par $M$. L'estimateur habituel de $M = N\bar{M}$ est l'estimateur stratifié sans biais défini par l'équation

$$\hat{M}_{2st} = \sum_h N_h \bar{\mu}_{2h}, \tag{2.11}$$

où $\bar{\mu}_{2h} = \sum_{i \in S_{2h}} \mu_i / n_{2h}$. L'estimateur correspondant de $B = N\bar{B}$ équivaut à $N$ fois le TÉN défini en (1.1). Pour les échantillons stratifiés, l'estimateur est

$$\hat{B}_{2st} = \hat{Y}_{2st} - \hat{M}_{2st}, \tag{2.12}$$

où $\hat{Y}_{2st} = \sum_h N_h \bar{y}_{2h}$ et $\bar{y}_{2h} = \sum_{i \in S_{2h}} y_i / n_{2h}$. Notons que (2.12) ne renferme pas l'information sur $y$ pour les unités désignées par $i \in S_{1h} \sim S_{2h}$. En revanche, l'estimateur ci-dessous utilise toute l'information sur $y$:

$$\hat{B}_{12st} = \hat{Y}_{1st} - \hat{M}_{2st}, \tag{2.13}$$

où $\hat{Y}_{1st} = \sum_h N_h \bar{y}_{1h}$ et $\bar{y}_{1h} = \sum_{i \in S_{1h}} y_i / n_{1h}$.

On peut définir un certain nombre d'estimateurs dépendants d'un modèle pour des plans à deux phases stratifiés. Ces estimateurs peuvent être de deux types: indépendants ou combinés (voir, par exemple, Cochran 1977, pp. 327-330). En outre, on peut appliquer les coefficients de correction aux estimateurs de strate de la phase 1 ou de la phase 2. Comme les tailles d'échantillons de strate sont généralement faibles dans les échantillons à deux phases, nous n'étudierons ici que des estimateurs combinés.

Étant donné que cet article insiste sur l'élaboration de méthodes d'estimation de l'erreur systématique de mesure par la prédiction modéliste et sur leur évaluation, nous allons considérer un cas simple, notamment un cas particulier du modèle (2.1), où $\gamma_0 = 0$ (c.-à-d. le modèle "sans ordonnée à l'origine"). Néanmoins, on peut étendre sans difficultés ce raisonnement aux modèles multidimensionnels avec ordonnée à l'origine; ce type de généralisation fera l'objet d'un article futur. Donc, si nous posons $\gamma_0 = 0$ dans (2.1), nous obtenons

$$y_i = \gamma \mu_i + \epsilon_i, \tag{2.14}$$

où $\gamma$ est une constante inconnue et, par hypothèse, $\epsilon_i \sim (0, \sigma_\epsilon^2 \mu_i)$. L'estimateur gaussien de $\gamma$ est $\hat{\gamma} = \bar{y}_{2st} / \bar{\mu}_{2st}$, où $\bar{y}_{2st} = \hat{Y}_{2st}/N$ et $\bar{\mu}_{2st} = \hat{M}_{2st}/N$. Par conséquent, un estimateur dépendant d'un modèle de $\mu_i$ est $y_i/\hat{\gamma} = \bar{\mu}_{2st} y_i / \bar{y}_{2st}$ et un estimateur équivalent de $M$ est

$$\hat{M}_{2stR} = \frac{\hat{M}_{2st}}{\hat{Y}_{2st}} \hat{Y}_{1st}. \tag{2.15}$$

Si nous utilisons cet estimateur de $M$, voici deux estimateurs de $B$ qui correspondent à (2.12) et à (2.13):

$$\hat{B}_{2stR} = \hat{Y}_{2st} - \hat{M}_{2stR} \tag{2.16}$$

et

$$\hat{B}_{12stR} = \hat{Y}_{1st} - \hat{M}_{2stR}. \tag{2.17}$$

On peut obtenir un troisième estimateur de $B$ à l'aide du modèle

$$y_i = \beta x_i + e_i, \tag{2.18}$$

où $\beta$ est une constante et $e_i \sim (0, \sigma_e^2 x_i)$. Cela donne un estimateur par quotient de $Y$,

$$\hat{Y}_{xstR} = \frac{\bar{y}_{1st}}{\bar{x}_{1st}} X. \tag{2.19}$$

Par conséquent, l'estimateur correspondant de $B$ est

$$\hat{B}_{x2stR} = \hat{Y}_{xstR} - \hat{M}_{2stR}. \tag{2.20}$$

Enfin, Särndal, Swensson et Wretman (1992, p. 360) proposent un estimateur général de $M$ pour l'échantillonnage à deux phases. En appliquant leur équation 9.7.2 au modèle (2.14) suivant un échantillonnage stratifié, nous obtenons

$$\hat{M}_{SSW} = \hat{M}_{2stR} + \frac{\bar{\mu}_{2st}}{\bar{x}_{2st}} (X - \hat{X}_{1st}). \tag{2.21}$$

Notons que l'estimateur ci-dessus n'est ni plus ni moins que l'estimateur (2.15), auquel on additionne l'estimateur non biaisé de la valeur nulle. Le nouvel estimateur peut avoir une variance moins élevée que celle de $\hat{M}_{2stR}$ si le second terme du membre de droite de l'équation (2.21) est corrélé négativement avec $\hat{M}_{2stR}$. De la même manière, l'estimateur de $Y$ que proposent les auteurs ci-dessus se ramène à $\hat{Y}_{xstR}$, défini en (2.19). Par conséquent, l'estimateur correspondant de $B$ est

$$\hat{B}_{SSW} = \hat{Y}_{xstR} - \hat{M}_{SSW}, \tag{2.22}$$

qui équivaut à $\hat{B}_{SSW} = B_{x2stR}$ plus le second terme du membre de droite de l'équation (2.21).

### 2.3.2 Estimateurs non pondérés de $M$ et $B$

Réécrivons $M$ sous la forme

$$M = \sum_{i \in S_2} \mu_i + \sum_{i \in S_1 - S_2} \mu_i + \sum_{i \in U - S_1} \mu_i \tag{2.23}$$

$$= M_{(2)} + M_{(1-2)} + M_{(-1)},$$

par exemple, où $S_g = \cup_{h=1}^{L} S_{gh}$, $g = 1, 2$. La façon de procéder pour l'estimation non pondérée basée sur un modèle est de remplacer $\mu_i$ dans $M_{(1-2)}$ et $M_{(-1)}$ par une prévision, $\hat{\mu}_i$, tirée d'un modèle.

Si nous utilisons le modèle (2.14), nous avons comme estimateur de $\mu_i$

$$\hat{\mu}_i = y_i / \bar{\gamma},$$

où, désormais, $\bar{\gamma} = \bar{y}_2 / \bar{\mu}_2$. Nous avons donc comme estimateur de $M_{(1-2)}$

$$\hat{M}_{(1-2)} = \frac{\bar{\mu}_2}{\bar{y}_2} \sum_{i \in S_1 - S_2} y_i \tag{2.24}$$

$$= \frac{\bar{\mu}_2}{\bar{y}_2} (n_1 \bar{y}_1 - n_2 \bar{y}_2),$$

où $\bar{y}_g = \sum_{i \in S_g} y_i / n_g$, $\bar{\mu}_2 = \sum_{i \in S_2} \mu_i / n_2$, et $n_g = \sum_h n_{gh}$, pour $g = 1, 2$. De plus, en utilisant le modèle

$$\mu_i = \delta x_i + \xi_i, \tag{2.25}$$

où $\delta$ est une constante et $\xi_i \sim (0, \sigma_\xi^2 x_i)$, nous obtenons

$$\hat{M}_{(-1)} = \frac{\bar{\mu}_2}{\bar{x}_2} X_{U-S_1}, \tag{2.26}$$

où $X_{U-S_1} = \sum_{i \in U - S_1} X_i$. Nous avons donc comme estimateur basé sur un modèle de $M$

$$\hat{M}_M = M_{(2)} + \hat{M}_{(1-2)} + \hat{M}_{(-1)} \tag{2.27}$$

$$= \hat{M}_{(1)} + \hat{M}_{(-1)},$$

où $\hat{M}_{(1)} = n_1 \bar{\mu}_2 \bar{y}_1 / \bar{y}_2$.

De la même manière, nous pouvons réécrire $Y$ sous la forme

$$Y = \sum_{i \in S_1} y_i + \sum_{i \in U - S_1} y_i \tag{2.28}$$

$$= Y_{(1)} + Y_{(-1)}$$

et nous cherchons à prédire $y_i$ dans $Y_{(-1)}$. En utilisant le modèle (2.18), nous avons comme estimateur basé sur un modèle de $Y_{(-1)}$

$$\hat{Y}_{(-1)} = \frac{\bar{y}_1}{\bar{x}_1} X_{U-S_1}$$

et, par conséquent, comme estimateur de $Y$

$$\hat{Y}_M = Y_{(1)} + \hat{Y}_{(-1)}. \tag{2.29}$$

$B$ est donc estimé au moyen de la formule

$$\hat{B}_M = \hat{Y}_M - \hat{M}_M. \tag{2.30}$$

On peut aussi construire des versions de $\hat{B}_{2stR}$, $\hat{B}_{12stR}$, $\hat{B}_{x2stR}$ et $\hat{B}_M$ qui résistent mieux aux valeurs aberrantes de modèle. On peut créer ces estimateurs, désignés par $\tilde{B}_{2stR}$, $\tilde{B}_{12stR}$, $\tilde{B}_{x2stR}$ et $\tilde{B}_M$ respectivement, en supprimant les points de données qui s'écartent sensiblement des prédictions du modèle et en calculant les estimateurs basés sur un modèle ou les estimateurs dépendants d'un modèle à l'aide des données restantes. À titre d'exemple, prenons l'estimateur $\hat{M}_{2stR}$ défini en (2.15). Pour cet estimateur, posons

$$(n_{2h} - 1) s_{res,h}^2 = \sum_{\mu_{hi} \neq 0} \frac{(y_{hi} - \hat{\gamma}\, \mu_{hi})^2}{\mu_{hi}}, \tag{2.31}$$

comme la somme des carrés des résidus pour le modèle (2.14). Ensuite, pour calculer l'estimateur de $\gamma$, on n'utilise que les unités désignées par $i \in \tilde{S}_{2h}$, où $\tilde{S}_{2h} = \{ i \in S_{2h} : | y_{ih} - \hat{\gamma} \mu_{ih} | \leq 3 s_{res,h} \sqrt{\mu_{hi}} \}$. Si l'on désigne cet estimateur par $\gamma$, l'estimateur de $M$ est $\tilde{M}_{2stR} = \hat{Y}_{1st} / \tilde{\gamma}$, où $\tilde{\gamma} = \tilde{y}_{2st} / \tilde{\mu}_{2st}$ et $\tilde{\mu}_{2st}$ et $\tilde{y}_{2st}$ sont les moyennes stratifiées de $\mu_i$ et $y_i$ pour $i \in \tilde{S}_{2h}$. Les autres estimateurs robustes se calculent de la même façon.

De nombreux autres estimateurs non pondérés basés sur un modèle peuvent être examinés par rapport à l'échantillonnage à deux phases. On peut, par exemple, ajouter une ordonnée à l'origine dans les modèles (2.14), (2.18) et (2.26). De plus, on peut définir séparément des paramètres de pente et d'ordonnée à l'origine pour chaque strate ou chaque combinaison de strates.

## 2.4 Estimation des erreurs quadratiques moyennes à l'aide d'estimateurs d'auto-amorçage

Bien qu'il soit possible, suivant les hypothèses basées sur un plan ou basées sur un modèle appropriées, de calculer des estimations en forme analytique de la variance des estimateurs considérés dans cette étude, nous avons préféré utiliser une méthode de rééchantillonnage intensif sur ordinateur. Primo, nous cherchons une méthode qui est facile à appliquer puisque notre étude est susceptible de porter sur un grand nombre d'estimateurs. Secundo, il est essentiel d'évaluer chaque estimateur à l'aide des mêmes critères et à cette fin, une méthode cohérente d'estimation de la variance est indispensable. Nous devons donc employer une méthode d'estimation de la variance qui

puisse s'appliquer à des estimateurs de tous degrés de complexité, suivant des hypothèses qui sont cohérentes et qui ne dépendent d'aucune hypothèse de modèle. Il est notoire que les méthodes d'estimation de la variance basées sur un modèle sont très sensibles à la défaillance du modèle (voir, par exemple, Royall et Herson 1973; Royall et Cumberland 1978; et Hansen, Madow et Tepping 1983). Dans Royall et Cumberland (1981), on examine plusieurs solutions dans la perspective de l'erreur systématique, notamment l'estimateur jackknife.

Notre approche ressemble à celle de Royall et Cumberland sauf que nous utilisons un estimateur d'auto-amorçage au lieu d'un estimateur jackknife. Pour des observations indépendantes et identiquement distribuées, Efron et Gong (1983) montrent que cet estimateur diffère de l'estimateur jackknife par un facteur de $n/(n-1)$ pour des échantillons de taille $n$. Par conséquent, les caractéristiques de robustesse que décrivent Royall et Cumberland pour l'estimateur jackknife s'appliquent aussi à l'estimateur utilisé.

D'autres caractéristiques de l'estimateur d'auto-amorçage nous ont amenés à le préférer à d'autres méthodes de rééchantillonnage. La méthode jackknife et la méthode BRR (balanced repeated replication) sont difficilement adaptables au plan de sondage à deux phases de notre étude. Au contraire, la méthode d'auto-amorçage s'adapte facilement à l'échantillonnage à deux phases. De plus, grâce à une étude de simulation, Rao et Wu (1988) ont démontré que le niveau d'efficacité des intervalles de confiance se compare avantageusement à celui des intervalles de confiance jackknife et BRR en ce qui concerne les plans d'échantillonnage complexes.

Dans le cheminement de l'étude, nous étendons la méthode élaborée par Bickel et Freedman (1984) pour l'échantillonnage à une phase stratifié à l'échantillonnage à deux phases stratifié. Puisque la méthode d'auto-amorçage est exécutée pour chaque strate prise individuellement, pour des raisons de simplicité nous allons décrire la méthode pour le cas à une strate.

### 2.4.1 Estimation de la variance

Étendre la méthode d'auto-amorçage à l'échantillonnage à deux phases n'équivaut pas simplement à soumettre les échantillons à une phase à un nouvel échantillonnage. Rappelons-nous que l'on connaît des valeurs vraies uniquement pour les unités de $S_2$ et qu'en conséquence, l'échantillonnage doit nécessairement se limiter aux unités contenues dans $S_2$. Désignons donc par $S_1$ et $S_2$ les échantillons de la phase 1 et de la phase 2 respectivement, tirés de $U$ au moyen d'un ÉASSR. Posons $S_{1\sim 2}$ comme l'ensemble d'unités, $S_1 \sim S_2$. Posons $\hat{\Theta} = \hat{\Theta}(S_{1\sim 2}, S_2)$ comme un estimateur de $\Theta$ qui peut être une fonction des observations relatives aux unités contenues dans $S_2$ aussi bien que $S_{1\sim 2}$. Définissons $N$, $n_1$, $n_2$ et $n_{1\sim 2}$ comme les tailles respectives des ensembles $U$, $S_1$, $S_2$ et $S_{1\sim 2}$. Voyons comment on applique la méthode d'auto-amorçage pour obtenir des estimations de $\mathrm{Var}(\hat{\Theta})$.

Le cas le plus simple est celui où $N/n_1$ est un nombre entier, disons $k$. Premièrement, nous constituons l'ensemble d'unités de la pseudo-population

$$U_A^* = U_{A(2)}^* \cup U_{A(1\sim 2)}^*, \tag{2.32}$$

où $U_{A(2)}^*$ est formé de $k$ répliques des unités de $S_2$ et $U_{A(1\sim 2)}^*$ est formé de $k$ répliques des unités de $S_{1\sim 2}$. Nous exécutons ensuite les trois étapes suivantes:

1. Tirer un ÉASSR de taille $n_2$ dans $U_{A(2)}^*$ et désigner cet ensemble par $S_2^*$.
2. Tirer un ÉASSR de taille $n_{1\sim 2}$ dans $U_{A(1\sim 2)}^*$ et désigner cet ensemble par $S_{1\sim 2}^*$.
3. Calculer $\hat{\Theta}_1^* = \hat{\Theta}_1(S_{1\sim 2}^*, S_2^*)$, qui a la même forme fonctionnelle que $\hat{\Theta}(S_{1\sim 2}, S_2)$ mais est calculé pour les $n_1 = n_{1\sim 2} + n_2$ unités contenues dans $S_1^* = S_{1\sim 2}^* \cup S_2^*$.

Nous répétons les étapes ci-dessus un nombre élevé de fois, $Q$, pour obtenir $\Theta_1^*, \ldots, \Theta_Q^*$. Nous pouvons ainsi définir un estimateur de $\mathrm{Var}(\hat{\Theta})$ par la formule

$$\mathrm{var}_{BSS}(\hat{\Theta}) = \sum_{q=1}^{Q} \frac{(\hat{\Theta}_q^* - \hat{\Theta}_\cdot^*)^2}{Q-1}, \tag{2.33}$$

où $\hat{\Theta}_\cdot^* = \sum_{q=1}^{Q} \hat{\Theta}_q^*/Q$.

En nous servant des méthodes de Rao et Wu (1988), nous pouvons maintenant montrer que $\mathrm{var}_{BSS}(\hat{\Theta})$ est un estimateur convergent de $\mathrm{Var}(\hat{\Theta})$. Si $N = kn_1 + r$, où $0 < r < n_1$, on doit procéder d'une autre manière en utilisant la méthode de Bickel et Freedman. Premièrement, on constitue, comme précédemment, la pseudo-population $U_A^*$, formée de $kn_1$ unités. De plus, on crée la pseudo-population $U_B^* = U_{B(1\sim 2)} \cup U_{B(2)}^*$, de taille $(k+1)n_1$, où $U_{B(1\sim 2)}^*$ et $U_{B(2)}$ sont composés de $(k+1)$ répliques des unités de $S_{1\sim 2}$ et de $S_2$ respectivement. Alors, pour un nombre $\alpha Q$ d'échantillons obtenus par la méthode d'auto-amorçage, tirons $S_1^* = S_{1\sim 2}^* \cup S_2^*$ dans $U_A^*$, et pour $(1-\alpha)Q$ échantillons, tirons $S_1^*$ dans la pseudo-population $U_B^*$ à l'aide de la procédure à trois étapes décrite ci-dessus, où

$$\alpha = \left(1 - \frac{r}{n_1}\right)\left(1 - \frac{r}{N-1}\right). \tag{2.34}$$

### 2.4.2 Estimation du biais et de l'EQM

La méthode d'auto-amorçage permet aussi d'estimer le biais d'un estimateur. L'estimateur habituel du biais (voir Efron et Gong 1983; Rao et Wu 1988) est $b(\hat{\Theta}) = \hat{\Theta}_\cdot^* - \hat{\Theta}$, où $\hat{\Theta}_\cdot^* = \sum_q \hat{\Theta}_q^*/Q$ et $\hat{\Theta}$ est l'estimation calculée à l'aide de l'échantillon complet. Notons que $\hat{\Theta}_q^*(q = 1, \ldots, Q)$ et $\hat{\Theta}$ ont la même forme fonctionnelle et reposent sur les mêmes hypothèses de modèle. Par conséquent, $b(\hat{\Theta})$ ne tient pas compte de l'effet de la défaillance du modèle sur le biais. Nous proposons donc un autre estimateur du biais qui, croyons-nous, est supérieur à $b(\hat{\Theta})$.

Rappelons-nous, d'après l'équation (2.4), que $\bar{B} = \mathrm{E}(y_i - \mu_i)$, où $\mathrm{E}(\cdot)$ désigne l'espérance mathématique par rapport à la distribution des erreurs d'échantillonnage et la distribution des erreurs de mesure. Nous pouvons

donc réécrire $\bar{B}$ sous la forme $\bar{B} = \sum_{i=1}^{N} (Y_i - \mu_i)/N$, où $Y_i = \mathrm{E}(y_i \mid i)$. Comme $Y_i$ est inconnu et inobservable pour tous $i \in U$, $\bar{B}$ est aussi inconnu et inobservable. Par conséquent, nous allons créer une pseudo-population semblable à $U$, que nous désignerons par $U^*$, de telle sorte que $\bar{B}^* = \mathrm{E}^*(y_i - \mu_i)$ soit connu, $\mathrm{E}^*(\cdot)$ étant l'espérance mathématique par rapport à la distribution des erreurs de mesure et à la distribution des erreurs d'échantillonnage rattachées à $U^*$.

Soit $U^* = \cup_{h=1}^{L} U_h^*$, où $U_h^*$ est formé de $k_h = N_h/n_{1h}$ répliques des unités de $S_{1h}$. Nous supposons ici que $k_h$ est un nombre entier, mais nous allons assouplir cette hypothèse plus loin. De plus, désignons par $y_i^*$ la valeur de la caractéristique pour l'unité $i \in U^*$. Cette valeur est égale à la valeur $y_i$ pour l'unité correspondante dans $S_1$. Par conséquent, le total des $y_i^*$ pour la population est $Y^* = \sum_{i \in U^*} y_i^* = \hat{Y}_{1st}$, $\hat{Y}_{1st}$ étant défini en (2.13). De la même manière, définissons la valeur vraie pour l'unité $i \in U^*$ comme $\mu_i^* = \mu_j$, pour $i \in U^*$ correspondant à $j \in S_2$. Pour $j \in S_{1\sim 2}$, $\mu_j$, $\mu_i^*$ est inconnu; cependant, nous pourrions, en ce qui concerne notre pseudo-population, générer des pseudo-valeurs pour $\mu_i^*$, de telle sorte que $M^* = \sum_{i \in U^*} \mu_i^* = \hat{M}_{2st}$, où $\hat{M}_{2st}$ est défini en (2.11). Par conséquent, pour $U^*$, $B^* = \hat{Y}_{1st} - \hat{M}_{2st} = \hat{B}_{12st}$, défini en (2.13). Comme nous le verrons, il n'est pas nécessaire de générer des pseudo-valeurs pour $\mu_i^*$ lorsqu'on veut évaluer le biais des estimateurs de $B^*$.

Notons que selon un plan d'échantillonnage stratifié, $U^* = U_A^*$, conformément à ce qui est mentionné dans la section 2.4. En outre, la méthode d'auto-amorçage décrite dans cette section équivaut à un échantillonnage répété dans $U^*$ et les estimateurs $\hat{\Theta}_1, \ldots, \hat{\Theta}_p$ de $B$ peuvent aussi être vus comme des estimateurs de $B^*$. Comme $B^*$ est connu, le biais de $\hat{\Theta}$, en tant qu'estimateur de $B^*$, est $\hat{B}^* = \hat{\Theta} - B^*$ et l'EQM correspondante peut être estimée par la formule

$$\widehat{\mathrm{EQM}} = \sum_q (\hat{\Theta}_q - B^*)^2/Q$$
$$\doteq \mathrm{var}_{BSS}(\hat{\Theta}) + (\hat{\Theta}^* - B^*)^2, \qquad (2.35)$$

où $\mathrm{var}_{BSS}(\hat{\Theta})$, $\hat{\Theta}_q$, et $\hat{\Theta}^*$ sont définis dans la section 2.4. On peut vérifier facilement que ces résultats sont aussi valides lorsque $k_h$ est un nombre non entier.

En conclusion, la méthode d'auto-amorçage offre un moyen d'évaluer l'EQM de divers estimateurs de $B^*$. En outre, la pseudo-population $U^*$ est une reconstitution de $U$ basée sur des répliques des valeurs relatives aux unités de $S_1$ et $S_2$. Il est donc raisonnable de se servir de $\hat{B}^*$ et de $\widehat{\mathrm{EQM}}^*$ pour évaluer divers estimateurs de $B$.

## 3. APPLICATION À L'ENQUÊTE AGRICOLE

### 3.1 Description de l'enquête

Le National Agricultural Statistics Service (NASS) effectue annuellement une série d'enquêtes qui est connue sous le nom de Agricultural Survey (AS) Program (programme d'enquêtes agricoles (EA)). Ces enquêtes visent à recueillir des données sur certains produits agricoles au niveau national et au niveau des États. De 1988 à 1990, des réinterviews ont été effectuées en décembre de chacune de ces années dans six États: l'Indiana, l'Iowa, le Minnesota, le Nebraska, l'Ohio et la Pennsylvanie; ces réinterviews avaient pour but d'évaluer l'erreur systématique de mesure dans les données recueillies par interview téléphonique assistée par ordinateur (ITAO). Les techniques de réinterview utilisées à ces trois occasions ressemblent beaucoup à celles appliquées par le Census Bureau des É.-U. (voir, par exemple, Forsman et Schreiner 1991). Cependant, contrairement au programme du Census Bureau, les études du NASS visent principalement à estimer l'erreur systématique de mesure plutôt qu'à évaluer le rendement des intervieweurs.

Comme nous l'avons mentionné plus haut, seules les unités répondantes qui avaient subi une interview téléphonique assistée par ordinateur lors des EA pouvaient être sélectionnées pour faire partie de l'échantillon de réinterview. Cette restriction était surtout motivée par des raisons d'économie, de temps et de commodité. Toutefois, comme une forte proportion des EA sont réalisées par ITAO, l'information qu'on a pu recueillir sur l'erreur systématique de mesure dans les EA grâce à l'échantillon de réinterview sera précieuse pour l'ensemble du programme d'enquêtes agricoles.

**Tableau 1**

Taille des échantillons selon la question d'enquête

| Question | $x$ | $y$ | $\mu$ |
|---|---|---|---|
| | $U$ | $S_1$ | $S_2$ |
| Stock de blé (toutes variétés confondues) | 108,267 | 8,176 | 1,157 |
| Superficie ensemencée en maïs | 225,269 | 8,211 | 1,157 |
| Stocks de maïs | 225,269 | 7,990 | 1,115 |
| Superficie cultivable | 278,045 | 8,274 | 1,141 |
| Capacité d'entreposage du grain | 207,460 | 8,126 | 1,104 |
| Superficie ensemencée en soja | 171,761 | 8,211 | 1,156 |
| Stocks de soja | 171,761 | 8,113 | 1,130 |
| Superficie totale des terres agricoles | 276,450 | 8,309 | 1,159 |
| Cheptel porcin | 248,571 | 8,247 | 1,142 |
| Semis de blé d'hiver | 108,267 | 8,211 | 1,150 |

Pour effectuer les réinterviews, le NASS comptait sur une équipe formée de surveillants des opérations sur le terrain et d'intervieweurs chevronnés. Cette équipe, qui était différente de celle affectée aux interviews téléphoniques assistées par ordinateur, a effectué des réinterviews directes auprès d'un sous-échantillon des participants aux EA en

reprenant un certain nombre de questions des enquêtes initiales. Afin de réduire au maximum les erreurs de mémoire, on a fait les réinterviews dans les 10 jours qui suivaient l'interview initiale. On s'efforçait de concilier les différences entre les réponses données à l'EA et celles données à la réinterview pour déterminer la valeur "vraie". Beaucoup d'efforts ont été consacrés à l'élaboration de procédures, à la formation et à la supervision du processus de réinterview pour faire en sorte que les réponses définitives soient toutes justes. Dans la grande majorité des cas, le libellé des questions posées dans la réinterview était le même que dans l'enquête d'origine. Les intervieweurs tentaient de communiquer avec la personne la plus avertie pour s'assurer de l'exactitude des réponses données.

Dans ce rapport, nous n'analysons que les données de 1990. Le Tableau 1 donne la taille des échantillons de réinterview utilisés pour cette étude.

### 3.2 Comparaison des estimateurs de $M$ et de $B$

En nous servant des données de l'enquête agricole de décembre 1990 et des données correspondantes recueillies à la réinterview, nous avons comparé les estimateurs élaborés dans la section précédente. Nous avons calculé les estimations de l'erreur type et de l'erreur quadratique moyenne à l'aide de la méthode d'auto-amorçage de Bickel-Freedman décrite dans la section 2.4 ($Q$ = 300 échantillons). Le Tableau 2 contient les résultats relatifs à six estimateurs, soit $\hat{B}_{2st}$, l'estimateur de différence classique; $\hat{B}_{x2stR}$, l'estimateur par quotient pondéré; $\tilde{B}_{x2stR}$, la version robuste (suppression de valeurs aberrantes) de $\hat{B}_{x2stR}$; $\hat{B}_{SSW}$, l'estimateur de Särndal, Swensson et Wretman; $\hat{B}_M$, l'estimateur à modèle non pondéré; et $\tilde{B}_M$, la version robuste (suppression de valeurs aberrantes) de $\hat{B}_M$.

**Tableau 2**

Comparaison des estimateurs étudiés, par rapport à $B^*$, la valeur du biais pour la pseudo-population †

| Caractéristique | $B^*$ | $\hat{B}_{2st}$ | $\hat{B}_{x2stR}$ | $\tilde{B}_{x2stR}$ | $\hat{B}_{SSW}$ | $\hat{B}_M$ | $\tilde{B}_M$ |
|---|---|---|---|---|---|---|---|
| Stocks de blé (toutes variétés) | 42.3 | −6.1 | 103.2 | −94.2 | −0.9‡ | 19.2‡ | 10.6‡ |
| | | (12.3) | (17.6) | (16.5) | (24.8) | (16.5) | (16.7) |
| Superficie ensemencée en maïs | −1.8 | 1.1‡ | 11.7 | 10.1 | 0.3‡ | −4.7‡ | −5.0 |
| | | (1.1) | (1.3) | (1.1) | (1.2) | (1.9) | (1.5) |
| Stocks de maïs | −6.4 | −5.4‡ | 2.4 | 0.2 | −6.5‡ | −7.9‡ | −9.3‡ |
| | | (1.5) | (1.6) | (1.3) | (1.6) | (2.4) | (2.2) |
| Superficie cultivable | 27.0 | −19.6 | −15.0 | 7.0 | −19.6 | −36.8 | −12.8 |
| | | (8.3) | (8.3) | (3.1) | (8.2) | (11.0) | (4.0) |
| Capacité d'entreposage du grain | −3.37 | 1.4‡ | 32.3 | 29.5 | −0.1‡ | −6.9 | −6.8 |
| | | (3.7) | (3.7) | (2.6) | (3.9) | (3.0) | (2.5) |
| Superficie ensemencée en soja | −4.4 | 0.8 | 13.0 | 9.9 | −0.3 | −2.9 | −2.7 |
| | | (0.8) | (1.0) | (0.9) | (1.0) | (1.1) | (1.0) |
| Stocks de soja | −0.01 | 2.8‡ | 21.3 | 5.0 | 0.2‡ | −11.0 | −8.8 |
| | | (3.1) | (2.9) | (2.3) | (3.5) | (3.6) | (3.4) |
| Superficie totale des terres agricoles | −20.0 | −24.7‡ | −18.8‡ | −2.6 | −25.7‡ | −44.5‡ | −21.2 |
| | | (10.4) | (12.5) | (7.6) | (10.7) | (13.4) | (5.8) |
| Cheptel porcin | −0.1 | −2.1 | 3.4 | −0.0‡ | −2.2‡ | −2.5‡ | −1.6‡ |
| | | (0.9) | (1.1) | (1.0) | (1.1) | (1.3) | (1.0) |
| Semis de blé d'hiver | −0.6 | −0.5‡ | 3.8 | 1.8 | −1.2‡ | 1.1 | 1.1 |
| | | (0.4) | (0.6) | (0.5) | (0.6) | (0.4) | (0.4) |
| Nombre de postes pour lesquels $B^*$ est incluse dans l'I.C. | | 6 | 1 | 1 | 8 | 5 | 3 |
| C.V. moyen | | 1.01 | .30 | 11.1 | 9.5 | .41 | .48 |
| Valeur moyenne de la racine carré de l'EQM | | 13.2 | 22.4 | 25.2 | 12.9 | 14.9 | 10.8 |
| Valeur moyenne du biais relatif en valeur absolue | | 30.8 | 220.0 | 53.4 | 4.9 | 113.1 | 91.3 |

† Erreurs types entre parenthèses.
‡ Un intervalle de confiance à 95% contient le paramètre de la pseudo-population.

### 3.3 Sommaire des résultats

Le Tableau 2 présente un sommaire des résultats de notre étude. La première colonne de chiffres indique la valeur connue de $B^* = \mathrm{E}(y_i^* - \mu_i^*)$, qui est le paramètre de biais pour la pseudo-population $U^*$. Les autres colonnes de chiffres contiennent les valeurs des estimateurs, avec les erreurs types entre parenthèses (e.t. $(\hat{\Theta}) = \sqrt{\mathrm{var}_{BSS}(\hat{\Theta})}$. Les quatre dernières lignes du tableau indiquent, respectivement,

a) le nombre de postes (sur 10) pour lesquels la valeur $B^*$ est incluse dans un intervalle de confiance à 95%,

b) le coefficient de variation (C.V.) moyen,

c) la moyenne de la racine carré de $\widehat{\mathrm{EQM}}^*$ et

d) la valeur moyenne du biais relatif en valeur absolue.

Un trait saillant de ces résultats est leur forte disparité d'un estimateur à l'autre pour toutes les caractéristiques étudiées, en particulier les stocks de blé (toutes variétés). Dans ce dernier cas, les estimations varient de $-94.2$ à 103.2. Le Tableau 2 indique aussi, par le symbole ‡, si le paramètre $B^*$ est contenu dans un intervalle de confiance à 95%, notamment $[\hat{\Theta} - 2\mathrm{e.t.}(\hat{\Theta}), \hat{\Theta} + 2\mathrm{e.t.}(\hat{\Theta})]$. Le meilleur estimateur à ce titre est $\hat{B}_{SSW}$, pour lequel nous avons des intervalles de confiance qui contiennent $B^*$ dans huit cas sur dix. L'estimateur $\hat{B}_{2st}$ vient au deuxième rang, avec six postes sur dix, et $\hat{B}_M$, au troisième rang, avec cinq postes sur dix. L'estimateur par quotient classique et sa version robuste donnent les pires résultats, la valeur de $B^*$ étant contenue dans l'intervalle de confiance dans un cas seulement.

Si on examine ces estimateurs du point de vue de l'erreur quadratique moyenne, on fait des constatations différentes. En effet, $\hat{B}_M$ s'impose comme l'estimateur pour lequel la valeur moyenne de la racine carré de l'EQM est la plus faible. Cependant, les estimateurs $\hat{B}_{SSW}$ et $\hat{B}_{2st}$ ne sont pas loin derrière. En outre, $\hat{B}_{SSW}$ est l'estimateur pour lequel la valeur moyenne du biais relatif en valeur absolue est la plus faible. Cet estimateur a produit un biais appréciable dans seulement deux cas sur dix. Les résultats nous permettent donc de croire que $\hat{B}_{SSW}$ est le meilleur estimateur, si l'on se sert du rendement global comme critère d'évaluation.

## 4. CONCLUSIONS ET RECOMMANDATIONS

Dans cet article, nous avons élaboré une méthode générale pour construire et évaluer des estimateurs de prédiction modéliste, pondérés ou non, de l'erreur systématique de mesure pour des plans d'échantillonnage aléatoire à deux phases stratifié. Les estimateurs proposés renferment de l'information sur les observations $y$ (information tirée de l'échantillon de la première phase) et sur une variable auxiliaire $x$. Nous avons aussi considéré et évalué des versions robustes de ces estimateurs. Le but ultime de l'estimation par la prédiction modéliste est de définir des estimateurs qui font un usage "optimal" des données $(y, \mu, x)$. La méthode générale d'estimation et d'évaluation utilisée pour atteindre ce but a été illustrée dans le cas du modèle de régression ordinaire, sans ordonnée à l'origine. Cependant, on peut étendre facilement cette méthode à des modèles multidimensionnels avec ordonnée à l'origine.

Les critères d'évaluation que nous proposons reposent sur des estimations du biais, de la variance et de l'erreur quadratique moyenne calculées à l'aide d'une méthode de rééchantillonnage basée sur l'auto-amorçage. À cette fin, nous avons étendu la méthode de Bickel et Freedman à l'échantillonnage à deux phases. Nous avons montré tant du point de vue analytique que du point de vue empirique que l'estimateur TÉN habituel n'est pas optimal suivant la méthode de prédiction modéliste. Nos analyses ont permis de déterminer que, parmi les six estimateurs étudiés, celui élaboré par Särndal et coll. (1992) était le meilleur estimateur d'après les critères d'évaluation de la méthode d'auto-amorçage. L'introduction d'information supplémentaire dans l'estimation de l'erreur systématique de mesure crée un certain nombre de problèmes pratiques qui peuvent rendre plus coûteuse et ralentir la production des estimations. Premièrement, il faut connaître les valeurs de la variable auxiliaire $x$, à tout le moins sous forme agrégée, pour tous les domaines socio-économiques et géographiques pour lesquels on cherche à obtenir des estimations par la prédiction modéliste. Cela peut représenter une énorme tâche sur le plan de la gestion de données. De plus, la complexité de l'estimateur de la variance qui utilise des méthodes analytiques s'accroît avec la complexité de l'estimateur du biais. Quoique plus simple, la méthode d'estimation de la variance par l'auto-amorçage peut être excessivement coûteuse s'il faut acheter du temps machine. Néanmoins, cet obstacle n'est pas insurmontable, surtout si on dispose d'un micro-ordinateur de forte puissance. Par ailleurs, compte tenu du coût des réinterviews destinées à estimer l'erreur systématique de mesure, une hausse, même modérée, de la précision des estimateurs du biais peut entraîner des économies substantielles.

La prédiction modéliste permet d'extraire des réinterviews le maximum d'information sur le biais dans les réponses. Les estimateurs de prédiction modéliste seront donc généralement plus efficaces que l'estimateur TÉN classique. De plus, la prédiction modéliste offre aussi un moyen d'étendre les estimations du biais à des domaines qui n'ont pas été sondés. Dans les enquêtes de réinterview du NASS par exemple, pour des raisons de commodité et d'économie, l'échantillon était constitué uniquement d'unités qui avaient participé à l'enquête initiale par ITAO. Or, en utilisant des modèles de prédiction qui sont des fonctions des réponses originales et des caractéristiques géographiques et démographiques locales, il serait possible de prédire l'erreur systématique de mesure pour les unités d'enquête qui participent autrement que par ITAO en se fondant sur les caractéristiques locales de ces unités, ce qui serait un genre d'estimation "synthétique". Bien que cette application de l'estimation basée sur un modèle n'ait pas été étudiée dans cet article, elle est un prolongement naturel de la méthode exposée et sera évaluée dans une étude ultérieure.

Nous avons aussi l'intention, dans des études futures, de considérer des modèles multidimensionnels avec ordonnée à l'origine pour l'estimation de l'erreur systématique de mesure. Comme les critères d'évaluation de la méthode d'auto-amorçage élaborés dans cet article sont de nature générale, il n'est pas nécessaire de modifier la méthode d'évaluation pour ajouter des variables dans les modèles d'estimation. En outre, nous prévoyons améliorer et évaluer les hypothèses de modèle et les méthodes de traitement des valeurs aberrantes dans un article ultérieur. Finalement, il nous faut étudier l'effet de la non-vérification des hypothèses de modèle sur l'estimation, en particulier l'hypothèse selon laquelle les données de réinterview ne sont entachées d'aucune erreur. Comme le montre Fuller (1991), si les données de la réinterview sont faillibles mais non biaisées, la variance des valeurs prédites s'accroît mais les valeurs prédites demeurent non biaisées. Par conséquent, on pourrait étudier, suivant ces hypothèses, la précision relative des divers estimateurs de l'erreur systématique de mesure afin de déterminer la robustesse de la méthode de prédiction modéliste.

## REMERCIEMENTS

Les auteurs remercient Manuel Cárdenas pour sa précieuse aide ainsi que les arbitres pour leurs remarques pertinentes.

## BIBLIOGRAPHIE

BICKEL, P., et FREEDMAN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

BIEMER, P., et STOKES, L. (1991). Approaches to the modeling of measurement errors. Dans *Measurement Errors in Surveys*, (Éds. P. Biemer, et coll.). New York: John Wiley and Sons.

COCHRAN, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons.

EFRON, B., et GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 31, 36-48.

FORSMAN, G., et SCHREINER, I. (1991). The design and analysis of reinterview: an overview. Dans *Measurement Errors in Surveys*, (Éds. P. Biemer, et coll.). New York: John Wiley and Sons.

FULLER, W.A. (1991). Regression estimation in the presence of measurement error. Dans *Measurement Errors in Surveys*, (Éds. P.P. Biemer, et coll.). New York: John Wiley and Sons, 617-636.

HANSEN, M., MADOW W., et TEPPING, B. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

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

ROYALL, R., et HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-893.

ROYALL, R., et CUMBERLAND, W. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-361.

ROYALL, R., et CUMBERLAND, W. (1981). The finite-population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

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

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 147–158
Statistique Canada



# Estimation modéliste des taux d'erreur liés au couplage d'enregistrements

J.B. ARMSTRONG et J.E. MAYDA[1]

RÉSUMÉ

On appelle couplage d'enregistrements l'appariement d'enregistrements contenant des données sur des particuliers, des entreprises ou des logements quand on ne dispose pas d'un identificateur unique. Les méthodes utilisées, en pratique, comportent la classification de paires d'enregistrements, comme constituant des liens ou des non-liens, à l'aide d'une procédure automatisée basée sur le modèle théorique présenté par Fellegi et Sunter (1969). L'estimation des taux d'erreur de classification constitue un problème important. Fellegi et Sunter présentent une méthode, afin de calculer des estimations des taux d'erreur de classification, qui découle directement du couplage. Ces estimations faites à l'aide de modèles sont plus faciles à produire que celles obtenues par appariement manuel d'échantillons, méthode généralement utilisée en pratique. Les propriétés des estimations du taux d'erreur de classification fondées sur un modèle, obtenues au moyen de trois estimateurs de paramètre de modèle, sont comparées.

MOTS CLÉS: Modèle mixte; modèle à variable latente; pondération itérative.

## 1. INTRODUCTION

Dans de nombreuses applications statistiques, on utilise des fichiers informatiques renfermant des renseignements sur des particuliers, sur des entreprises et sur des logements. Il faut souvent effectuer le couplage d'enregistrements qui se rapportent à la même entité. L'opération,qui consiste à coupler des enregistrements portant sur la même entité s'appelle appariement exact. Si l'on a attribué un identificateur unique à tous les enregistrements utilisés dans une application, l'appariement exact ne pose aucun problème. Les méthodes de couplage d'enregistrements s'attaquent au problème de l'appariement exact quand on ne dispose pas d'un identificateur unique. Dans ce cas, chaque enregistrement inclut généralement un certain nombre de champs de données renfermant des renseignements d'identification qui pourraient être utilisés pour effectuer l'appariement. Les problèmes rencontrés au cours de l'appariement sont dus à des erreurs dans ces données ou au fait que la même valeur dans un champ particulier est valable pour plus d'une entité.

Les applications du couplage d'enregistrements comprennent l'élimination des doubles comptes dans des listes de logements ou d'entreprises obtenues de diverses sources afin de créer des bases de sondage. De plus, le couplage d'enregistrements est largement utilisé dans des applications portant sur la santé et l'épidémiologie. Le travail dans ce domaine comporte généralement l'appariement d'enregistrements renfermant des renseignements sur des particuliers dans des cohortes d'industries ou de professions avec des enregistrements renfermant des renseignements sur la maladie ou sur le décès de particuliers. Par exemple, dans Fair, Newcombe et Lalonde (1988), on traite des méthodes de couplage d'enregistrements utilisées dans des études de suivi portant sur des personnes exposées à la radiation.

Le problème du couplage d'enregistrements peut être formulé à l'aide de deux fichiers de données qui correspondent à deux populations. Chaque fichier peut renfermer soit de l'information sur toutes les entités dans la population correspondante, soit de l'information pour un échantillon aléatoire d'entités. Le fichier A contient $N_A$ enregistrements et le fichier B en renferme $N_B$. L'ensemble de paires d'enregistrements formées comme produit croisé de $A$ et de $B$, est représenté par $C = \{ (a,b); a \in A, b \in B \}$. $C$ contient $N = N_A \cdot N_B$ paires d'enregistrements. L'objectif du couplage d'enregistrements est de partager l'ensemble $C$ en deux ensembles disjoints, l'ensemble des concordances vraies, représenté par $M$ et l'ensemble des non-concordances vraies, $U$.

De nombreux travaux d'application sont basés sur le modèle théorique présenté par Fellegi et Sunter (1969). Pour chaque paire d'enregistrements, on prend une décision afin de déterminer si les enregistrements se rapportent ou non à la même entité, après avoir examiné les données enregistrées dans les fichiers A et B. Les décisions possibles sont les suivantes: lien ($A_1$), non-lien ($A_3$) et cas indéterminé ($A_2$). Il y a deux types d'erreurs. Premièrement, la décision $A_1$ peut être prise pour une paire d'enregistrements qui fait partie de $U$, l'ensemble des non-concordances vraies. Deuxièmement, la décision $A_3$ peut être prise pour une paire d'enregistrements qui fait partie de l'ensemble $M$, l'ensemble des concordances vraies. Des niveaux acceptables d'erreur de classification sont précisés avant le couplage des fichiers. Une paire d'enregistrements est classée comme un cas indéterminé si les données ne

[1] J.B. Armstrong et J.E. Mayda, Statistique Canada, Division des méthodes d'enquêtes-entreprises, 11-Immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario), K1A 0T6.

fournissent pas une preuve suffisante pour justifier la classification de la paire comme un lien ou un non-lien à des niveaux d'erreur inférieurs ou égaux aux niveaux précisés. Il faut estimer avec précision les taux d'erreur de classification associés à diverses règles de décision afin de déterminer une règle appropriée. Le taux d'erreur de classification pour les non-concordances vraies est $P(A_1 \mid U)$. Le taux d'erreur de classification pour les concordances vraies est $P(A_3 \mid M)$.

On peut obtenir des estimations des taux d'erreur de classification en choisissant un échantillon de paires d'enregistrements de l'ensemble $C$ et en déterminant manuellement l'état véritable vis-à-vis de la concordance des paires échantillonnées. Des applications de cette méthode sont décrites dans Bartlett et coll. (1993). L'échantillonnage peut être à la fois coûteux et peu facile à réaliser, particulièrement quand on doit effectuer le même couplage pour un certain nombre de paires de fichiers, chacune avec des caractéristiques légèrement différentes. Belin et Rubin (1991) décrivent une autre méthode d'estimation des taux d'erreur pour lesquelles il faut disposer de l'état véritable de la concordance pour des paires d'enregistrements dans une étude-pilote. Par opposition à la méthode d'échantillonnage simple, la méthode de Belin et Rubin fournit un modèle pour l'application de renseignements, obtenus à l'aide de l'étude-pilote, à des couplages plus importants dans lesquels des données semblables sont utilisées.

Le modèle de Fellegi-Sunter fournit une méthode pour calculer des estimations des taux d'erreur à l'aide d'estimations des probabilités qu'il y aura concordance entre des paires d'enregistrements pour diverses combinaisons de champs de données. Le calcul de ces estimations du taux d'erreur obtenues à l'aide de modèles est simple et la détermination manuelle de l'état de paires d'enregistrements par rapport à la concordance vraie n'est pas nécessaire. Cependant, ces estimations ont souvent des propriétés qui laissent à désirer dans des applications. Voir, par exemple, Belin (1990). Dans le présent article, on démontre que les propriétés des estimations du taux d'erreur obtenues à l'aide de modèles peuvent être améliorées en faisant une estimation soignée des probabilités d'accord.

Trois méthodes d'estimation qui peuvent être employées sont évaluées. Les façons de procéder décrites n'utilisent que les renseignements dans les fichiers A et B. Elles n'emploient pas de renseignements auxiliaires. Les estimations des taux d'erreur obtenues à l'aide de modèles pour chacune des méthodes évaluées sont comparées aux taux d'erreur réels à l'aide tant de données synthétiques qui incorporent des caractéristiques importantes de données tirées d'applications du couplage d'enregistrements dans le domaine de la santé que de renseignements obtenus à partir d'une application réelle du couplage d'enregistrements.

Voici la structure du présent article. La section 2 contient des détails sur la méthode d'estimation de l'erreur de classification à l'aide d'un modèle présentée par Fellegi et Sunter. Le modèle utilisé pour les probabilités d'accord qui forme la base de la discussion ultérieure des méthodes d'estimation est aussi précisé. Deux méthodes d'estimation qui sont fondées sur une hypothèse d'indépendance importante sont décrites dans la section 3. On traite d'une troisième méthode pour laquelle l'indépendance n'est pas exigée dans la section 4. Les résultats des comparaisons des trois méthodes à l'aide de données synthétiques sont présentés dans la section 5. Les résultats du travail d'évaluation effectué avec des renseignements tirés d'une application réelle sont décrits dans la section 6. La section 7 renferme des conclusions.

## 2. CONCEPTS THÉORIQUES

Nous résumons, dans la présente section, les aspects pertinents de la théorie du couplage d'enregistrements élaborée par Fellegi et Sunter (1969). Dans le modèle de Fellegi-Sunter, les estimations des taux d'erreur de classification sont calculées à l'aide d'estimations des probabilités d'accord pour diverses combinaisons de champs de données. Les applications de la théorie de Fellegi et Sunter supposent habituellement l'hypothèse que la probabilité qu'il y aura accord, pour une paire d'enregistrements, au niveau d'un champ de données particulier est indépendante des résultats des comparaisons pour d'autres champs. La théorie est néanmoins très flexible, pouvant tenir compte de tous les types de dépendance entre les résultats de comparaisons pour différents champs de données. Un paramétrage de la dépendance en fonction d'effets log-linéaires est présenté.

### 2.1 Estimation des taux d'erreur de classification fondée sur un modèle

Pour obtenir des renseignements portant sur la classification d'une paire d'enregistrements comme un lien ($A_1$), un non-lien ($A_3$) ou un cas indéterminé ($A_2$), on compare des champs de données renfermant des renseignements d'identification. Dans une application comportant des enregistrements relatifs à des personnes, on pourrait faire des comparaisons distinctes des noms de famille, des prénoms et des dates de naissance. Le résultat d'une comparaison est un code numérique représentant un énoncé du genre: "les noms concordent", "les noms ne concordent pas", "un nom manque dans un des fichiers ou dans les deux", "les noms concordent et ils sont tous deux Georges" ou "les noms ne concordent pas mais leurs deux premiers caractères concordent". Les codes de résultat utilisés dans les travaux d'application diffèrent selon les applications et selon les comparaisons dans la même application. Le plus petit nombre de codes de résultat qui peut être utilisé pour toute comparaison est deux, ce qui correspond à l'accord et au désaccord. Dans les applications, il faut habituellement disposer d'un code de résultat correspondant à "manquant dans un des deux fichiers ou dans les deux". Le résultat de l'accord peut être remplacé par un certain nombre de résultats ayant une valeur particulière (telle que "les noms concordent et ils sont tous deux Georges"). Certains désaccords peuvent être codés comme des accords partiels (tels que "les noms ne concordent pas mais leurs deux premiers caractères concordent").

Pour nos fins, nous ne tenons compte que des résultats de l'accord et du désaccord. Dans le cas de $K$ champs d'appariement, nous définissons le vecteur de résultats $\underline{x}^j = (x_1^j, x_2^j, \ldots, x_K^j)$ pour la paire d'enregistrements $j$. Nous avons $x_k^j = 1$ s'il y a concordance pour la paire d'enregistrements $j$ à propos du champ de données $k$ et $x_k^j = 0$ s'il y a non-concordance pour la paire d'enregistrements $j$ à propos du champ de données $k$.

Newcombe et coll. (1959) ont proposé l'idée que les décisions relatives au fait qu'une paire d'enregistrements représente ou non la même entité devraient être basées sur le rapport

$$R(\underline{x}) = P(\underline{x} \mid M)/P(\underline{x} \mid U), \qquad (1)$$

où $\underline{x} = (x_1, x_2, \ldots, x_K)$ est le vecteur de résultats générique, $P(\underline{x} \mid M)$ est la probabilité que les comparaisons pour une paire d'enregistrements où il y a concordance vraie produiront le vecteur de résultats $\underline{x}$, et $P(\underline{x} \mid U)$ est la probabilité de $\underline{x}$ pour une paire d'enregistrements pour laquelle il y a non-concordance vraie. L'optimalité des méthodes de couplage d'enregistrements qui font appel à ce rapport a été démontrée par Fellegi et Sunter.

Dans le modèle de Fellegi-Sunter, une règle d'appariement attribue une probabilité pour chaque décision de classification ($A_1$, $A_2$ et $A_3$) à chaque vecteur de résultats. La fonction de décision correspondant au vecteur de résultats $\underline{x}$ est $d(\underline{x}) = (P(A_1 \mid \underline{x}), P(A_2 \mid \underline{x}), P(A_3 \mid \underline{x}))$. Des taux d'erreur de classification acceptables pour les non-concordances vraies et les concordances vraies sont précisés avant que le couplage ne soit effectué. Nous désignons ces taux d'erreur précisés à l'avance par $\mu$ et $\lambda$ respectivement. Fellegi et Sunter définissent, parmi la classe de règles d'appariement d'enregistrements qui satisfont aux relations $P(A_1 \mid U) \leq \mu$ et $P(A_3 \mid M) \leq \lambda$ pour des valeurs fixes de $\mu$ et de $\lambda$, la règle d'appariement optimale comme étant la règle qui minimise $P(A_2)$, la probabilité qu'une paire d'enregistrements sera classée comme un cas indéterminé. La règle optimale a la forme

$$\begin{aligned}
d(\underline{x}^j) &= (1,0,0) \quad \text{si} \quad \omega^j > \tau_1 \\
d(\underline{x}^j) &= (P_\mu, 1 - P_\mu, 0) \quad \text{si} \quad \omega^j = \tau_1 \\
d(\underline{x}^j) &= (0,1,0) \quad \text{si} \quad \tau_2 < \omega^j < \tau_1 \qquad (2) \\
d(\underline{x}^j) &= (0, 1 - P_\lambda, P_\lambda) \quad \text{si} \quad \omega^j = \tau_2 \\
d(\underline{x}^j) &= (0,0,1) \quad \text{si} \quad \omega^j < \tau_2
\end{aligned}$$

où $\tau_1 \geq \tau_2$, le "poids" $\omega^j$ est défini comme $\omega^j = \log(R(\underline{x}^j))$ et $P_\mu$ et $P_\lambda$ sont des constantes positives sur l'intervalle $[0,1)$. (Pour tous les détails, consulter Fellegi et Sunter (1969)). Pour déterminer $\tau_1$ et $\tau_2$, il faut estimer les taux d'erreur de classification correspondant à divers choix pour ces valeurs seuil, ce qui souligne l'importance d'une estimation précise des taux d'erreur de classification dans le modèle de Fellegi-Sunter.

On peut calculer des estimations des taux d'erreur de classification fondées sur un modèle à l'aide d'estimations des probabilités de résultats pour les concordances vraies et pour les non-concordances vraies. Utilisons $\hat{P}(\underline{x} \mid M)$ et $\hat{P}(\underline{x} \mid U)$ pour représenter les estimations des probabilités du vecteur de résultats $\underline{x}$ pour les concordances vraies et pour les non-concordances vraies et représentons le rapport de ces estimations par $\hat{R}(\underline{x})$. L'estimation du taux d'erreur de classification pour les concordances vraies fondée sur un modèle basé sur la règle de décision (2) est

$$\hat{\lambda} = \sum_{\underline{x} \in L(\tau_2)} \hat{P}(\underline{x} \mid M) + P_\lambda \sum_{\underline{x} \in Q(\tau_2)} \hat{P}(\underline{x} \mid M) \qquad (3)$$

où $L(\tau_2) = \{\underline{x};\ \log(\hat{R}(\underline{x})) < \tau_2\}$ et $Q(\tau_2) = \{\underline{x};\ \log(\hat{R}(\underline{x})) = \tau_2\}$.

L'estimation du taux d'erreur de classification pour les non-concordances vraies fondée sur un modèle est

$$\hat{\mu} = \sum_{\underline{x} \in G(\tau_1)} \hat{P}(\underline{x} \mid U) + P_\mu \sum_{\underline{x} \in Q(\tau_1)} \hat{P}(\underline{x} \mid U) \qquad (4)$$

où $G(\tau_1) = \{\underline{x};\ \log(\hat{R}(\underline{x})) > \tau_1\}$ et $Q(\tau_1) = \{\underline{x};\ \log(\hat{R}(\underline{x})) = \tau_1\}$.

### 2.2 Un modèle pour les probabilités de résultats

Pour calculer les estimations du taux d'erreur de classification fondées sur un modèle, il faut estimer $P(\underline{x} \mid M)$ et $P(\underline{x} \mid U)$, pour chacune des $2^K$ valeurs possibles de $\underline{x}$. La densité de probabilité pour $\underline{x}$ est une combinaison de deux densités de probabilité données par

$$f(\underline{x}) = pP(\underline{x} \mid M) + (1 - p)\, P(\underline{x} \mid U), \qquad (5)$$

où $p$ est la probabilité qu'il y a concordance vraie pour une paire d'enregistrements choisie au hasard. Les probabilités de résultats dépendent de la distribution de fréquences des identificateurs pour des entités représentées dans les fichiers A et B, ainsi que des probabilités qu'il y a introduction d'erreurs quand les identificateurs sont enregistrés dans les fichiers. Fellegi et Sunter (1969, pp. 1192-1194) décrivent, pour estimer les probabilités d'accord, une méthode qui fait appel à leur définition en fonction des distributions de fréquences et des probabilités d'erreur. Ils recommandent d'utiliser la méthode quand on dispose d'informations préalables.

Dans le présent article, nous considérons des situations où les données dans les fichiers A et B, ainsi que les vecteurs de résultats $\underline{x}^j$, $j = 1, 2, \ldots, N$, représentent les seuls renseignements disponibles pour l'estimation des probabilités de résultats. Une structure log-linéaire pour les probabilités de résultats constitue le paramétrage le plus général. Le modèle log-linéaire saturé pour les probabilités de résultats dans le cas des concordances vraies est

$$\log(P(\underline{x} \mid M)) = M(0) + M(1)_{x_1} + M(2)_{x_2} + \ldots$$
$$+ M(K)_{x_K} + M(1)\, M(2)_{x_1,x_2} + \ldots$$
$$+ M(K-1)\, M(K)_{x_{K-1},x_K} + \ldots$$
$$+ M(1)\, M(2) \ldots M(K)_{x_1,x_2,\ldots,x_K}, \quad (6)$$

avec les restrictions habituelles

$$\sum_{x_J} M(J)_{x_J} = 0, \quad J = 1, 2, \ldots, K,$$

$$\sum_{x_{J_1}} M(J_1)M(J_2)_{x_{J_1},x_{J_2}} = \sum_{x_{J_2}} M(J_1)M(J_2)_{x_{J_1},x_{J_2}} = 0,$$
$$\forall\, J_1, J_2, \quad etc.,$$

ainsi que la restriction

$$\sum_{\underline{x}} P(\underline{x} \mid M) = 1.$$

Le modèle saturé pour $P(\underline{x} \mid U)$ est analogue.

Si l'on utilise des modèles log-linéaires saturés pour $P(\underline{x} \mid M)$ et pour $P(\underline{x} \mid U)$, la fonction de densité comprend $2^{K+1} - 1$ paramètres inconnus. On ne peut identifier tous ces paramètres quand on ne dispose pas de renseignements auxiliaires. Afin d'obtenir un modèle qui peut être identifié et pour simplifier le problème d'estimation, on fait souvent l'hypothèse que les résultats des comparaisons pour différents champs de données sont indépendants. Quand on suppose qu'il y a indépendance, on désigne les probabilités d'accord parmi les paires d'enregistrements qui sont des concordances vraies et des non-concordances vraies, respectivement, par

$$m_k = P(x_k = 1 \mid M), \quad k = 1, 2, \ldots, K,$$

$$u_k = P(x_k = 1 \mid U), \quad k = 1, 2, \ldots, K.$$

Les probabilités de résultats peuvent être écrites sous la forme:

$$P(\underline{x} \mid M) = \prod_{k=1}^{K} m_k^{x_k} (1 - m_k)^{(1-x_k)},$$

$$P(\underline{x} \mid U) = \prod_{k=1}^{K} u_k^{x_k} (1 - u_k)^{1-x_k}.$$

Ce modèle comprend $2 \cdot K + 1$ paramètres inconnus, nommément $(\underline{m}, \underline{u}, p)$, où $\underline{m} = (m_1, m_2, \ldots, m_k)$, $\underline{u} = (u_1, u_2, \ldots, u_k)$. Il y a, bien entendu, un certain nombre de modèles intermédiaires entre le modèle saturé et le modèle où l'on fait appel à l'indépendance. Des méthodes qui peuvent être utilisées pour estimer le modèle qui fait appel à l'indépendance sont décrites dans la section 3. On traite de l'estimation de modèles intermédiaires dans la section 4.

## 3. ESTIMATION FONDÉE SUR L'HYPOTHÈSE DE L'INDÉPENDANCE

### 3.1 Méthode des moments

On peut employer un estimateur de $P(\underline{x} \mid M)$ et de $P(\underline{x} \mid U)$ obtenu à l'aide de la méthode des moments quand il y a indépendance. L'estimateur est basé sur un système de $2 \cdot K + 1$ équations qui fournissent des expressions pour des moments fonctionnellement indépendants de $\underline{x}$ en fonction des paramètres. Les équations sont les suivantes:

$$E\left(\prod_{k \neq i}^{K} x_k\right) = pN \prod_{k \neq i}^{K} m_k + (1 - p)\, N \prod_{k \neq i}^{K} u_k,$$
$$i = 1, 2, \ldots, K$$

$$E(x_i) = pNm_i + (1 - p)\, Nu_i, \quad i = 1, 2, \ldots, K, \quad (7)$$

$$E\left(\prod_{k=1}^{K} x_k\right) = pN \prod_{k=1}^{K} m_k + (1 - p)\, N \prod_{k=1}^{K} u_k.$$

Pour obtenir des estimations des paramètres à l'aide de la méthode des moments, il faut résoudre les équations une fois que les valeurs espérées ont été remplacées par des moyennes calculées à l'aide de paires d'enregistrements dans $C$. Le système d'équations pour $K = 3$ a été présenté par Fellegi et Sunter, qui ont aussi calculé une solution en forme analytique fermée qui existe si certaines conditions modérées sont satisfaites. Leur article comprenait une mise en garde pour ce qui est de l'utilisation de la méthode s'il n'y a pas indépendance. Pour $K > 3$, on ne dispose pas d'une solution en forme analytique fermée, mais on peut employer les méthodes numériques courantes. Les estimations de paramètres obtenues à l'aide de la méthode des moments sont statistiquement convergentes si l'hypothèse d'indépendance est vérifiée.

### 3.2 Méthode itérative

La méthode itérative a été élaborée par des personnes qui effectuent des couplages d'enregistrements. Bien que la méthode ne soit pas basée sur la distribution de probabilité du vecteur de résultats, elle fait appel à l'hypothèse d'indépendance. L'application de la méthode itérative est décrite par plusieurs auteurs, y compris Newcombe (1988). Le logiciel de couplage d'enregistrements de Statistique Canada, CANLINK, est conçu de façon à faciliter l'utilisation de la méthode itérative.

La méthode exige des estimations initiales des probabilités d'accord pour les concordances vraies et pour les non-concordances vraies. Pour les concordances vraies, on doit utiliser des estimations au jugé basées sur l'expérience acquise. Pour obtenir des estimations initiales des probabilités d'accord parmi les paires d'enregistrements qui correspondent à des non-concordances vraies, on suppose habituellement que ces probabilités sont égales aux probabilités d'accord parmi les paires d'enregistrements choisies de façon aléatoire, nommément que:

$$u_k = P(x_k = 1), \quad k = 1, 2, \ldots, K.$$

Supposons que $J(k)$ valeurs différentes apparaissent, pour le champ de données $k$, dans le fichier A et/ou dans le fichier B. Désignons les fréquences de ces valeurs dans le fichier A par $f_{k1}, f_{k2}, \ldots, f_{kJ(k)}$ et désignons les fréquences pour le fichier B par $g_{k1}, g_{k2}, \ldots, g_{kJ(k)}$. Pour une valeur particulière, un des chiffres, mais non les deux, peut être nul. L'estimation initiale de $u_k$ est

$$\hat{u}_k^0 = \sum_{j=1}^{J(k)} (f_{kj}\, g_{kj})/N. \tag{8}$$

Compte tenu de ces estimations de probabilités, des ensembles initiaux de concordances et de non-concordances, désignés par $M^0$ et par $U^0$, respectivement, sont obtenus à l'aide d'une règle de décision

$$j \in M^0 \quad \text{si} \quad \omega^j > \tau_1^0,$$

$$j \in U^0 \quad \text{si} \quad \omega^j < \tau_2^0.$$

Puis, on utilise les chiffres des fréquences parmi les paires d'enregistrements dans les ensembles $M^0$ et $U^0$ comme nouvelles estimations des probabilités d'accord. Ces estimations sont employées pour obtenir de nouveaux ensembles de concordances et de non-concordances et le processus itératif est repris jusqu'à ce que des estimations consécutives des probabilités d'accord soient suffisamment rapprochées.

Dans la majorité des applications, l'hypothèse que la probabilité d'accord parmi les paires d'enregistrements qui correspondent à des concordances vraies est égale à la probabilité d'accord parmi toutes les paires d'enregistrements est justifiée et l'itération ne mène pas à des modifications importantes dans les estimations des probabilités d'accord pour les non-concordances vraies. Toutefois, il arrive souvent que la première itération produise des variations considérables dans les estimations de la probabilité d'accord pour les concordances vraies. Généralement, il n'y a pas de variations importantes lors de la deuxième itération.

Il faudrait remarquer que les propriétés statistiques de la méthode itérative ne sont pas connues avec précision. En pratique, le rendement de la méthode dépendra du choix des seuils initiaux $\tau_1^0$, $\tau_2^0$. Ces seuils sont généralement choisis de façon subjective. Les simulations mentionnées dans la section 5 fournissent des renseignements à propos des effets des divers seuils initiaux.

## 4. ASSOUPLISSEMENT DE L'HYPOTHÈSE DE L'INDÉPENDANCE - ESTIMATION À L'AIDE DE LA PONDÉRATION ITÉRATIVE

On peut utiliser des méthodes d'estimation pour les modèles à variable latente afin d'estimer les probabilités d'accord quand on effectue le paramétrage, en fonction d'effets log-linéaires, de la dépendance entre les résultats de comparaisons pour différents champs d'appariement. Winkler (1989) et Thibaudeau (1989) ont estimé les probabilités d'accord avec des modèles log-linéaires qui comprennent tous les termes d'interaction jusqu'au troisième ou quatrième ordre afin de paramétrer les dépendances. La formulation présentée ici facilite l'utilisation de modèles log-linéaires, y compris certaines interactions. On peut considérer l'état par rapport à la concordance comme une variable latente avec deux niveaux (concordance vraie et non-concordance vraie). Représentons par $c_{0,\underline{x}}$ et $c_{1,\underline{x}}$, le nombre de non-concordances vraies et de concordances vraies, respectivement, avec vecteur de résultats $\underline{x}$ dans une application de couplage d'enregistrements pour laquelle on utilise $K$ variables d'appariement. Bien entendu, ces chiffres ne peuvent être observés puisque la valeur de la variable latente pour chaque paire d'enregistrements est inconnue. On observe plutôt $c_{\underline{x}} = c_{0,\underline{x}} + c_{1,\underline{x}}$.

À l'aide du paramétrage de la dépendance réalisé au moyen d'effets log-linéaires et d'un modèle saturé pour les concordances vraies, nous pouvons poser

$$\begin{aligned}\log(c_{1,\underline{x}}/(Np)) = {} & M(0) + M(1)_{x_1} + M(2)_{x_2} + \ldots \\ & + M(K)_{x_K} + M(1)M(2)_{x_1,x_2} + \ldots \\ & + M(K-1)M(K)_{x_{K-1},x_K} + \ldots \\ & + M(1)M(2) \ldots M(K)_{x_1,x_2,\ldots,x_K},\end{aligned}$$

avec les restrictions habituelles. Nous disposons d'une expression semblable pour les non-concordances vraies. Le modèle à variable latente correspondant à ces modèles log-linéaires saturés est:

$$\begin{aligned}\log(c_{s,\underline{x}}/w_s) = {} & G(0) + Z_s + G(1)_{x_1} + \ldots \\ & + G(K)_{x_K} + ZG(1)_{s,x_1} + \ldots + ZG(K)_{s,x_K} \\ & + \ldots + G(1)G(2)\ldots G(K)_{x_1,x_2,\ldots,x_K} \\ & + ZG(1)G(2) \ldots G(K)_{s,x_1,x_2,\ldots x_K},\end{aligned}$$

où la valeur de l'indice $s$ est zéro pour les non-concordances vraies et un pour les concordances vraies,

$w_0 = (1 - p)N$ et $w_1 = pN$. Les paramètres sont analogues aux paramètres d'un modèle log-linéaire saturé pour un tableau de contingence de dimensions $2^{K+1}$. Les restrictions habituelles s'appliquent. Par exemple, le terme $ZG(1)_{s,x_1}$ représente l'interaction de la variable latente et de la première variable d'appariement et

$$\sum_s ZG(1)_{s,x_1} = \sum_{x_1} ZG(1)_{s,x_1} = 0.$$

Ce modèle est conforme au modèle à variable latente général de Haberman (1979, p. 561). On doit imposer des restrictions additionnelles pour identifier et estimer les paramètres. Pour plus de simplicité, nous ne considérerons que les modèles hiérarchiques. De plus, nous n'étudierons que des modèles qui permettent à tous les effets non nuls d'interagir avec la variable latente.

Dans la discussion qui suit, nous désignerons les modèles à variable latente par les symboles $G(1)$, $G(2)$, ... les modèles log-linéaires pour les concordances vraies par $M(1), M(2)$, ... et les modèles log-linéaires pour les non-concordances vraies par $U(1)$, $U(2)$, ... . Dans le cas de quatre variables d'appariement, par exemple, le modèle $G(1)G(2)$, $G(3)$, $G(4)$ est un modèle à variable latente qui comprend un terme de niveau général, des effets principaux pour les quatre variables d'appariement et un terme pour l'interaction des variables d'appariement un et deux, ainsi qu'un terme pour les effets principaux pour la variable latente (l'interaction du terme de niveau général et de la variable latente), des termes pour l'interaction de chaque variable d'appariement et de la variable latente ainsi qu'un terme pour l'interaction des variables d'appariement un et deux et de la variable latente. Le modèle comprend douze paramètres qui doivent être estimés. Le nombre de paramètres qui doivent être estimés dans un des modèles à variable latente examinés dans le présent document est deux fois plus élevé que le nombre de paramètres dans le modèle log-linéaire correspondant.

On peut utiliser la méthode de pondération itérative de Haberman (1976) pour estimer des modèles à variable latente. La méthode d'estimation de Haberman consiste à soumettre à un balayage des tableaux qui renferment des chiffres estimés pour chaque résultat parmi les concordance vraies et les non-concordances vraies. Représentons les chiffres estimés pour le vecteur de résultats $\underline{x}$ après $i$ itérations de l'algorithme de Haberman par $\hat{C}^i_{1,\underline{x}}$ et par $\hat{C}^i_{0,\underline{x}}$ pour les concordances vraies et les non-concordances vraies, respectivement. On peut construire les valeurs initiales $\hat{C}^0_{1,\underline{x}}$ et $\hat{C}^0_{0,\underline{x}}$ à l'aide d'estimations des probabilités d'accord et de la proportion de concordances vraies obtenues quand on applique l'hypothèse de l'indépendance. Chaque itération de l'algorithme comporte une série d'opérations de balayage appliquées au tableau courant des concordances vraies ainsi que les opérations analogues sur le tableau courant des non-concordances vraies. À l'aide de la notation pour les modèles hiérarchiques présentée plus haut, on effectue un ensemble d'opérations de balayage pour chacun des termes d'interaction qui définissent le modèle. Pour quatre variables d'appariement ainsi que le modèle $G(1)G(2)$, $G(3)G(4)$, on effectue deux ensembles d'opérations de balayage, une pour l'interaction $G(1)G(2)$ et la seconde pour l'interaction $G(3)G(4)$. Pour chaque interaction, une opération de balayage est effectuée pour chaque niveau de la variable de classification correspondante. Représentons par $S_{gl}$ l'ensemble de vecteurs de résultats au niveau $l$ du terme $g$. L'opération de balayage appliquée au tableau des concordances vraies lors de l'itération $i$ pour le niveau $l$ du terme $g$ comporte le calcul de

$$\gamma_{1,\underline{x}} = c_{\underline{x}}\hat{c}^{i-1}_{1,\underline{x}} / (\hat{c}^{i-1}_{1,\underline{x}} + \hat{c}^{i-1}_{0,\underline{x}}),$$

$$\hat{c}^i_{1,\underline{x}} = \hat{c}^{i-1}_{1,\underline{x}} \sum_{\underline{x} \in s_{gl}} \gamma_{1,\underline{x}} \Big/ \sum_{\underline{x} \in s_{gl}} \hat{c}^{i-1}_{1,\underline{x}}, \quad \forall\, \underline{x} \in S_{gl}.$$

Le traitement de l'algorithme prend fin quand les variations entre les chiffres estimés pour des itérations consécutives sont inférieures à une tolérance donnée.

Haberman (1976) fait remarquer qu'il se peut que l'algorithme de pondération itérative converge vers un maximum local de la fonction de vraisemblance plutôt que vers l'estimation obtenue à l'aide de la méthode du maximum de vraisemblance. Des expériences réalisées avec des valeurs initiales différentes qui utilisent des ensembles de données employés dans l'évaluation dont on traite à la section 5 n'ont permis d'obtenir aucun exemple de ce problème.

## 5. COMPARAISON DES MÉTHODES D'ESTIMATION - DONNÉES SYNTHÉTIQUES

On présente, dans cette section, les résultats de comparaisons des méthodes d'estimation décrites dans les sections 3 et 4. Les comparaisons comprenaient l'application de chaque méthode à une série d'ensembles de données synthétiques produits à l'aide de méthodes de Monte Carlo.

On a employé des enregistrements de données synthétiques renfermant quatre identificateurs de personne (nom de famille, initiale d'un deuxième prénom, prénom, date de naissance). Les renseignements sur les valeurs possibles de chaque identificateur ainsi que leurs fréquences relatives, ont été tirés de la Base canadienne de données sur la mortalité pour 1988. Cette base de données, qui est souvent utilisée dans des applications de couplage d'enregistrements dans le domaine de la santé, renferme un enregistrement distinct pour chaque mortalité.

L'hypothèse d'indépendance n'était pas respectée parmi les concordances vraies dans chaque ensemble de données synthétiques. Pendant la production des données, on a utilisé des renseignements, sur la fréquence des vecteurs de résultats pour les concordances vraies, obtenus à partir de divers projets de couplage d'enregistrements réalisés par le Centre canadien d'information sur la santé de Statistique Canada. La majorité des projets comportaient l'appariement d'un fichier de cohorte à la Base canadienne de données

sur la mortalité. La fréquence de chaque vecteur de résultats parmi les concordances vraies est présentée au tableau 1. L'absence d'indépendance dans ces données est évidente. Bien que pour environ 88.3% des concordances vraies, il y ait accord pour le prénom, la probabilité d'un accord pour le prénom quand il y a désaccord pour l'initiale d'un deuxième prénom et accord pour le nom de famille et l'année de naissance n'est que de 381/1366, soit environ 27.9%. La valeur de la statistique du test du rapport des vraisemblances pour l'hypothèse d'indépendance est 3604. Cette valeur est extrêmement élevée par rapport à la distribution de référence chi carré avec 10 degrés de liberté. (Il faut remarquer qu'un degré de liberté est perdu parce que la fréquence pour la case (1,0,0,0) est zéro.)

**Tableau 1**

Fréquences des résultats, ensemble de concordances vraies, données synthétiques

| Résultat selon l'identificateur: 0 = Désaccord, 1 = Accord | | | | Fréquence | |
|---|---|---|---|---|---|
| Prénom | Initiale d'un deuxième prénom | Nom de famille | Année de naissance | Chiffre | Pourcentage |
| 0 | 0 | 0 | 0 | 7 | 0.03 |
| 0 | 0 | 0 | 1 | 33 | 0.12 |
| 0 | 0 | 1 | 0 | 125 | 0.45 |
| 0 | 0 | 1 | 1 | 985 | 3.54 |
| 0 | 1 | 0 | 0 | 5 | 0.02 |
| 0 | 1 | 0 | 1 | 39 | 0.14 |
| 0 | 1 | 1 | 0 | 202 | 0.73 |
| 0 | 1 | 1 | 1 | 1,848 | 6.65 |
| 1 | 0 | 0 | 0 | 0 | 0.0 |
| 1 | 0 | 0 | 1 | 13 | 0.05 |
| 1 | 0 | 1 | 0 | 50 | 0.18 |
| 1 | 0 | 1 | 1 | 381 | 1.37 |
| 1 | 1 | 0 | 0 | 44 | 0.16 |
| 1 | 1 | 0 | 1 | 451 | 1.62 |
| 1 | 1 | 1 | 0 | 1,751 | 6.30 |
| 1 | 1 | 1 | 1 | 21,860 | 78.65 |
| | | | Total | 27,794 | 100 |

Pour chaque ensemble de données synthétiques, on a produit les enregistrements du fichier A en choisissant des identificateurs selon les fréquences relatives dans la Base canadienne de données sur la mortalité pour 1988. Afin de simplifier le processus de production des données, le choix des noms de famille était limité aux 100 noms de famille non francophones les plus courants et aux 100 noms de famille francophones les plus courants qui figuraient dans le fichier de 1988. Le choix des prénoms était limité aux 50 prénoms francophones les plus courants et aux 50 prénoms non francophones les plus courants. On n'a pas tenu compte des variantes orthographiques lors du choix des noms. On a toutefois pris en considération toutes les initiales d'un deuxième prénom et de toutes les années de naissance qui figuraient dans le fichier pour 1988. La probabilité que des enregistrements avec des prénoms anglophones se voient attribuer un nom de famille anglophone était plus élevée que dans le cas des prénoms francophones (ce qui reflète la répartition des noms dans la population canadienne). À part ces restrictions, les identificateurs étaient choisis indépendamment.

Le point de départ pour le fichier B était une copie exacte du fichier A. Pour chaque enregistrement du fichier B, il y avait une concordance vraie avec exactement un enregistrement du fichier A. Pour introduire une absence d'indépendance parmi les concordances vraies, on a tiré un vecteur de résultats de la distribution de fréquences du tableau 1 pour chaque enregistrement du fichier B. Les identificateurs correspondant à des zéros dans le vecteur de résultats ont été choisis à nouveau. Par conséquent, l'ensemble de vecteurs de résultats pour les concordances vraies était un échantillon de la distribution du tableau 1. Les ensembles de données synthétiques comprenaient aussi de légers écarts par rapport à l'hypothèse d'indépendance pour les non-concordances vraies puisque la sélection des prénoms et des noms de famille n'était pas complètement indépendante.

Chaque ensemble de résultats de simulations mentionné plus loin est basé sur 50 essais de Monte Carlo. Chaque essai comportait la production de fichiers A et B comprenant 500 personnes, l'estimation de $\underline{m}$ et de $\underline{u}$, la détermination de seuils correspondant à diverses estimations du taux d'erreur de classification fondées sur un modèle et le calcul de taux d'erreur réels correspondant aux seuils. La même série de 50 ensembles de données synthétiques était utilisée pour chaque ensemble de simulations. Il faut remarquer que l'ensemble C renferme 250,000 paires d'enregistrements, y compris 249,500 non-concordances vraies pour chaque essai de Monte Carlo. Afin de réduire le temps de calcul nécessaire pour effectuer les simulations, on n'a utilisé que 49,500 non-concordances vraies pour chaque essai. (On a effectué un essai à petite échelle afin de s'assurer que la réduction du nombre de non-concordances vraies avait un effet négligeable sur les probabilités d'accord estimées.) On a supprimé les non-concordances vraies contenues dans C en divisant les fichiers A et B en cinq blocs correspondants de taille 100 et en excluant les paires d'enregistrements dans lesquelles on trouvait des enregistrements provenant de blocs qui ne correspondaient pas.

Le système d'équations utilisé pour la méthode des moments a été résolu à l'aide d'une variation de la méthode de Newton, décrite en détail dans Moré et coll. (1980). Un logiciel fourni par IMSL (1987) a été utilisé. On a employé des probabilités d'accord de 0.9 pour les concordances vraies et de 0.1 pour les non-concordances vraies, pour tous les champs d'appariement, comme valeurs initiales pour la solution du système d'équations. La méthode ne semblait pas sensible aux valeurs initiales.

**Tableau 2**

Taux d'erreur de classification, non-concordances vraies, données synthétiques (erreurs-types Monte Carlo entre parenthèses)

| Taux estimé (× 99) | Taux réel (× 99) | | | | |
|---|---|---|---|---|---|
| | Méthode des moments | Méthode itérative $\mu^0 = 0.0000625$ | Méthode itérative $\mu^0 = 0.00025$ | Méthode itérative $\mu^0 = 0.001$ | Pondération itérative |
| 0.02 | 0.0188 (0.0008) | 0.0208 (0.0008) | 0.0208 (0.001) | 0.0207 (0.001) | 0.0195 (0.001) |
| 0.04 | 0.0381 (0.001) | 0.0408 (0.0013) | 0.0407 (0.0016) | 0.0405 (0.0016) | 0.0397 (0.0016) |
| 0.06 | 0.057 (0.0012) | 0.0626 (0.0015) | 0.0615 (0.0018) | 0.0602 (0.0019) | 0.059 (0.0018) |
| 0.08 | 0.076 (0.0015) | 0.0855 (0.0017) | 0.0838 (0.0019) | 0.0804 (0.0022) | 0.0785 (0.0019) |
| 0.10 | 0.095 (0.0019) | 0.1086 (0.0021) | 0.1061 (0.0022) | 0.1007 (0.0026) | 0.0978 (0.0021) |

Les propriétés de la méthode itérative dépendent des définitions des ensembles initiaux de concordances et de non-concordances, $M^0$ et $U^0$. Il faut se rappeler que, compte tenu de probabilités initiales, les paires d'enregistrements sont classées selon la règle suivante:

$$j \in M^0 \quad \text{si} \quad \omega^j > \tau_1^0,$$

$$j \in U^0 \quad \text{si} \quad \omega^j < \tau_2^0.$$

Quand la méthode itérative a été appliquée pour les simulations mentionnées ici, on a donné à $\tau_2^0$ la valeur de $\tau_1^0$. Pour chaque essai de Monte Carlo, $\tau_1^0$ a été fixé de façon à ce que

$$\hat{P}(j \in U \mid \omega^j > \tau_1^0) + \gamma \cdot \hat{P}(j \in U \mid \omega^j = \tau_1^0) = \mu^0,$$

pour un $\gamma \in [0,1)$, où les probabilités estimées sont basées sur les estimations itératives initiales de $\underline{u}$. Les paires d'enregistrements avec poids $\tau_1^0$ ont été classifiés dans $M^0$ avec probabilité $\gamma$. C'est-à-dire que l'ensemble initial de concordances utilisé par la méthode itérative était défini de façon à ce que le taux de fausses concordances estimé correspondant soit $\mu^0$. Les valeurs initiales pour $m_k$, $k = 1, 2, \ldots, 4$, ont été fixées à 0.9.

Le chiffre zéro dans le tableau 1 (accord pour le prénom, désaccord pour tous les autres identificateurs) était traité comme un zéro structurel pendant la production des données. Parmi les modèles log-linéaires pour lesquels on n'utilisait pas plus de six paramètres, c'est le modèle $M(1)M(2), M(3), M(4)$ qui donne le meilleur ajustement avec les données du tableau 1. Ce modèle, dans lequel on utilise la dépendance pour les résultats de comparaisons pour le prénom et l'initiale d'un deuxième prénom, ne donne pas un très bon ajustement. La statistique du test du rapport des vraisemblances pour le manque d'ajustement est 57.95, une valeur extrême par rapport à la distribution de référence chi carré avec 9 degrés de liberté. Le modèle à variable latente $G(1)G(2), G(3), G(4)$ était estimé, à l'aide de la pondération itérative, pour chaque ensemble de données synthétiques. L'ajustement de ce modèle avec les ensembles de données synthétiques était légèrement meilleur que l'ajustement du modèle $M(1)M(2), M(3), M(4)$ avec les données sur les concordances vraies. La plus élevée des statistiques du test du manque d'ajustement parmi les cinquante ensembles de données synthétiques était 25.03 et le modèle n'était rejeté que dix fois au niveau de signification 5%.

Les moyennes des estimations du taux d'erreur de classification obtenues à l'aide des ensembles de données synthétiques ainsi que les erreurs-types de Monte Carlo correspondantes sont présentées au tableau 2 pour les non-concordances vraies et au tableau 3 pour les concordances vraies. Après multiplication par 99, les taux d'erreur pour les non-concordances vraies représentent le nombre de non-concordances vraies mal classées divisé par le nombre de concordances vraies. Les résultats sont présentés pour la méthode des moments et pour la pondération itérative, ainsi que pour la méthode itérative avec $\mu^0 = 0.0000625$, 0.00025 et 0.001. Les biais dans les taux d'erreur estimés pour les non-concordances vraies sont généralement faibles. La méthode itérative avec $\mu^0 = 0.001$ fournit les meilleures estimations, vient ensuite la pondération itérative. Pour les concordances vraies, le rendement de la méthode itérative dépend beaucoup du choix de $\mu^0$. Bien que la méthode itérative donne de bons résultats pour $\mu^0 = 0.001$, les biais pour $\mu^0 = 0.0000625$ et $\mu^0 = 0.00025$ sont considérables. Les estimations du taux d'erreur de classification pour les concordances vraies obtenues à l'aide de la méthode des moments comprennent aussi des biais importants. Les biais dans les estimations obtenues par pondération itérative sont relativement faibles.

**Tableau 3**

Taux d'erreur de classification, concordances vraies, données synthétiques
(erreur-types Monte Carlo entre parenthèses)

| Taux estimé | Taux réel | | | | |
|---|---|---|---|---|---|
| | Méthode des moments | Méthode itérative $\mu^0 = 0.0000625$ | Méthode itérative $\mu^0 = 0.00025$ | Méthode itérative $\mu^0 = 0.001$ | Pondération itérative |
| 0.02 | 0.0580 (0.0013) | 0.1179 (0.0041) | 0.0507 (0.0014) | 0.0149 (0.0008) | 0.025 (0.0012) |
| 0.04 | 0.0773 (0.0014) | 0.1362 (0.004) | 0.0735 (0.0012) | 0.0359 (0.0018) | 0.0455 (0.0016) |
| 0.06 | 0.0966 (0.0014) | 0.1542 (0.0038) | 0.0954 (0.0012) | 0.0660 (0.0014) | 0.0646 (0.0018) |
| 0.08 | 0.1159 (0.0014) | 0.1722 (0.0036) | 0.1165 (0.0012) | 0.0866 (0.0017) | 0.0841 (0.0019) |
| 0.10 | 0.1348 (0.0014) | 0.1904 (0.0035) | 0.1319 (0.0014) | 0.1025 (0.002) | 0.1043 (0.002) |

**Tableau 4**

Taux d'erreur de classification, non-concordances vraies, données synthétiques modifiées
(erreur-types Monte Carlo entre parenthèses)

| Taux estimé (× 99) | Taux réel (× 99) | |
|---|---|---|
| | Méthode des moments | Pondération itérative |
| 0.02 | 0.0189 (0.0008) | 0.0194 (0.001) |
| 0.04 | 0.0385 (0.0011) | 0.0396 (0.0016) |
| 0.06 | 0.0577 (0.0013) | 0.0589 (0.0019) |
| 0.08 | 0.0767 (0.0016) | 0.0785 (0.002) |
| 0.10 | 0.0957 (0.002) | 0.0978 (0.0021) |

**Tableau 5**

Taux d'erreur de classification, concordances vraies, données synthétiques modifiées
(erreurs-types Monte Carlo entre parenthèses)

| Taux estimé | Taux réel | |
|---|---|---|
| | Méthode des moments | Pondération itérative |
| 0.02 | 0.0553 (0.0014) | 0.0208 (0.0011) |
| 0.04 | 0.0747 (0.0014) | 0.0415 (0.0016) |
| 0.06 | 0.094 (0.0014) | 0.0608 (0.0018) |
| 0.08 | 0.1134 (0.0014) | 0.0805 (0.002) |
| 0.10 | 0.1325 (0.0015) | 0.1007 (0.002) |

Les renseignements présentés dans les tableaux 4 et 5 sont basés sur une série d'ensembles de données synthétiques produits à l'aide d'une version modifiée du tableau 1. Les valeurs probables des chiffres dans les cases du tableau 1 selon le modèle $M(1)M(2), M(3), M(4)$ ont été utilisées pour produire les données. Les biais dans les estimations du taux d'erreur de classification fondées sur un modèle obtenues à l'aide de la méthode des moments sont fortement réduits quand on utilise le modèle à variable latente $G(1)G(2), G(3), G(4)$ estimé à l'aide de la pondération itérative, particulièrement dans le cas des concordances vraies.

## 6. COMPARAISON DES MÉTHODES D'ESTIMATION - DONNÉES RÉELLES

Les résultats des comparaisons des trois méthodes d'estimation effectuées à l'aide de données provenant d'une application de couplage d'enregistrements sont présentées dans cette section. Deux fichiers de données utilisés dans un travail empirique présenté par Fair et Lalonde (1987) ont été employés. Le premier fichier renfermait des renseignements sur les mineurs ontariens obtenus de la Commission des accidents du travail. Le deuxième fichier comprenait des renseignements tirés de la Base canadienne de données sur la mortalité (BCDM) pour les décès de

particuliers pendant la période allant de 1964 à 1977 inclusivement. Le fichier des mineurs n'incluait que les enregistrements avec un numéro d'assurance sociale valable. Le deuxième fichier renfermait des enregistrements qui avaient été retenus après une comparaison initiale visant à éliminer les enregistrements qui n'avaient aucune similitude avec un quelconque des enregistrements dans les fichiers des mineurs. Le statut vital de chaque mineur, à la fin de 1977, avait été classé comme "décès confirmé", "survie confirmée" ou "non retrouvé lors du suivi", basé sur un couplage antérieur, combiné avec des procédures de suivi complètes, y compris un examen manuel. Les enregistrements dans le fichier des mineurs, pour les personnes dont le statut vital est "décès confirmé", incluaient le numéro d'enregistrement du décès dans la BCDM. On peut trouver plus de renseignements sur la construction des fichiers et sur les procédures utilisées pour déterminer l'état véritable du couplage dans Fair et Lalonde.

Quatre identificateurs, le premier prénom, le code "NYSIIS" du nom de jeune fille de la mère, le jour de naissance et le mois de naissance, ont été choisis comme champs d'appariement pour la comparaison. Les enregistrements dans le fichier des mineurs pour lesquels le statut vital était "non retrouvé lors du suivi" ont été éliminés. Après que les enregistrements pour lesquels des valeurs manquaient soit dans au moins un champ d'appariement, soit dans le champ de l'année de naissance, aient aussi été supprimés, le fichier A (basé sur le fichier des mineurs) renfermait 45,638 enregistrements et le fichier B (basé sur la BCDM) comprenait 24,597 enregistrements. En limitant les comparaisons des deux fichiers aux enregistrements pour lesquels il y avait accord pour le code "NYSIIS" du nom de jeune fille de la mère et la différence entre les années de naissance est d'au plus un, il y avait 26,500 non-concordances vraies et 2,063 concordances vraies.

Les fréquences des résultats pour les concordances vraies et pour les non-concordances vraies sont présentées au tableau 6. Tous les modèles log-linéaires correspondant à un modèle à variable latente non saturé (c'est-à-dire, tous les modèles comprenant moins de huit paramètres) sont rejetés par les données sur les fréquences pour les non-concordances vraies à un niveau de signification très faible. Parmi les modèles comprenant moins de huit paramètres, le modèle $U(1)$, $U(2)U(4)$, $U(3)U(4)$ correspond à la statistique la plus faible du test du rapport des vraisemblances pour le manque d'ajustement, soit 35.29. Le modèle $M(1)$, $M(2)M(4)$, $M(3)M(4)$ fournit un ajustement adéquat avec les données pour la concordance vraie (la statistique du test du rapport des vraisemblances est 10.29).

Les estimations des probabilités d'accord ont été calculées à l'aide de la méthode des moments, de la méthode itérative et de la pondération itérative, en utilisant le modèle à variable latente $G(1)$, $G(2)G(4)$, $G(3)G(4)$. La statistique du test du rapport des vraisemblances pour le modèle fondé sur l'hypothèse d'indépendance correspondant à l'estimateur de la méthode des moments est 108 (six degrés de liberté). Le modèle fondé sur l'hypothèse d'indépendance est rejeté par les données à un niveau de signification très faible. Par contre, la statistique du test du rapport des vraisemblances pour le modèle à variable latente $G(1)$, $G(2)G(4)$, $G(3)G(4)$ est 1.44 (deux degrés de liberté), ce qui laisse supposer un ajustement adéquat. Les estimations, fondées sur un modèle, du taux d'erreur de classification correspondant à chaque ensemble d'estimations de probabilité ont été calculées pour divers seuils. Les taux d'erreur de classification réels sont comparés aux estimations fondées sur un modèle pour les non-concordances vraies dans le tableau 7 et pour les concordances vraies dans le tableau 8. On a modifié l'échelle des taux d'erreur pour les non-concordances vraies afin que le nombre de concordances vraies se trouve au dénominateur.

**Tableau 6**

Fréquences des résultats, données réelles

| Résultat selon l'indenticateur: 0 = Désaccord, 1 = Accord | | | | Fréquence | |
|---|---|---|---|---|---|
| Premier prénom | Code NYSIIS du nom de jeune fille de la mère | Jour de naissance | Mois de naissance | Concordances vraies | Non-concordances vraies |
| 0 | 0 | 0 | 0 | 4 | 22,100 |
| 0 | 0 | 0 | 1 | 3 | 888 |
| 0 | 0 | 1 | 0 | 11 | 2,322 |
| 0 | 0 | 1 | 1 | 128 | 211 |
| 0 | 1 | 0 | 0 | 3 | 199 |
| 0 | 1 | 0 | 1 | 7 | 19 |
| 0 | 1 | 1 | 0 | 27 | 27 |
| 0 | 1 | 1 | 1 | 242 | 13 |
| 1 | 0 | 0 | 0 | 9 | 576 |
| 1 | 0 | 0 | 1 | 10 | 32 |
| 1 | 0 | 1 | 0 | 52 | 94 |
| 1 | 0 | 1 | 1 | 392 | 4 |
| 1 | 1 | 0 | 0 | 27 | 13 |
| 1 | 1 | 0 | 1 | 32 | 1 |
| 1 | 1 | 1 | 0 | 115 | 0 |
| 1 | 1 | 1 | 1 | 1,001 | 1 |
| | | | Total | 2,063 | 26,500 |

Les estimations, fondées sur un modèle, du taux d'erreur de classification obtenues à l'aide de la méthode itérative sont très imprécises, particulièrement pour les non-concordances vraies, quelle que soit la valeur de $\mu^0$. Les estimations du taux d'erreur obtenues à l'aide de la

**Tableau 7**

Taux d'erreur de classification, non-concordances vraies, données réelles

| Taux estimé (× 12.84) | Taux réel (× 12.84) | | | | |
|---|---|---|---|---|---|
| | Méthode des moments | Méthode itérative $\mu^0 = 0.0000625$ | Méthode itérative $\mu^0 = 0.00025$ | Méthode itérative $\mu^0 = 0.001$ | Pondération itérative |
| 0.02 | 0.0368 | 1.311 | 0.1859 | 0.186 | 0.0339 |
| 0.04 | 0.0796 | 1.314 | 0.1888 | 0.193 | 0.0649 |
| 0.06 | 0.1224 | 1.317 | 0.1917 | 0.1967 | 0.0684 |
| 0.08 | 0.1573 | 1.323 | 0.1990 | 0.1994 | 0.1106 |
| 0.10 | 0.1863 | 1.333 | 0.60 | 0.4066 | 0.1282 |

**Tableau 8**

Taux d'erreur de classification, concordances vraies, données réelles

| Taux estimé | Taux réel | | | | |
|---|---|---|---|---|---|
| | Méthode des moments | Méthode itérative $\mu^0 = 0.0000625$ | Méthode itérative $\mu^0 = 0.00025$ | Méthode itérative $\mu^0 = 0.001$ | Pondération itérative |
| 0.02 | 0.0166 | 0.0141 | 0.0193 | 0.0225 | 0.0105 |
| 0.04 | 0.0318 | 0.0264 | 0.029 | 0.0278 | 0.0263 |
| 0.06 | 0.0598 | 0.0383 | 0.0472 | 0.0326 | 0.0529 |
| 0.08 | 0.0782 | 0.0416 | 0.1372 | 0.0488 | 0.0784 |
| 0.10 | 0.0966 | 0.045 | 0.1393 | 0.1371 | 0.0958 |

pondération itérative sont un peu moins précises que celles basées sur la méthode des moments pour les concordances vraies. Toutefois, elles sont beaucoup plus précises que les estimations obtenues avec la méthode des moments pour les non-concordances vraies.

Une mise en garde s'impose. Bien que le modèle $U(1)$, $U(2)U(4)$, $U(3)U(4)$ ne décrive pas de façon adéquate les dépendances parmi les non-concordances vraies, l'algorithme de pondération itérative a permis d'obtenir un bon ajustement à l'aide d'une estimation de la proportion d'enregistrements appariés (0.0747) qui diffère un peu de la valeur vraie (0.0722). On peut aussi obtenir un ajustement semblable à l'aide du modèle $G(1)G(2)$, $G(1)G(3)$, $G(4)$ ainsi qu'une estimation de 0.077 pour la proportion des appariements. Les estimations du taux d'erreur basées sur le modèle $G(1)G(2)$, $G(1)G(3)$, $G(4)$ ne sont pas meilleures que les estimations obtenues à l'aide de la méthode des moments.

## 7. CONCLUSIONS

Dans cet article, on a traité de la question de l'estimation des taux d'erreur de classification pour le couplage d'enregistrements. Le modèle de Fellegi-Sunter permet de calculer des estimations des taux d'erreur de classification à l'aide d'estimations des probabilités d'accord. Ces estimations fondées sur un modèle ont généralement de mauvaises propriétés en pratique. Il a été démontré que leurs propriétés peuvent être améliorées en estimant avec soin les probabilités d'accord. Trois méthodes d'estimation ont été évaluées à l'aide de données synthétiques ainsi que de renseignements provenant d'une application réelle.

Pour deux des trois méthodes, on a utilisé l'hypothèse que les résultats des comparaisons pour différents champs de données sont indépendants. Cette hypothèse n'était pas valable soit pour les données synthétiques, soit pour les données réelles. Les données synthétiques incluaient de fortes dépendances pour les concordances vraies et des dépendances mineures pour les non-concordances vraies. Les dépendances pour les données réelles étaient particulièrement importantes dans le cas des non-concordances vraies. Les estimations du taux d'erreur de classification obtenues à l'aide de la méthode des moments, qui est fondée sur l'hypothèse d'indépendance, comportaient des biais considérables pour les données synthétiques et étaient relativement imprécises pour les données réelles. L'importance du biais dans les estimations des taux d'erreur de classification obtenues à l'aide de la méthode itérative dépendait de la définition d'un ensemble initial de concordances. Bien que certaines définitions de l'ensemble initial de concordances aient mené à des biais relativement faibles, d'autres ont produit des estimations avec des biais

beaucoup plus considérables que ceux obtenus à l'aide des autres méthodes. Pour les données réelles, toutes les définitions de l'ensemble initial d'appariements considérées ont mené à des estimations très imprécises du taux d'erreur. Il n'existe pas de règles mathématiques qui permettent de choisir un ensemble initial de concordances pour la méthode itérative. Rien dans les résultats présentés dans cet article ne permet de recommander l'utilisation de cette dernière méthode.

La troisième méthode est basée sur un paramétrage de dépendances entre les résultats de comparaisons, pour différents champs de données, à l'aide d'effets log-linéaires. Avec ce paramétrage, on peut obtenir des estimations des probabilités d'accord qui ne sont pas fondées sur l'hypothèse d'indépendance en utilisant la méthode de pondération itérative pour estimer les paramètres d'un modèle à variable latente. Pour les ensembles de données synthétiques, avec absence d'indépendance, les estimations des taux d'erreur de classification fondées sur un modèle calculé par pondération itérative comprenaient des biais de beaucoup inférieurs à ceux qui s'appliquaient aux estimations basées sur l'hypothèse d'indépendance. Bien que l'ajustement obtenu à l'aide du modèle à variable latente pour la plupart des ensembles de données synthétiques ait été meilleur que celui obtenu pour un modèle fondé sur l'hypothèse d'indépendance, on relevait parfois un manque d'ajustement important avec le premier modèle. Quand les données synthétiques ont été modifiées afin d'améliorer l'ajustement obtenu avec le modèle à variable latente, rien ne montrait qu'il y avait un biais dans les estimations du taux d'erreur de classification fondées sur un modèle. Pour les données réelles, il y avait des écarts importants par rapport à l'hypothèse d'indépendance tant pour les concordances vraies que pour les non-concordances vraies. Les estimations du taux d'erreur fondées sur un modèle obtenues à l'aide de la pondération itérative étaient un peu moins précises que les estimations fondées sur la méthode des moments pour les concordances vraies et beaucoup plus précises pour les non-concordances vraies.

Les résultats présentés ici montrent que l'on peut améliorer les propriétés des estimations du taux d'erreur de classification fondées sur un modèle quand on utilise un estimateur approprié des probabilités d'accord. Les modèles à variable latente ainsi que la pondération itérative fournissent une méthode pour incorporer des dépendances entre des résultats de comparaisons pour différents champs de données pendant l'estimation des probabilités d'accord.

## REMERCIEMENTS

Les auteurs désirent remercier William Winkler qui a fourni le code machine sur lequel est basé le programme d'estimation par pondération itérative que nous avons utilisé pour obtenir nos résultats, ainsi que Fritz Scheuren et trois arbitres anonymes pour leurs commentaires, sur une version antérieure de cet article, qui ont mené à des améliorations considérables tant pour ce qui est du contenu que de la présentation. Nous désirons aussi remercier Martha Fair et Pierre Lalonde qui nous ont permis d'obtenir les données sur les mineurs ontariens ainsi que les données sur la fréquence des résultats pour les concordances vraies.

## BIBLIOGRAPHIE

BARTLETT, S., KREWSKI, D., WANG, Y., et ZIELINSKI, J.M. (1992). Évaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisé. *Techniques d'enquête*, 19, 3-13.

BELIN, T.R. (1990). A proposed improvement in computer matching techniques. Dans *Statistics of Income and Related Administrative Record Research*: 1988-1989, U.S. Internal Revenue Service, 167-172.

BELIN, T.R., et RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 657-668.

FAIR, M.E., et LALONDE, P. (1987). Identificateurs manquants et justesse de l'observation suivie. *Recueil: Les utilisations statistiques des données administratives, Statistique Canada*, 111-125.

FAIR, M.E., NEWCOMBE, H.B., et LALONDE, P. (1988). Improved mortality searches for Ontario miners using social insurance index identifiers. Rapport de recherche, Commission de contrôle de l'énergie atomique.

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

HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.

HABERMAN, S.J. (1979). *Analysis of Qualitative Data*. London: Academic Press.

IMSL (1987). Math/Library FORTRAN subroutines for mathematical applications. Houston: IMSL Inc.

MORÉ, J., GARBOW, B., et HILLSTROM, K. (1980). User guide for MINPACK-1. Argonne National Labs Report ANL-80-74.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., et JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.

THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.

WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 145-155.

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 159–174
Statistique Canada



# Modélisation conjointe robuste de séries de données sur l'activité pour de petites régions

D. PFEFFERMANN et S.R. BLEUER[1]

RÉSUMÉ

Dans cet article, nous présentons les résultats de l'application d'un modèle d'espace d'états aux taux de chômage canadiens. Le modèle suppose une décomposition additive des valeurs de la population en une tendance, une composante saisonnière et une composante irrégulière, ainsi que des relations autorégressives distinctes pour les six séries d'erreurs de l'enquête correspondant aux six estimateurs de panel mensuels. Le modèle tient compte des effets des groupes de renouvellement et permet que changent, dans le temps, les variances liées au plan qui affectent les erreurs de l'enquête. L'ajustement du modèle est effectué au niveau de petites régions, mais il tient compte de corrélations entre les séries des composantes pour différentes régions. On obtient la robustesse des estimateurs produits par le modèle en imposant, à titre de contrainte, que les estimateurs globaux mensuels fondés sur le modèle visant un groupe de petites régions pour lequel la taille d'échantillon totale est suffisamment grande coïncident avec les estimateurs directs correspondants de l'enquête. La performance du modèle, dans le cas d'une application aux provinces de l'Atlantique, est évaluée au moyen de diverses statistiques de diagnostic et de graphiques de résidus, ainsi que par des comparaisons avec des estimateurs actuellement en usage.

MOTS CLÉS: Variance liée au plan; filtre de Kalman; enquête par panel; biais lié au renouvellement; modèle d'espace d'états.

## 1. INTRODUCTION

Un modèle de série chronologique appliqué à des données d'enquête est la combinaison de deux modèles distincts: le "modèle du recensement", qui décrit l'évolution dans le temps des valeurs de la population finie, et le modèle des erreurs de l'enquête, qui décrit les relations qui existent dans la série chronologique des erreurs des estimateurs de l'enquête. Il y a au moins quatre raisons principales qui incitent à modéliser les estimateurs bruts de l'enquête:

a) Les estimateurs des valeurs de la population qui résultent du processus de modélisation (estimateurs fondés sur le modèle) ont en général des variances moindres que les estimateurs de l'enquête, notamment dans les petites régions, où les tailles des échantillons sont peu élevées.

b) Le modèle que nous employons permet d'obtenir des estimateurs pour les effets saisonniers, ainsi que pour les variances de ces estimateurs, comme sous-produit du processus d'estimation.

c) Le modèle permet d'établir des prévisions des valeurs de la population, de la tendance et des composantes saisonnières à l'égard de périodes postérieures à la période d'observation de l'échantillon pour laquelle les estimateurs directs de l'enquête sont disponibles. De telles prévisions sont importantes pour l'évaluation de la performance du modèle, de même que pour l'élaboration de politiques.

d) Le modèle peut servir à détecter des points tournants dans le niveau des séries chronologiques et à en évaluer les conséquences. (Les travaux réalisés dans ce domaine seront exposés dans un article distinct.)

La méthodologie décrite dans le présent article intègre les méthodologies présentées dans Pfeffermann et Burck (1990) et Pfeffermann (1991), avec certaines modifications et extensions. Les principales caractéristiques du modèle sont les suivantes:

1) Le modèle décompose les valeurs de la population en trois composantes non observables: tendance, saisonnalité et termes irréguliers. Des prédicteurs lissés de ces composantes (et donc des valeurs de la population) fondés sur l'ensemble des données disponibles, ainsi que les erreurs-types des erreurs de prédiction, sont obtenus directement par l'application du filtre de Kalman. Les erreurs types sont modifiées pour tenir compte de la variation additionnelle causée par l'utilisation de valeurs estimées de paramètres.

2) Le modèle utilise comme données d'entrée les différents estimateurs de panel mensuels. L'utilisation des estimateurs propres à chaque panel offre deux avantages importants par rapport à l'emploi des estimateurs moyens:

[1] D. Pfeffermann, Department of Statistics, Hebrew University, Jerusalem 91905; S.R. Bleuer, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6.

i) elle facilite la détermination du modèle de série chronologique reflétant les erreurs de l'enquête, en permettant l'analyse des contrastes entre les estimateurs des différents panels, et ii) elle donne des estimateurs plus efficaces pour les paramètres du modèle, et donc de meilleurs prédicteurs des composantes non observables du modèle.

3) Le modèle tient compte des changements des variances des erreurs de l'enquête au fil du temps, ainsi que des effets possibles des groupes de renouvellement.

4) Le modèle peut être appliqué simultanément aux estimateurs des panels de petites régions distinctes. Dans ce cas, le modèle du recensement est étendu de manière à tenir compte des corrélations croisées entre les composantes non observables des valeurs de la population qui existent dans ces régions.

5) Une modification visant à assurer la robustesse des estimateurs des petites régions à l'égard d'éventuelles défaillances du modèle est incorporée aux équations du modèle. La modification consiste à contraindre les estimateurs fondés sur le modèle d'agrégats de valeurs de la population pour un groupe de petites régions dont la taille totale de l'échantillon est suffisamment grande à coïncider avec les estimateurs de l'enquête pour les agrégats correspondants. Ainsi, les variations soudaines du niveau de la série sont répercutées sans retard sur les estimateurs fondés sur le modèle.

Le modèle, ainsi que les modifications apportées pour en assurer la robustesse, sont décrits de façon plus détaillée à la section 2. Les résultats empiriques de l'application du modèle aux quatre provinces de l'Atlantique du Canada sont présentés à la section 3. La section 4 présente un bref sommaire ainsi que des suggestions quant à la direction que pourraient emprunter les recherches futures.

Avant de conclure la présente section, mentionnons qu'aux États-Unis, les estimations du chômage sont produites pour la plupart des États à l'aide de modèles de série chronologique ayant une structure semblable à celle du modèle de notre étude. Voir Tiller (1992) pour plus de détails. Une différence importante entre les deux, toutefois, est qu'aux États-Unis, le modèle postulé pour les valeurs de la population inclut également des variables explicatives, de sorte que la tendance et la composante saisonnière ne reflètent que la tendance et les variations saisonnières non prises en compte par les variables explicatives. Les modèles ajustés aux erreurs de l'enquête sont, comme dans notre cas, du type ARMMI, et ils tiennent compte eux aussi des changements des variances des erreurs de l'enquête. Les modèles sont par ailleurs différents en raison des modes très distincts de renouvellement de l'échantillon utilisés dans les deux pays. Une autre différence notable entre les deux modèles vient du fait qu'aux États-Unis, les modèles sont ajustés à chaque État séparément et que les données d'entrée comprennent seulement les estimations moyennes de l'enquête, c'est-à-dire une seule observation chaque mois. Les modèles, par conséquent, ne tiennent pas compte des biais liés aux groupes de renouvellement.

## 2. UN MODÈLE D'ESPACE D'ÉTATS POUR LA SÉRIE CHRONOLOGIQUE DES DONNÉES SUR LE CHÔMAGE AU CANADA

### 2.1 L'enquête sur la population active du Canada

Les données canadiennes sur le chômage sont recueillies dans le cadre de l'enquête sur la population active (EPA) réalisée par Statistique Canada. L'EPA canadienne est une enquête par panel mensuelle avec renouvellement, dans laquelle chaque nouveau panel de ménages observé demeure dans l'échantillon pendant six mois consécutifs avant d'être remplacé par un autre panel de la même unité primaire d'échantillonnage (UPÉ) ou de la même strate. Les UPÉ sont délimitées géographiquement (pâtés de maisons ou centres urbains dans les régions urbaines, et groupes de secteurs de dénombrement dans les régions rurales). Les strates sont des groupes homogènes d'UPÉ délimités géographiquement (p. ex. secteurs de recensement, subdivisions de recensement et secteurs de dénombrement). Dans les régions urbaines (environ les deux tiers de l'échantillon), chaque UPÉ est représentée dans un seul panel. Dans les régions rurales, les UPÉ sont représentées dans tous les panels, mais les secteurs de dénombrement sont différents d'un panel à l'autre. Par conséquent, les estimateurs de panel distincts peuvent être présumés indépendants, propriété qui a été validée et utilisée dans d'autres études; voir par exemple Lee (1990). Pour une description récente du plan de l'EPA et de la construction des estimateurs directs de l'enquête, le lecteur est invité à consulter Singh et coll. (1990).

### 2.2 Le modèle du recensement

Dans les paragraphes qui suivent, nous examinons le cas d'une petite région unique, tandis que dans la section 2.4, nous aborderons la modélisation conjointe des estimations des panels pour un groupe de petites régions. Le modèle postulé pour les valeurs de la population est le modèle structurel de base (MSB), représenté par l'ensemble suivant d'équations:

$$Y_t = L_t + S_t + \epsilon_t; \quad L_t = L_{t-1} + R_{t-1} + \eta_{Lt};$$

$$R_t = R_{t-1} + \eta_{Rt}; \quad \sum_{j=0}^{11} S_{t+j} = \eta_{St}. \qquad (2.1)$$

Dans 2.1, $Y_t$ est la valeur de la population (taux de chômage "vrai") au temps $t$, $L_t$ est le niveau de la tendance, $R_t$ est l'accroissement, $S_t$ est l'effet saisonnier et $\epsilon_t$ est le terme irrégulier qu'on suppose être un bruit blanc avec moyenne zéro et variance $\sigma_\epsilon^2$. Ainsi, la première équation de (2.1) suppose la décomposition classique d'une série chronologique en trois composantes: tendance, variations saisonnières et terme irrégulier. Une telle décomposition est inhérente aux méthodes courantes de désaisonnalisation des données; voir par exemple Dagum (1980).

Notons toutefois que dans le présent cas, la série $\{Y_t\}$ est elle-même non observable. Les séries $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ et $\{\eta_{St}\}$ sont des perturbations de bruit blanc indépendantes de moyennes zéro et de variances $\sigma_L^2$, $\sigma_R^2$ et $\sigma_S^2 \times g(t)$ respectivement. Par conséquent, la deuxième et la troisième équation de (2.1) définissent une approximation locale d'une tendance linéaire, tandis que la dernière équation décrit l'évolution des effets saisonniers de telle manière que la somme de n'importe quel ensemble de 12 effets successifs fluctue autour de zéro. Notons que les variances des termes d'erreur $\eta_{St}$ dépendent du temps. Les fonctions $g(t)$ sont définies à la fin de la section 3.1.

Les propriétés théoriques du MSB en comparaison d'autres modèles sont examinées dans Harrison et Stevens (1976), Harvey (1984) et Maravall (1985). Les résultats empiriques illustrant la performance du modèle sont présentés dans Harvey et Todd (1983), Morris et Pfeffermann (1984) et Pfeffermann (1991). Bien qu'il soit plus limité que la famille des modèles ARMMI, le MSB est maintenant considéré comme suffisamment flexible pour donner une approximation du comportement d'une grande variété de séries chronologiques.

### 2.3 Le modèle des erreurs de l'enquête

Le modèle représentant les erreurs de l'enquête a été déterminé initialement par une analyse distincte des séries de pseudo-erreurs $e_{t,p}^{(j)} = (y_t^{(j)} - \bar{y}_t)$, $t = 1, \ldots, N$, où $y_t^{(j)}$ est l'estimateur de $Y_t$ d'après le $j$-ième panel, $j = 1, \ldots, 6$, (le panel observé pour le $j$-ième mois consécutif) et $\bar{y}_t = \sum_{j=1}^{6} y_t^{(j)}/6$ est l'estimateur moyen. Notons que $(y_t^{(j)} - \bar{y}_t) = (e_t^{(j)} - \sum_{j=1}^{6} e_t^{(j)}/6)$, où $e_t^{(j)} = (y_t^{(j)} - Y_t)$ sont les vraies erreurs de l'enquête. Ainsi, la caractéristique notable des contrastes $(y_t^{(j)} - \bar{y}_t)$ est qu'ils sont fonction uniquement des erreurs de l'enquête, peu importe le modèle représentant les valeurs de la population.

Il y a deux facteurs principaux à considérer dans le choix d'un modèle pour les erreurs de l'enquête:

a) Le modèle devrait tenir compte d'éventuels biais liés aux groupes de renouvellement ou, plus généralement, permettre des moyennes différentes pour les erreurs propres à différents panels.

b) Le modèle devrait permettre que les variances des erreurs de l'enquête changent au fil du temps.

Des biais liés aux groupes de renouvellement peuvent survenir lorsque des répondants ne donnent pas les mêmes informations à des interviews différentes; ces effets peuvent dépendre de la durée de participation à l'enquête des répondants, ou de la méthode de collecte des données (p. ex. par téléphone ou par interviews à domicile). (Dans l'EPA canadienne, les membres du premier panel sont interviewés à domicile, tandis que ceux des autres panels sont interviewés par téléphone.) Une autre cause possible d'écarts entre les moyennes des erreurs propres aux panels réside dans les différences entre les profils de non-réponse d'un panel à l'autre. Voir Pfeffermann (1991) pour une analyse plus approfondie du problème et des références concernant les études antérieures effectuées dans ce domaine.

Il survient au fil du temps des changements des variances des erreurs de l'enquête lorsque les variances sont fonction du niveau de la série. En fait, comme le montre la figure 1 de la section 3, les estimations des écarts-types des erreurs de l'enquête sont soumis à des effets saisonniers, selon un profil qui suit l'évolution saisonnière des valeurs de la population. Des changements des variances des erreurs de l'enquête peuvent aussi s'expliquer par des modifications du plan d'échantillonnage. Par exemple, en 1985-1986, la taille totale de l'échantillon de l'EPA canadienne a été ramenée de 55,000 à 48,000 ménages. Cette réduction s'est accompagnée d'autres modifications du plan. Voir Singh et coll. (1990) pour plus de détails.

L'application de méthodes simples d'estimation de modèles et de diagnostic aux pseudo-erreurs de l'enquête suggère un modèle autorégressif (AR) de 3$^{e}$ ordre pour les erreurs de l'enquête standardisées $\tilde{e}_t^{(j)} = (e_t^{(j)} - \beta_j)/SD(e_t^{(j)})$, soit:

$$\tilde{e}_t^{(j)} = \phi_{j1}\, \tilde{e}_{t-1}^{(j-1)} + \phi_{j2}\, \tilde{e}_{t-2}^{(j-2)} + \phi_{j3}\, \tilde{e}_{t-3}^{(j-3)} + u_t^{(j)},\ j = 1, \ldots, 6, \quad (2.2)$$

où les $\beta_j = E(e_t^{(j)})$ sont les biais liés aux groupes de renouvellement, les $SD(e_t^{(j)})$ sont les écarts-types liés au plan et les $u_t^{(j)}$ sont des termes de bruit blanc indépendants de moyenne zéro et de variances $\sigma_j^2$. On suppose que $\sum_{j=1}^{6} \beta_j = 0$, d'où il résulte que l'estimateur moyen de l'enquête, $\bar{y}_t$, est non biaisé. Voir Pfeffermann (1991) pour une analyse du besoin d'imposer une contrainte aux coefficients de biais. L'analyse subséquente de l'ajustement du modèle combiné défini par (2.1) et (2.2) (voir la section 2.4) permet de valider ce modèle, avec l'observation additionnelle selon laquelle les coefficients $(\phi_{j1}, \phi_{j2}, \phi_{j3})$ peuvent être supposés égaux pour $j = 4, 5, 6$. En outre, pour le premier panel, un modèle AR(1) donne déjà un bon ajustement, tandis que pour le deuxième et le troisième panel, un modèle AR(2) est approprié, bien que les coefficients soient différents. Ces relations sont valables pour chacune des quatre provinces de l'Atlantique.

L'un des arbitres ayant revu le présent article s'est demandé si le modèle AR(3) défini par (2.2) est suffisamment souple pour tenir compte des corrélations entre les estimations des panels à des décalages importants, qu'on croit être élevées en raison des "effets des UPÉ". Comme il a été mentionné à la section 2.1, les panels qui sortent de l'échantillon sont remplacés par des panels des mêmes UPÉ, et il faut habituellement plusieurs années avant qu'une UPÉ ne soit épuisée et remplacée par une UPÉ voisine. Lee (1990) présente deux ensembles de corrélations entre estimations de panels pour l'EPA canadienne. Le premier ensemble, dénoté par $\rho_j$, comprend les corrélations entre les estimations produites par le même panel, de sorte que $j$ varie de 1 à 5. Le deuxième ensemble, dénoté par $\gamma_j$, comprend les corrélations entre les estimations produites par un panel et son prédécesseur, de sorte que $j$ varie de 1 à 11. Les corrélations $\rho$ sont généralement élevées, comme il faut s'y attendre, mais il convient de souligner qu'elles sont plus faibles pour les données du

chômage que pour celles de l'emploi, ce qui reflète la mobilité élevée de la population active en chômage. Les corrélations $\gamma$ sont beaucoup plus faibles que les corrélations $\rho$, mais comme le signale l'auteur, le calcul de ces corrélations est beaucoup moins fiable et leur comportement est quelque peu erratique, affichant parfois une tendance à la hausse. Nous avons calculé les corrélations sérielles d'après les modèles (2.2), en remplaçant les coefficients $\phi$ par leurs valeurs estimées et nous avons observé en général un ajustement étroit avec les corrélations $\rho$ à tous les décalages (de 1 à 5). Les corrélations à des décalages plus élevés sont différentes des corrélations $\gamma$ correspondantes, mais, fait intéressant, elles sont la plupart du temps plus élevées et diminuent toujours à mesure que $j$ augmente.

Une autre question ayant trait au modèle (2.2) qu'ont soulevée les arbitres avait trait à la possibilité d'appliquer la transformation logarithmique aux données brutes de façon à stabiliser les variances des erreurs de l'enquête, plutôt que de modéliser les erreurs standardisées. Deux raisons principales incitent à ne pas recourir à la transformation logarithmique dans notre cas. En premier lieu, l'utilisation de cette transformation entraînerait une décomposition multiplicative des taux de chômage de la population, ce qui va à l'encontre de la pratique courante qui consiste à supposer une décomposition additive. À Statistique Canada, les taux de chômage des deux plus grandes provinces sur les quatre examinées dans notre étude sont désaisonnalisés sur la base d'une décomposition additive. Aux États-Unis, les modèles ajustés aux séries du chômage des États reposent eux aussi sur une décomposition additive. Voir Tiller (1992). La deuxième raison est que les changements des variances des erreurs de l'enquête peuvent résulter de modifications du plan d'échantillonnage et, en particulier, de changements des tailles des échantillons. De tels changements produisent des mouvements discrets des variances qui ne peuvent être traités efficacement par la transformation logarithmique. Comme l'a signalé par ailleurs l'un des arbitres, le fait de transformer les données a l'inconvénient de produire une non-linéarité dans l'agrégation des estimations sur les divers panels et/ou petites régions.

Le modèle défini en (2.2) tient compte des deux facteurs à considérer énoncés plus haut. L'application réelle du modèle, toutefois, exige deux modifications:

1. Pour les trois premiers panels, il n'y a pas assez de données antérieures pour permettre l'ajustement d'un modèle AR(3). Par exemple, l'erreur de l'enquête $e_t^{(1)}$ correspond au panel qui fait partie de l'enquête pour la première fois. Pour surmonter ce problème, nous remplaçons les erreurs de l'enquête manquantes par les erreurs de l'enquête correspondant aux panels précédemment choisis dans les mêmes UPÉ ou strates. Par exemple, le modèle AR(2) ajusté à $\tilde{e}_t^{(2)}$ est

$$\tilde{e}_t^{(2)} = \phi_{21}\,\tilde{e}_{t-1}^{(1)} + \phi_{22}\,\tilde{e}_{t-2}^{(6)} + u_t^{(2)}. \qquad (2.3)$$

Notons que le panel observé pour la deuxième fois au mois $t$ remplace au temps $(t-1)$ le panel observé pour la sixième fois au mois $(t-2)$, de sorte que les deux panels représentent les mêmes UPÉ ou strates. L'utilisation de valeurs substituts pour les erreurs de l'enquête dans le cas des trois premiers panels peut expliquer les modèles différents obtenus pour ces panels, par rapport au modèle obtenu pour les trois autres panels.

2. Les écarts-types vrais des erreurs de l'enquête sont inconnus, tandis que les estimations de l'enquête des écarts-types sont elles-mêmes l'objet d'erreurs d'échantillonnage. Pour surmonter ce problème, nous utilisons des valeurs lissées des écarts-types estimés, obtenues par l'ajustement de la relation

$$(\widetilde{SD})_t = \hat{\gamma}(\widehat{SD})_{t-1} + \hat{\gamma}_0\, t + \sum_{i=1}^{12} \hat{\gamma}_i D_{it}, \qquad (2.4)$$

où les coefficients $\gamma$ sont estimés par les moindres carrés ordinaires. La notation $(\widehat{SD})_t$ définit l'estimation brute, non lissée, de l'écart-type lié au plan de l'estimateur moyen de l'enquête, $\bar{y}_t$, au mois $t$ et les $\{D_{it}\}$ sont des variables fictives tenant compte des effets saisonniers mensuels de telle sorte que $D_{it} = 1$ quand $t = 12k + i$, $k = 0, 1, \ldots, i = 1, \ldots, 12$ et $D_{it} = 0$ dans les autres cas. Les écarts-types lissés des erreurs propres aux panels sont donnés par $\widetilde{SD}(e_t^{(j)}) = \sqrt{6}(\widetilde{SD})_t$. Ces dernières estimations sont utilisées comme substituts des écarts-types vrais, qui sont inconnus.

### 2.4 Représentation sous forme d'espace d'états et estimation du modèle représentant les estimateurs de l'enquête

Il découle de (2.1) que les estimateurs des panels peuvent être modélisés par

$$y_t^{(j)} = L_t + S_t + \epsilon_t + e_t^{(j)}, \quad j = 1, \ldots, 6, \qquad (2.5)$$

où

$$L_t = L_{t-1} + R_{t-1} + \eta_{Lt}; \quad R_t = R_{t-1} + \eta_{Rt};$$

$$\sum_{j=0}^{11} S_{t+j} = \eta_{St}, \qquad (2.6)$$

avec $\{\epsilon_t\}$, $\{\eta_{Lt}\}$, $\{\eta_{Rt}\}$ et $\{\eta_{St}\}$ définis comme en (2.1). Les modèles distincts définis par (2.5), (2.6) et (2.2) peuvent être fondus en une représentation compacte d'espace d'états avec $\underline{y}_t' = (y_t^{(1)}, \ldots, y_t^{(6)})$ comme données d'entrée, semblable à la représentation décrite dans Pfeffermann (1991). Suivant cette représentation, les erreurs de l'enquête (et, dans la présente étude, les termes irréguliers du recensement également) sont inclus dans le vecteur d'états, de façon qu'il n'y ait pas de termes résiduels dans l'équation des observations définie par (2.5). Contrairement à la situation décrite dans Pfeffermann (1991), toutefois, la matrice de transition et la matrice de

variance-covariance (V-C) des termes d'erreur des états ne sont pas fixes dans le temps, car elles dépendent des variances liées au plan des erreurs de l'enquête qui, comme il a été expliqué à la section 2.3, changent avec le temps.

La représentation du modèle sous forme d'espace d'états nous permet de mettre à jour, de lisser ou de prédire les vecteurs d'états et, par conséquent, la tendance, la composante saisonnière et la valeur de la population à n'importe quel mois $t$, au moyen du filtre de Kalman. Désignons par $\underset{\sim}{\alpha}_t$ le vecteur d'états correspondant au mois $t$. Le vecteur d'états comprend le niveau de la tendance, l'accroissement et les effets saisonniers, les biais liés aux groupes de renouvellement et les erreurs de l'enquête. Voir Pfeffermann (1991) pour plus de détails. Par "mise à jour", nous entendons l'estimation de $\underset{\sim}{\alpha}_t$, au mois $t$, d'après toutes les données existant jusqu'au mois $t$ inclusivement. Par "lissage", nous entendons l'estimation de $\underset{\sim}{\alpha}_t$ d'après toutes les données disponibles pour l'ensemble des mois qui précèdent et qui suivent le mois $t$. Le lissage est nécessaire pour améliorer des estimations passées, par exemple lorsqu'on estime les effets saisonniers ou qu'on estime des variations des valeurs de la population ou de la tendance. Quant à la "prédiction" des vecteurs d'états relatifs à des mois postérieurs à la période d'observation de l'échantillon, elle est importante pour l'élaboration de politiques. Les prédictions portant sur la période d'observation de l'échantillon permettent d'évaluer la performance du modèle, p. ex. en comparant les estimations de panel découlant de la prédiction de vecteurs d'états avec les estimations réelles. Voir la section 3 pour plus de détails. La théorie des modèles d'espace d'états et du filtre de Kalman est élaborée dans plusieurs publications; voir Pfeffermann (1991) pour les équations de filtrage et de lissage, avec mention de références. Notons que les équations de filtrage et de lissage fournissent non seulement les trois ensembles d'estimateurs pour n'importe quel mois donné $t$, mais aussi les matrices V-C des erreurs d'estimation correspondantes.

L'application réelle du filtre de Kalman exige l'estimation des paramètres inconnus du modèle et l'initialisation du filtre, c'est-à-dire l'estimation du vecteur d'états initial $\underset{\sim}{\alpha}_0$ et de la matrice V-C correspondante des erreurs d'estimation. Pour une petite région unique, les paramètres inconnus du modèle sont les quatre variances des termes d'erreur du modèle du recensement (2.1), ainsi que les huit coefficients d'autorégression et les six variances résiduelles des modèles des erreurs propres aux panels (2.2). (Les moyennes des groupes de renouvellement sont incluses dans les vecteurs d'états à titre de coefficients fixes, invariables dans le temps.) Afin de réduire le nombre de paramètres inconnus dans le modèle d'espace d'états combiné, nous supposons que $\sigma_j^2 = \sigma^2 \times \tilde{\sigma}_j^2, j = 1, \ldots, 6$, où les $\{\sigma_j^2\}$ sont les variances résiduelles dans (2.2) et les $\tilde{\sigma}_j^2$ sont les estimations des variances résiduelles obtenues par l'ajustement des équations d'autorégression aux pseudo-erreurs de l'enquête, $e_{t,p}^{(j)}$, définies à la section 2.3. Cette hypothèse réduit le nombre de paramètres inconnus de 18 à 13. (Les estimations $\tilde{\sigma}_j^2$ sont très voisines pour $j = 4, 5, 6$ et ont été supposées égales.)

Si l'on suppose que les termes d'erreur dans les modèles du recensement et des erreurs de l'enquête ont une distribution normale, les paramètres inconnus du modèle peuvent être estimés par une maximisation de la vraisemblance. Voir Pfeffermann et Burck (1991) pour une brève description de la façon d'utiliser l'algorithme de maximisation de la méthode de notation et pour des détails sur l'initialisation du filtre. Cet article inclut des références à des analyses plus approfondies.

### 2.5 Rajustements pour tenir compte de l'utilisation de valeurs estimées des paramètres

Une fois que les paramètres inconnus du modèle ont été estimés, les équations du filtre de Kalman peuvent être appliquées, les valeurs réelles des paramètres étant remplacées par les estimations. Comme il a été indiqué à la section 2.4, le filtre de Kalman produit non seulement des estimations des vecteurs d'états, mais aussi des matrices V-C des erreurs d'estimation correspondantes. L'utilisation de ces matrices V-C peut toutefois poser un problème du fait que ces dernières ne tiennent pas compte de la variation additionnelle attribuable au fait que les paramètres sont estimés, ce qui peut engendrer une sous-estimation des variances vraies.

En notation mathématique, supposons que $\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}})$ soit l'estimateur de $\underset{\sim}{\alpha}_t$, au mois $t$, d'après toutes les données disponibles jusqu'à un mois donné $n$, où $\hat{\underset{\sim}{\lambda}}$ représente les estimateurs des paramètres inconnus du modèle. L'erreur d'estimation peut être ainsi décomposée:

$$[\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \underset{\sim}{\alpha}_t] = [\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) - \underset{\sim}{\alpha}_t] + [\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda})]. \quad (2.7)$$

Il s'agit de la somme de l'erreur qu'on obtiendrait si $\underset{\sim}{\lambda}$ était connu et de l'erreur attribuable à l'estimation de $\underset{\sim}{\lambda}$. Les deux termes du côté droit de (2.7) sont non corrélés. Un moyen simple de vérifier cette propriété consiste à observer que $\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) = E(\underset{\sim}{\alpha}_t \mid Y, \underset{\sim}{\lambda})$, où $Y$ représente l'ensemble des données disponibles. Puisque la condition porte sur $Y$ et $\underset{\sim}{\lambda}$, $[\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda})]$, est non stochastique, tandis que $E\{[\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) - \underset{\sim}{\alpha}_t] \mid Y, \underset{\sim}{\lambda}\} = \underset{\sim}{0}$. Il s'ensuit, d'après (2.7), que

$$\begin{aligned} Q_t &= E\{[\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \underset{\sim}{\alpha}_t][\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \underset{\sim}{\alpha}_t]'\} \\ &= E\{[\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) - \underset{\sim}{\alpha}_t][\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}) - \underset{\sim}{\alpha}_t]'\} \\ &\quad + E\{[\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda})][\hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}}) - \hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda})]'\} \\ &= A_t + B_t. \end{aligned} \quad (2.8)$$

Pour estimer $A_t$ et $B_t$, nous faisons porter la condition sur $Y$ et nous suivons la méthode proposée par Hamilton (1986). Selon cette méthode, des réalisations $\underset{\sim}{\lambda}_{(k)}$, $k = 1, \ldots, K$ sont produites à partir de la distribution postérieure normale asymptotique de $\underset{\sim}{\lambda}$, c.-à-d. à partir d'une distribution $N(\hat{\underset{\sim}{\lambda}}, \hat{\Lambda})$ où $\hat{\underset{\sim}{\lambda}}$ est l'estimateur du maximum de vraisemblance de $\underset{\sim}{\lambda}$ et $\hat{\Lambda}$ est la matrice V-C asymptotique de $\hat{\underset{\sim}{\lambda}}$. ($\hat{\underset{\sim}{\lambda}}$ et $\hat{\Lambda}$ sont tous deux obtenus par la méthode de notation.) Le filtre de Kalman est alors appliqué avec

chacune de ces réalisations, ce qui donne des estimations $\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}_{(k)})$ avec matrices V-C $P_t(\underset{\sim}{\lambda}_{(k)})$. Les matrices $A_t$ et $B_t$ sont ainsi estimées:

$$\hat{A}_t = \frac{1}{k} \sum_{k=1}^{K} P_t(\underset{\sim}{\lambda}_{(k)});$$

$$\hat{B}_t = \frac{1}{k} \sum_{k=1}^{K} [\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}_{(k)}) - \hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}})][\hat{\underset{\sim}{\alpha}}_t(\underset{\sim}{\lambda}_{(k)}) - \hat{\underset{\sim}{\alpha}}_t(\hat{\underset{\sim}{\lambda}})]'. \tag{2.9}$$

Ansley et Kohn (1986) proposent un estimateur pour $B_t$ fondé sur une approximation de série de Taylor de premier ordre. L'utilisation de leur estimateur exige moins de calculs, mais la méthode proposée par Hamilton est un peu plus flexible sur le plan des hypothèses posées, et elle donne une meilleure idée de la mesure dans laquelle les résultats produits par le filtre de Kalman sont sensibles aux erreurs des estimateurs des paramètres.

### 2.6 Modélisation conjointe pour un groupe de petites régions

Le modèle examiné jusqu'ici concerne une seule région. Quand les tailles d'échantillon des diverses régions sont peu élevées, on peut souvent obtenir des estimateurs plus efficaces en modélisant en outre les relations transversales entre les valeurs de la population de plusieurs régions. De toute évidence, l'accroissement d'efficacité résultant d'une telle modélisation conjointe dépend de la taille des échantillons des petites régions et de la similitude, entre les régions, du comportement des valeurs de la population dans le temps.

Les erreurs de l'enquête étant indépendantes entre les régions, toute modélisation conjointe des estimateurs de l'enquête s'applique seulement au modèle du recensement. Pour modéliser les taux de chômage dans les quatre provinces de l'Atlantique, nous adoptons la démarche de Pfeffermann et Burck (1990) et permettons des corrélations contemporaines non nulles entre les termes d'erreur correspondants des modèles du recensement valables dans ces provinces. Par conséquent, si $\underset{\sim}{y}'_{ta} = (\epsilon_t^{(a)}, \eta_{Lt}^{(a)}, \eta_{Rt}^{(a)}, \eta_{St}^{(a)})$ dénote le vecteur des termes d'erreur au temps $t$ associé au modèle du recensement valable dans la région $a$, on suppose que $C_{a,b} = E(\underset{\sim}{y}_{ta}\, \underset{\sim}{y}'_{tb})$ est diagonale, mais comporte peut-être des covariances non nulles sur la diagonale principale. La conséquence réelle de cette hypothèse est que si, par exemple, il y a une augmentation sensible du niveau de la tendance dans une province, on peut s'attendre à ce que des augmentations semblables surviennent dans d'autres provinces.

Le modèle conjoint résultant, valable pour les quatre provinces (ou plus généralement pour un groupes de régions), peut encore être présenté sous forme d'un espace d'états; voir les équations (2.7) et (2.8) dans Pfeffermann et Burck (1990). L'ajustement de ce modèle pose toutefois un problème important, du fait que l'estimation conjointe de tous les paramètres inconnus exige trop de l'ordinateur en temps de calcul et en espace de stockage. (Le programme informatique rédigé pour l'application de la méthode de notation utilise des dérivées du premier ordre numériques, de sorte que le calcul de chaque dérivée exige un balayage distinct de la totalité des données. Chaque balayage nécessite le calcul des équations du filtre de Kalman pour chaque mois inclus dans la période d'observation de l'échantillon.)

Pour solutionner ce problème, nous avons d'abord ajusté les modèles définis par (2.5), (2.6) et (2.2) séparément pour chacune des provinces. Nous avons également supposé des corrélations égales entre les termes d'erreur correspondants des modèles du recensement des différentes provinces, c'est-à-dire

$$\phi_{a,b} = C_{a,a}^{-1/2}\, C_{a,b}\, C_{b,b}^{-1/2} = \phi \qquad 1 \le a,b \le 4, \tag{2.10}$$

où $C_{a,a} = E(\underset{\sim}{y}_{ta}\, \underset{\sim}{y}'_{ta})$. Les quatre corrélations maximisant la vraisemblance du modèle conjoint ont été déterminées par une méthode de recherche par quadrillage, les autres paramètres du modèle étant maintenus fixes à leur valeur préalablement estimée.

L'hypothèse des corrélations égales réduit de beaucoup le nombre de paramètres inconnus. Elle se justifie aussi par le petit nombre de régions examinées dans cette étude, qui permet de supposer qu'aucune autre structure préexistante régissant ces corrélations ne pourrait être détectée de façon sûre. Plus concrètement, une simple ventilation de la population active par industrie (tableau 1 de la section 3) montre des fréquences relatives très semblables dans les quatre provinces, ce qui laisse croire à un degré élevé d'homogénéité de leurs économies.

### 2.7 Modifications pour se prémunir contre les défaillances du modèle

L'utilisation d'un modèle pour la production de statistiques officielles soulève la question de la protection contre d'éventuelles défaillances du modèle. Comme il est indiqué ci-après, il n'est pas possible de tester le modèle chaque fois que de nouvelles données deviennent disponibles. C'est pourquoi il faut doter le modèle d'un mécanisme intrinsèque visant à assurer la robustesse des estimateurs lorsque le modèle ne tient plus.

Pour la modélisation des séries sur la population active dans de petites régions, nous avons utilisé la modification proposée par Pfeffermann et Burck (1990). En vertu de cette modification, les estimations des vecteurs d'états mis à jour à n'importe quel temps $t$ sont soumises à la contrainte suivante:

$$\sum_{a=1}^{A} w_{ta}\, \hat{Y}_{ta} = \sum_{a=1}^{A} w_{ta}\, \bar{y}_{ta} \qquad t = 1, 2, \ldots, \tag{2.11}$$

où $\hat{Y}_{ta}$ est l'estimateur fondé sur le modèle de la valeur de la population $Y_{ta}$ dans la région $a$, $\bar{y}_{ta} = 1/6 \sum_{j=1}^{6} y_{ta}^{(j)}$ est l'estimateur de l'enquête correspondant et $w_{ta} = M_{ta}/M_t$ est la taille relative de la population active dans cette région, c'est-à-dire que $M_t = \sum_{a=1}^{4} M_{ta}$ et $\sum_{a=1}^{4} w_{ta} = 1$.

Notons que $\sum_{a=1}^{A} w_{ta} \hat{Y}_{ta}$ et $\sum_{a=1}^{A} w_{ta} \bar{y}_t$ sont respectivement l'estimateur fondé sur le modèle et l'estimateur direct de l'enquête de la valeur totale de la population pour le groupe de régions examiné. La condition 2.11 peut aussi être écrite sous la forme $\sum_{a=1}^{A} w_{ta} \bar{e}_{ta} = 0$ où $\bar{e}_{ta} = \sum_{j=1}^{6} e_{ta}^{(j)}/6$ est l'erreur de l'enquête moyenne pour la région $a$. Pfeffermann et Burck (1990) montrent comment modifier les équations du filtre de Kalman de façon qu'elles produisent l'estimateur des vecteurs d'états assujetti à cette contrainte et sa matrice V-C appropriée selon le modèle (sans la contrainte), pour chaque mois $t$.

La justification de cette modification est simple. On suppose que la taille totale de l'échantillon de l'ensemble des régions est suffisamment grande et, par conséquent, que les estimateurs globaux de l'enquête sont fiables. Cette hypothèse, en fait, dicte le niveau d'agrégation requis; voir ci-dessous. En obligeant les estimateurs globaux fondés sur le modèle à coïncider avec les estimateurs globaux de l'enquête, l'analyste s'assure que toute variation réelle des valeurs de la population reflétée dans les estimateurs de l'enquête sera aussi reflétée dans les estimateurs fondés sur le modèle. Signalons qu'en l'absence d'une telle contrainte imposée aux estimateurs, s'il survenait par exemple des variations soudaines du niveau de la série, celles-ci ne se répercuteraient sur les estimateurs fondés sur le modèle que plusieurs mois plus tard, car ces estimateurs dépendent non seulement des données courantes, mais aussi des données passées. Par contre, si aucune variation prononcée ne survient, on peut s'attendre à ce que les estimateurs fondés sur le modèle se conforment approximativement aux contraintes, même si ces dernières ne sont pas imposées explicitement. Par conséquent, au cours de périodes normales, les estimateurs assujettis à la contrainte devraient être presque aussi performants que les estimateurs exempts de contrainte.

L'hypothèse selon laquelle la taille totale de l'échantillon de l'ensemble des régions est élevée et donc que l'estimateur global de l'enquête est suffisamment proche de la valeur correspondante de la population est critique. Elle garantit (probabilité élevée) que la modification interviendra uniquement s'il survient des changements réels dans les valeurs de la population, et non par suite de vastes erreurs d'échantillonnage. Il convient de signaler, comme l'a d'ailleurs mentionné l'un des arbitres, que dans l'application de la méthode aux provinces de l'Atlantique décrite à la section 3, l'estimateur global ne se fonde que sur quatre provinces, de sorte que son erreur-type est d'environ 50% des erreurs-types des estimateurs de l'enquête pour les provinces, selon les tailles des échantillons des provinces. (Les estimateurs de l'enquête pour les provinces sont conditionnellement indépendants, les conditions étant les valeurs de la population pour les provinces correspondantes.) Par conséquent, si les contraintes doivent être utilisées en pratique, l'agrégation devrait être faite sur un plus grand nombre de provinces ou sur d'autres petites régions.

Les deux autres méthodes suivantes ont été suggérées pour aborder le problème de la robustesse:

i) Procéder à une détection des valeurs aberrantes de la série chronologique, comme il es proposé par exemple dans Chang, Tiao et Chen (1988).

ii) Modéliser les séries chronologiques des proportions $\{\hat{\pi}_{ta} = \bar{y}_{ta}/\sum_{a=1}^{A} \bar{y}_{ta},\ a = 1, \ldots, (A-1)\}$, si ces séries chronologiques affichent un profil plus lisse que les séries $\{\bar{y}_{ta}\}$.

La détection des valeurs aberrantes est un important aspect de toute activité de modélisation, mais il reste à trouver la façon de modifier les estimations des valeurs de la population une fois que des observations (estimations de l'enquête) sont détectées à titre de valeurs aberrantes. Notons à ce propos que notre intérêt porte principalement sur les estimations courantes, c'est-à-dire les estimations les plus récentes disponibles. Dans Chang, Tiao et Chen (1988), la détection des valeurs aberrantes a pour objet de *retrancher* leur effet des observations, de façon à permettre de mieux comprendre la structure sous-jacente de la série et d'améliorer l'estimation des paramètres du modèle. Mais si la cause d'une valeur aberrante est un mouvement réel du niveau des valeurs de la population, ce mouvement ne doit pas être retranché, mais plutôt pris en compte dans les estimateurs fondés sur le modèle. Harrison et Stevens (1976) proposent de tenir compte de tels changements en modifiant la distribution antérieure des vecteurs d'états, p. ex. en accroissant les variances des erreurs des vecteurs d'états de façon à permettre des variations plus rapides des estimateurs des vecteurs d'états. Un exemple de cette façon de procéder est présenté dans Morris et Pfeffermann (1984). Notre méthode consistant à imposer la contrainte selon laquelle les estimateurs fondés sur le modèle doivent coïncider avec les estimateurs globaux de l'enquête offre un processus plus automatique, qui n'exige pas la modification de données antérieures.

La deuxième méthode suggérée pour aborder le problème de la robustesse est attrayante, car on peut s'attendre à ce que les variations brusques des valeurs de la population s'annulent dans les rapports $\hat{\pi}_{ta}$. Le principal inconvénient du recours à cette méthode est que le modèle valable pour les rapports "vrais" $\pi_{ta}$ est naturellement très différent du modèle représentant les valeurs de la population $Y_t$ tel qu'il est défini en (2.1) et que, notamment, il n'offre plus d'estimations de la tendance et des effets saisonniers, qui sont parmi les éléments les plus importants de notre approche, comme il est mentionné dans l'introduction. Il n'est pas évident, non plus, de trouver comment extraire les estimations des valeurs de la population $Y_t$ du modèle valable pour les rapports $\hat{\pi}_{ta}$ sans faire des hypothèses additionnelles, comme par exemple notre hypothèse selon laquelle l'estimateur global de l'enquête est suffisamment voisin de la valeur correspondante de la population.

L'utilisation de contraintes de la forme (2.11) a été envisagée précédemment par Battese, Harter et Fuller (1988), ainsi que par Pfeffermann et Barnard (1991), pour l'analyse d'enquêtes transversales. Pfeffermann et Burck (1990) présentent des résultats empiriques illustrant la bonne performance des estimateurs modifiés au cours de périodes anormales. Voir aussi la section 3.

## 3. AJUSTEMENT DU MODÈLE AUX PROVINCES DE L'ATLANTIQUE: RÉSULTATS EMPIRIQUES

Le modèle défini par (2.2), (2.5), (2.6) et (2.10) a été ajusté aux estimateurs de panel mensuels pour les quatre provinces de l'Atlantique, en deux étapes. À la première étape, le modèle défini par (2.2), (2.5) et (2.6) a été ajusté à chacune des provinces séparément. À la deuxième étape, les corrélations définissant la matrice $\phi$ de (2.10) ont été estimées au moyen d'une recherche par quadrillage (voir la section 2.6). Les estimateurs obtenus sont $\text{Diag}(\phi) = (0.5, 0.25, 0.80, 0.0)$. Les données ayant servi à l'estimation du modèle portent sur les années 1982 à 1988. Les données de 1989 ont servi à faire le diagnostic du modèle, en comparant les résultats obtenus à l'intérieur et à l'extérieur de la période d'observation de l'échantillon.

### 3.1 Analyse préliminaire

Le tableau 1 présente une ventilation de la population active des quatre provinces, par industrie. Les chiffres du tableau reflètent la situation qui existait en mars 1991. Les tailles (attendues) des échantillons de l'EPA sont également indiquées. Comme on peut le voir, les répartitions en pourcentage dans les quatre provinces sont très voisines, ce qui justifie l'hypothèse de corrélations égales entre les termes d'erreur des modèles du recensement pour les différentes provinces. La similitude des répartitions permet également d'espérer une meilleure efficacité des estimateurs fondés sur le modèle, dans le cas du modèle conjoint, comparativement aux estimateurs qui ne tiennent pas compte des corrélations transversales entre les valeurs de la population pour les différentes provinces.

Deux autres importants facteurs à considérer, mentionnés à la section 2.3, sont les suivants: le modèle devrait tenir compte d'effets possibles liés aux groupes de renouvellement, ainsi que des changements des variances des erreurs de l'enquête avec le temps. Pour obtenir des estimations initiales des effets liés aux groupes de renouvellement, nous avons fait la moyenne des pseudo-erreurs de l'enquête, $e_{t,p}^{(j)} = (y_t^{(j)} - \bar{y}_t)$, $j = 1, \ldots, 6$, sur l'ensemble des mois de la période visée par l'échantillon. Nous avons ensuite divisé les moyennes par les estimations classiques des erreurs-types. (Les erreurs $e_{t,p}^{(j)}$ sont corrélées dans le temps, mais les corrélations sont faibles, car sauf pour les décalages 6, 12, *etc.*, les données de n'importe quel panel donné se rapportent à des UPÉ différentes dans les régions urbaines et à des secteurs de dénombrement différents dans les régions rurales. Voir la section 2.1.) Notons qu'en l'absence d'effets liés aux groupes de renouvellement, $E(e_{t,p}^{(j)}) = 0$ pour tous les $j$ et $t$, peu importe le modèle postulé pour les valeurs de la population.

Cette analyse préliminaire (indépendante du modèle) donne des résultats semblables aux résultats obtenus avec le modèle complet, présentés au tableau 2 de la section 3.3.

Examinons maintenant les variances des erreurs de l'enquête.

La figure 1 présente le graphique des effets saisonniers des estimateurs globaux de l'enquête dans les quatre provinces, ainsi que des effets saisonniers des erreurs-types de ces estimations (multipliés par 100). Comme précédemment, désignons par $w_{ta}$ la taille relative de la population active dans la province $a$ au temps $t$. L'estimateur global de l'enquête est défini par $y_t^* = \sum_{a=1}^{4} w_{ta}\bar{y}_{ta}$ (équation 2.11). L'erreur-type de $y_t^*$ est $(SD^*)_t = [\sum_{a=1}^{4} w_{ta}^2 (\widehat{SD})_{ta}^2]^{1/2}$. Les effets saisonniers ont été estimés par l'application du mode additif de la procédure X-11, afin qu'ils ne soient liés à aucun modèle particulier. Nous avons choisi le modèle additif puisque nous supposons une décomposition additive pour les estimateurs de l'enquête. (Comme le montre la figure 4, les effets saisonniers des estimateurs globaux de l'enquête produits par la procédure X-11 sont très voisins des effets saisonniers obtenus avec le modèle.)

La figure 1 montre que les erreurs-types sont assujetties à des variations saisonnières, affichant un profil saisonnier qui est très voisin de celui des estimateurs de l'enquête et donc des valeurs correspondantes de la population.

**Tableau 1**

Population active par industrie dans les provinces de l'atlantique, mars 1991

| | Nouvelle-Écosse | | Nouveau-Brunswick | | Terre-Neuve | | Île-du-Prince-Édouard | |
|---|---|---|---|---|---|---|---|---|
| Taille de l'échantillon | 4,409 | | 3,843 | | 2,970 | | 1,421 | |
| | Milliers | % | Milliers | % | Milliers | % | Milliers | % |
| Agriculture | 7 | 1.7 | 7 | 2.3 | 0.5 | 0.2 | 6.0 | 9.8 |
| Autres industries primaires | 18 | 4.4 | 13 | 4.2 | 18.0 | 7.7 | 4.0 | 6.6 |
| Secteur manufacturier | 44 | 10.7 | 37 | 11.9 | 23.0 | 9.9 | 6.0 | 9.8 |
| Construction | 24 | 5.9 | 21 | 6.8 | 18.0 | 7.7 | 4.0 | 6.6 |
| Transp. et communication | 35 | 8.6 | 30 | 9.6 | 20.0 | 8.6 | 5.0 | 8.3 |
| Commerce | 81 | 19.8 | 61 | 19.6 | 41.0 | 17.6 | 10.0 | 16.4 |
| Finances | 20 | 4.9 | 12 | 3.9 | 6.0 | 2.6 | 0.5 | 0.8 |
| Services | 143 | 35.0 | 107 | 34.4 | 83.0 | 35.6 | 19.0 | 31.1 |
| Administration publique | 36 | 8.8 | 22 | 7.0 | 23.0 | 9.9 | 6.0 | 9.8 |
| Non classé | 1 | 0.2 | 1 | 0.3 | 0.5 | 0.2 | 0.5 | 0.8 |
| Total | 409 | 100.0 | 311 | 100.0 | 233.0 | 100.0 | 61.0 | 100.0 |

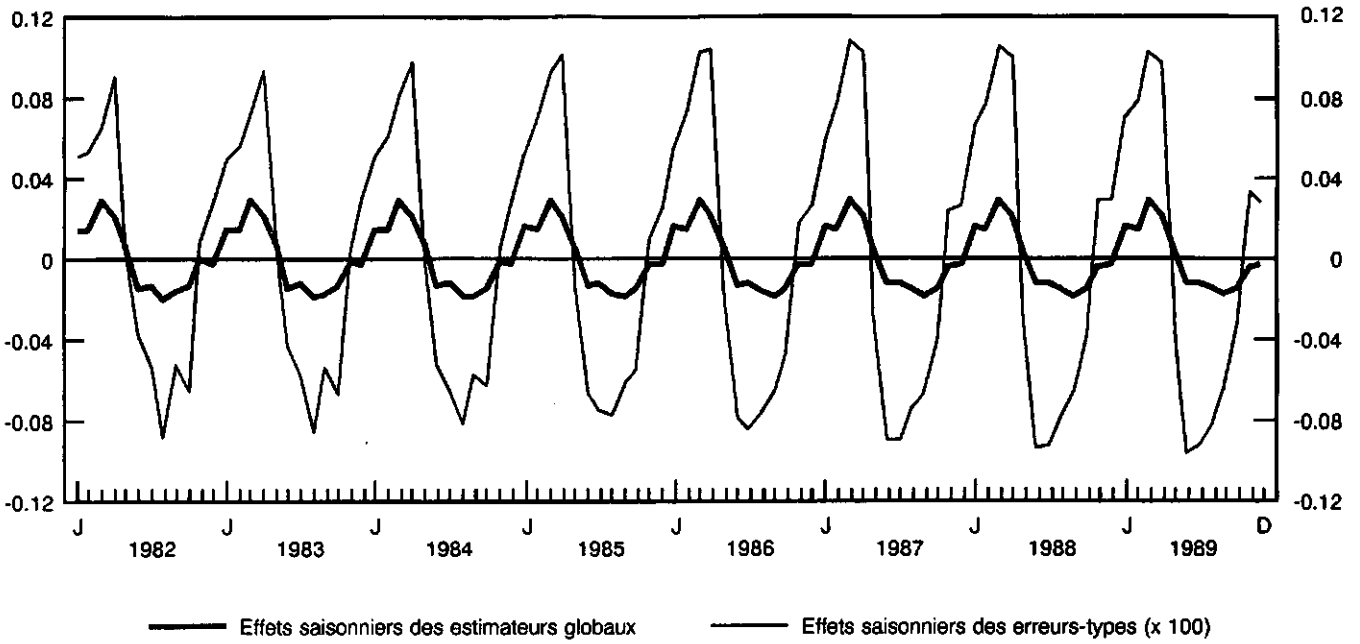


**Figure 1.** Effets saisonniers des estimateurs globaux de l'enquête et des erreurs-types des estimateurs globaux de l'enquête (× 100)

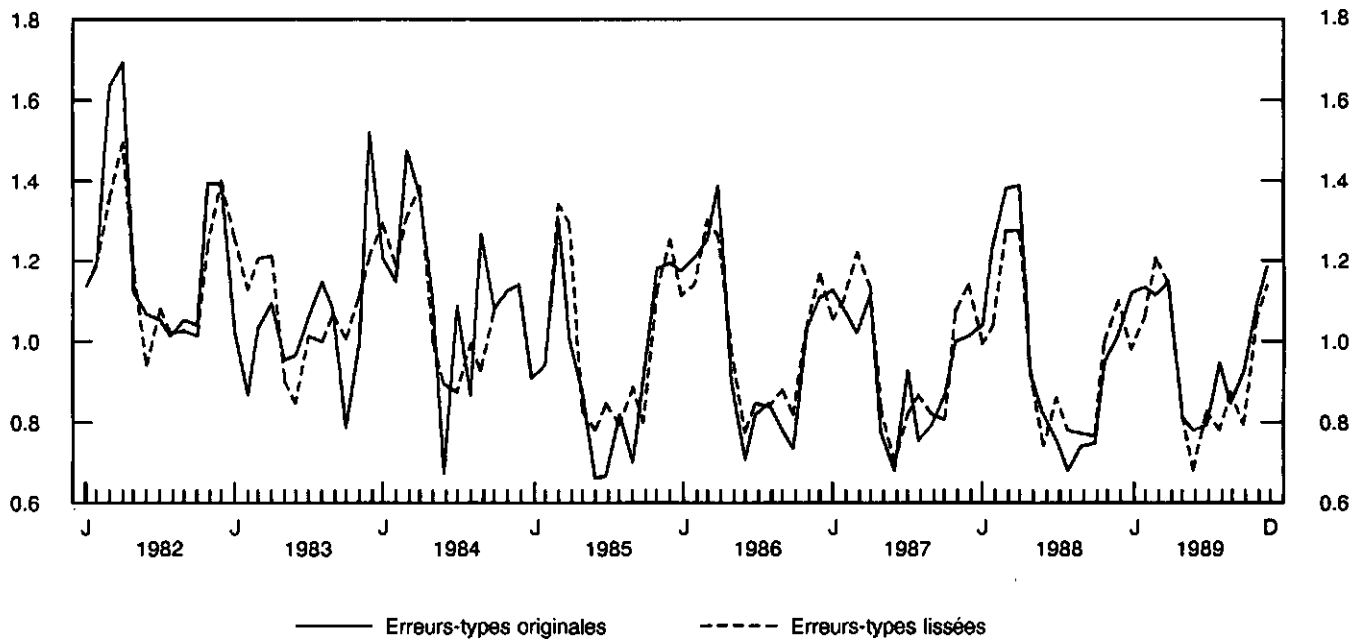


**Figure 2.** Erreurs-types originales et lissées des estimateurs de l'enquête (× 100) – Île-du-Prince-Édouard

Comme il a été mentionné à la section 2.3, plutôt que d'utiliser les estimations originales des erreurs-types dues au plan dans les modèles ajustés aux erreurs propres aux panels, nous utilisons des valeurs lissées, ce qui réduit les effets des erreurs d'échantillonnage sur les estimateurs de l'enquête. La figure 2 présente le graphique des deux ensembles d'estimateurs pour l'Île-du-Prince-Édouard (Î.-P.-É.), qui est la plus petite province de la région de l'Atlantique, et donc celle où les tailles des échantillons sont les plus faibles. Comme on peut le voir, l'effet du lissage est d'éliminer les estimations brutes extrêmes, mais, par ailleurs, les valeurs lissées suivent en gros le même profil que les estimations brutes. Les graphiques des autres provinces présentent des profils semblables, mais les écarts entre les estimations brutes et les estimations lissées sont moins prononcés en raison du fait que les tailles des échantillons sont plus élevées dans ces provinces.

Nous concluons cette section en définissant les modèles postulés pour les effets saisonniers dans les quatre provinces. Notre modèle initial supposait des variances fixes pour les termes d'erreur $\eta_{St} = \sum_{j=0}^{11} S_{t+j}$, $t = 1, 2, \ldots$ (voir l'équation 2.1). Les erreurs prédites $\hat{\eta}_{St} = \sum_{j=0}^{11} \hat{S}_{t+j}$ fournies par ce modèle diminuaient en valeur absolue en fonction du temps dans trois des quatre provinces, et augmentaient dans la dernière. Notons qu'en vertu du modèle défini par (2.1), avec des variances constantes des termes d'erreur des états, le filtre de Kalman converge vers un état stable dans lequel les matrices V-C des estimateurs des vecteurs d'états, et donc les $\hat{\eta}_{st}$, sont constants. Donc, nous avons modifié le modèle initial de telle façon que $\mathrm{VAR}(\eta_{St}) = \sigma_S^2 \times g(t)$ où, pour la Nouvelle-Écosse, Terre-Neuve et l'Î.-P.-É., $g(t) = t^{(-3/2)}$, tandis que pour le Nouveau-Brunswick, $g(t) = t^{1/2}$.

### 3.2 Résultats

#### 3.2.1 Biais liés aux groupes de renouvellement

Le tableau 2 montre les biais liés aux groupes de renouvellement (BGR) et leurs erreurs-types estimées (E.-T.) pour les quatre provinces, en vertu du modèle complet défini par (2.3), (2.5), (2.6) et (2.10).

**Tableau 2**

Biais liés aux groupes de renouvellement et erreurs-types pour les quatre provinces (× 100)

| Panels | Nouvelle-Écosse | | Nouveau-Brunswick | | Terre-Neuve | | Î.-P.-É. | |
|---|---|---|---|---|---|---|---|---|
| | BGR | E.-T. | BGR | E.-T. | BGR | E.-T. | BGR | E.-T. |
| 1 | −0.20 | 0.10 | −0.02 | 0.11 | −0.47 | 0.13 | 0.32 | 0.17 |
| 2 | 0.18 | 0.09 | 0.40 | 0.10 | 0.42 | 0.12 | 0.18 | 0.15 |
| 3 | 0.32 | 0.08 | 0.24 | 0.09 | 0.47 | 0.12 | 0.31 | 0.15 |
| 4 | 0.06 | 0.07 | 0.01 | 0.09 | 0.18 | 0.12 | 0.03 | 0.15 |
| 5 | −0.03 | 0.08 | −0.15 | 0.10 | −0.10 | 0.13 | −0.25 | 0.16 |
| 6 | −0.34 | 0.08 | −0.50 | 0.11 | −0.50 | 0.14 | −0.60 | 0.16 |

Les biais liés aux groupes de renouvellement affichent un profil relativement uniforme entre les provinces. Ainsi, les biais relatifs au 3e et au 6e panel sont tous hautement significatifs selon la statistique $t$ classique, affichant un signe positif pour le 3e panel et un signe négatif pour le 6e panel. Les biais relatifs au 4e et au 5e panel ont eux aussi le même signe dans toutes les provinces, et sont tous non significatifs.

Pour le 2e panel, tous les biais sont positifs, mais celui de l'Î.-P.-É. est non significatif (l'Î.-P.É. est la province ayant la taille d'échantillon la plus faible). C'est également pour l'Î.-P.-É. que le signe du biais du 1er panel est différent de celui des autres provinces.

Comme il a été indiqué à la section 2.3, il y a plus d'une raison possible pour expliquer l'existence de biais liés aux groupes de renouvellement, mais les résultats présentés dans le tableau incitent fortement à penser que quelle que soit la raison, les biais observés pour certains des panels sont réels, et non pas attribuables uniquement aux erreurs d'échantillonnage. Un inconvénient de la présente analyse, toutefois, est que les biais liés aux groupes de renouvellement sont supposés fixes dans le temps. Un modèle plus souple est proposé à la section 4.

#### 3.2.2 Qualité de l'ajustement

A. TEST DE LA NORMALITÉ

Définissons par $I_{ta}^{(j)} = (y_{ta}^{(j)} - y_{ta|(t-1)}^{(j)})$ l'erreur de prédiction qu'on obtient en prédisant l'estimateur du $j$-ième panel un mois à l'avance, et posons $I'_{ta} = (I_{ta}^{(1)}, \ldots, I_{ta}^{(6)})$. L'utilisation de l'estimation du maximum de vraisemblance dans cette étude suppose que les vecteurs $I_{ta}$ sont des écarts aléatoires normaux (voir la section 2.4). Pour tester cette hypothèse, nous avons calculé la distribution empirique des erreurs de prédiction standardisées $\{(SI)_{ta}^{(j)} = [I_{ta}^{(j)}/\widehat{SD}(I_{ta}^{(j)})], t = (k+1), \ldots, N\}$, et nous l'avons comparée à la distribution normale réduite au moyen de la statistique de test de Kolmogorov-Smirnov. La statistique de test a été calculée pour chacun des six panels et chacune des quatre provinces, et a donné des valeurs $P$ supérieures à 0.15 dans 21 cas sur 24. (Les tests ont été faits au moyen de la procédure PROC UNIVARIATE du logiciel SAS. Dans cette procédure, si la taille de l'échantillon est plus grande que cinquante, comme dans notre cas, les données sont testées à l'égard d'une distribution normale avec moyenne et variance égales à la moyenne et à la variance de l'échantillon.) L'application de la même procédure de test aux erreurs de prédiction standardisées $\{(SI)_{ta} = [I_{ta}/\widehat{SD}(I_{ta})], t = (k+1), \ldots, N\}$ où $I_{ta} = [\sum_{j=1}^{6} I_{ta}^{(j)}/6]$, donne des valeurs $P$ supérieures à 0.15 dans chacune des quatre provinces.

Les estimateurs des écarts-types des erreurs de prédiction ayant servi aux tests sont ceux produits par le filtre de Kalman, sans prise en considération de la composante de variance résultant de l'estimation des paramètres (voir la section 2.5). Cette dernière composante est négligeable même pour l'Î.-P.-É., qui compte les échantillons ayant la taille la plus faible parmi les quatre provinces. Nous reviendrons sur cette observation à la section 3.4.

## B. ERREURS DE PRÉDICTION AFFECTANT DIFFÉRENTS PRÉDICTEURS

Le tableau 3 présente des statistiques sommaires comparant le profil des erreurs de prédiction entre les quatre provinces, selon les résultats fournis par trois différents ensembles d'estimateurs des vecteurs d'états: 1) Les estimateurs obtenus en vertu des modèles séparés (MS) définis par (2.2), (2.5) et (2.6); 2) les estimateurs obtenus en vertu du modèle conjoint (MC) défini par (2.2), (2.5), (2.6) et (2.10); 3) les estimateurs obtenus avec les contraintes imposées au modèle conjoint (ROB) pour les fins de la robustesse (2.11). Nous définissons ci-dessous les statistiques sommaires en utilisant encore une fois la notation $I_{ta}^{(j)} = (y_{ta}^{(j)} - \hat{y}_{ta|(t-1)}^{(j)})$ pour l'erreur de prédiction obtenue lorsqu'on prédit l'estimateur du $j$-ième panel un mois à l'avance.

$MB_a = \sum_{t=k+1}^{N} (\sum_{j=1}^{6} I_{ta}^{(j)}/6)/(N-k)$ - biais moyen de la prédiction de l'estimateur moyen de l'enquête $\bar{y}_{ta} = \sum_{j=1}^{6} y_{ta}^{(j)}/6$.

$MAB_a = \sum_{j=1}^{6} | \sum_{t=k+1}^{N} I_{ta}^{(j)}/(N-k) | /6$ - biais absolu moyen de la prédiction des estimateurs propres aux panels.

$SQRE_a = \{\sum_{t=k+1}^{N} [1/6 \sum_{j=1}^{6} I_{ta}^{(j)}/\bar{y}_{ta})]^2/(N-k)\}^{1/2}$ - racine carrée de l'erreur de prédiction relative carrée moyenne quand on prédit l'estimateur moyen de l'enquête.

Les statistiques sommaires ci-dessus sont présentées séparément pour la période d'observation de l'échantillon allant de juillet 1983 à décembre 1988, et pour la période postérieure allant de janvier 1989 à décembre 1989. Dans ce dernier cas, les données ont été ajoutées un mois à la fois, c'est-à-dire que pour prédire l'estimateur de l'enquête de février 1989, par exemple, nous avons utilisé les données observées jusqu'à janvier 1989, et ainsi de suite.

Les principales conclusions qui découlent du tableau 3 sont les suivantes:

1) Les résultats obtenus dans le cas des trois ensembles de prédicteurs sont en général très semblables, ce qui indique que pour les données analysées, l'utilisation du modèle conjoint ne produit qu'une légère amélioration par rapport à l'utilisation des modèles séparés, et qu'il n'y a pas de variations brusques du niveau des séries au cours des années examinées.

2) Les erreurs de prédiction des estimateurs de l'enquête sont faibles tant à l'intérieur qu'à l'extérieur de la période d'observation de l'échantillon, ce qui laisse supposer un bon ajustement du modèle. Notons que sauf pour l'Î.-P.-É., les erreurs de prédiction relatives mesurées par la statistique $SQRE_a$ sont toutes inférieures à 7%.

3) Les biais des erreurs de prédiction dans la période postérieure à l'observation de l'échantillon sont plus élevés que dans la période d'observation de l'échantillon, les écarts étant particulièrement prononcés au Nouveau-Brunswick et à l'Î.-P.-É. Ce résultat à lui seul pourrait laisser croire à une certaine défaillance du modèle au cours de l'année 1989. Toutefois, l'examen des erreurs de prédiction mensuelles des panels dans les quatre provinces pour cette année-là (non présentées dans le tableau) indique que même si les erreurs sont en général positives, les biais relativement élevés résultent principalement d'une ou deux erreurs extrêmes, qui ont un effet important sur les statistiques sommaires moyennes, puisqu'il n'y a que 12 mois de données. Il est à signaler, en outre, que les taux de chômage estimés dans les quatre provinces au cours de l'année 1989 se situent entre 0.11 et 0.18, de telle sorte qu'un biais de prédiction de .005, ou même de .009 comme dans le cas de l'Î.-P.-É., n'est pas élevé. De toute évidence, le modèle peut être modifié pour tenir compte de ces biais, si ceux-ci persistent avec l'ajout de nouvelles données. Par ailleurs, rappelons que l'analyse qui précède ne concerne que le biais des erreurs de prédiction, puisque le biais des estimateurs fondés sur le modèle des valeurs de la population correspondantes est régi par les contraintes imposées pour assurer la robustesse (2.11).

**Tableau 3**

Erreurs de prédiction pour les quatre provinces - statistiques sommaires (× 100)

| | Nouvelle-Écosse | | | Nouveau-Brunswick | | | Terre-Neuve | | | Île-du-Prince-Édouard | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MS | MC | ROB | MS | MC | ROB | MS | MC | ROB | MS | MC | ROB |
| | | | | | | 7.83 - 12.88 | | | | | | |
| *MB* | -.11 | -.07 | -.06 | -.12 | -.09 | -.06 | -.25 | -.18 | -.08 | .06 | .14 | .15 |
| *MAB* | .12 | .11 | .10 | .14 | .12 | .11 | .29 | .24 | .20 | .20 | .23 | .23 |
| *SQRE* | 5.76 | 5.62 | 5.70 | 5.48 | 5.47 | 5.47 | 7.03 | 6.91 | 6.96 | 9.34 | 9.13 | 9.17 |
| | | | | | | 1.89 - 12.89 | | | | | | |
| *MB* | .14 | .11 | .04 | .47 | .47 | .46 | .36 | .33 | .17 | .84 | .85 | .86 |
| *MAB* | .32 | .32 | .30 | .51 | .51 | .50 | .39 | .37 | .29 | .84 | .85 | .86 |
| *SQRE* | 6.39 | 6.27 | 6.82 | 6.25 | 6.25 | 6.32 | 5.92 | 5.90 | 5.61 | 9.45 | 9.26 | 9.30 |

Compte tenu des résultats très semblables obtenus pour les trois ensembles de prédicteurs examinés, et afin de faire ressortir la performance des contraintes relatives à la robustesse, nous avons délibérément appliqué une correction à la baisse de 33% aux taux de chômage de la période allant de mars 1985 à mars 1987, une correction à la baisse de 25% des taux de la période d'avril 1987 à novembre 1988 et une correction à la hausse de 33% des taux de la période de décembre 1988 à décembre 1989. Ces opérations ont pour effet d'introduire des mouvements soudains des données aux mois $t = 39$, $t = 64$ et $t = 84$. La figure 3 présente les erreurs de prédiction un mois d'avance globales (APE), $I_t^a = \sum_{a=1}^{4} w_{ta} [\sum_{j=1}^{6} (y_{ta}^{(j)} - \hat{y}_{ta|(t-1)}^{(j)})/6]$, selon le modèle conjoint avec et sans les contraintes relatives à la robustesse, ainsi que selon les modèles séparés.

Il ressort clairement de la figure 3 que lorsque les contraintes sont imposées, les APE pour les périodes qui suivent les trois mois affichant des mouvements soudains sont moins élevées que les APE obtenues sans les contraintes. Par exemple, en mars 1985 ($t = 39$), les APE sont très élevées en valeur absolue avec et sans les contraintes, ce qui est évident étant donné que les prédicteurs ne sont fondés que sur les données allant jusqu'à février 1985. Les APE correspondant aux prédicteurs robustes reviennent toutefois à leur niveau normal beaucoup plus vite que les APE relatives aux prédicteurs non robustes. Une situation semblable peut être observée dans les deux autres périodes. Un autre résultat notable du graphique est que dans les périodes qui suivent les mois affichant des mouvements soudains, le modèle conjoint a une meilleure performance que les modèles séparés, même sans l'imposition des contraintes visant la robustesse. Par conséquent, en regroupant l'information des différentes provinces, le modèle conjoint s'adapte plus rapidement au nouveau niveau de la série. Pour d'autres illustrations de la performance des contraintes relatives à la robustesse, voir Pfeffermann et Burck (1990).

## C. COMPARAISONS AVEC LES ESTIMATEURS PRODUITS PAR LA PROCÉDURE X-11

À titre d'évaluation finale du bien-fondé du modèle, nous comparons les estimations des effets saisonniers et des niveaux de la tendance produits par le modèle aux estimations produites par la procédure X-11 (Dagum 1980). Cette dernière est reconnue pour sa moins grande dépendance à l'égard des hypothèses particulières du modèle. Elle est la procédure couramment utilisée partout dans le monde pour la désaisonnalisation. La figure 4 présente les effets saisonniers moyens pour les quatre provinces obtenus avec la procédure X-11 et en vertu du modèle. La figure 5 montre les estimations correspondantes des niveaux de la tendance. Les moyennes sont calculées à l'aide des poids ($w_{ta}$) employés dans les analyses précédentes. Les estimations fondées sur le modèle indiquées dans les deux figures sont les estimations lissées qui, comme celles de la procédure X-11, utilisent toutes les données de la période d'observation de l'échantillon.

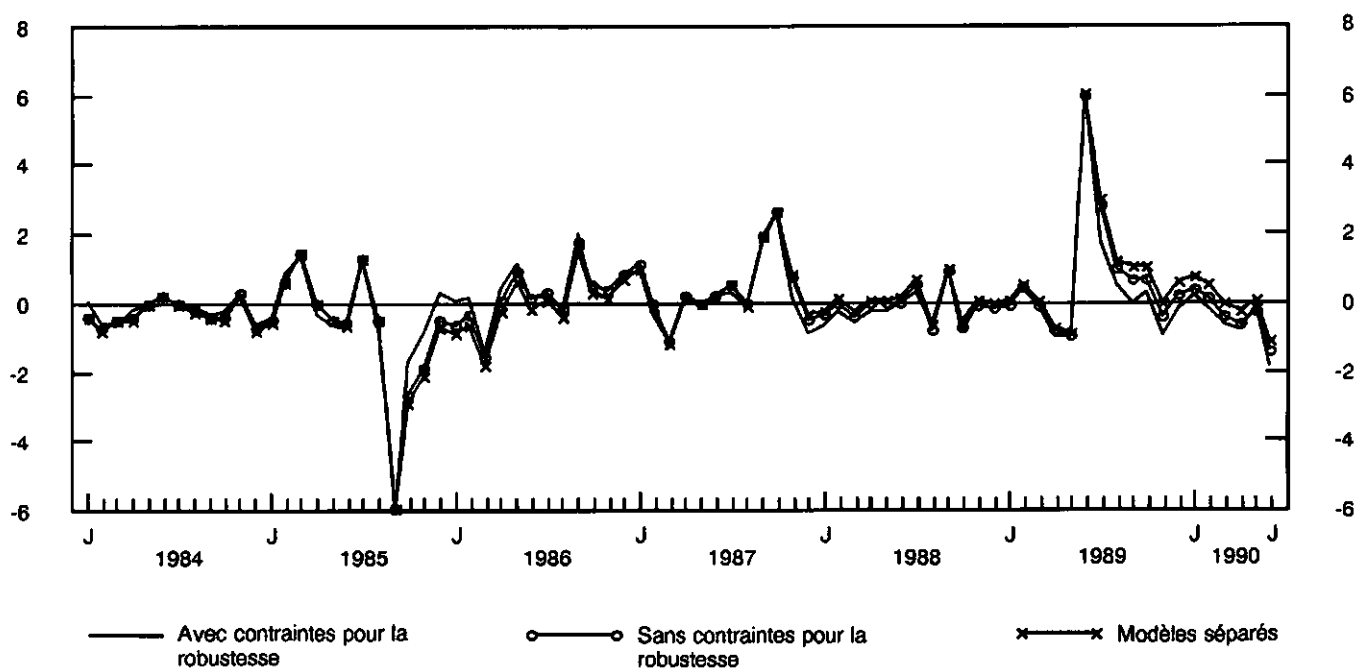


**Figure 3.** Erreurs de prédiction un mois d'avance globales pour les trois ensembles de prédicteurs (× 100) – Données contaminées

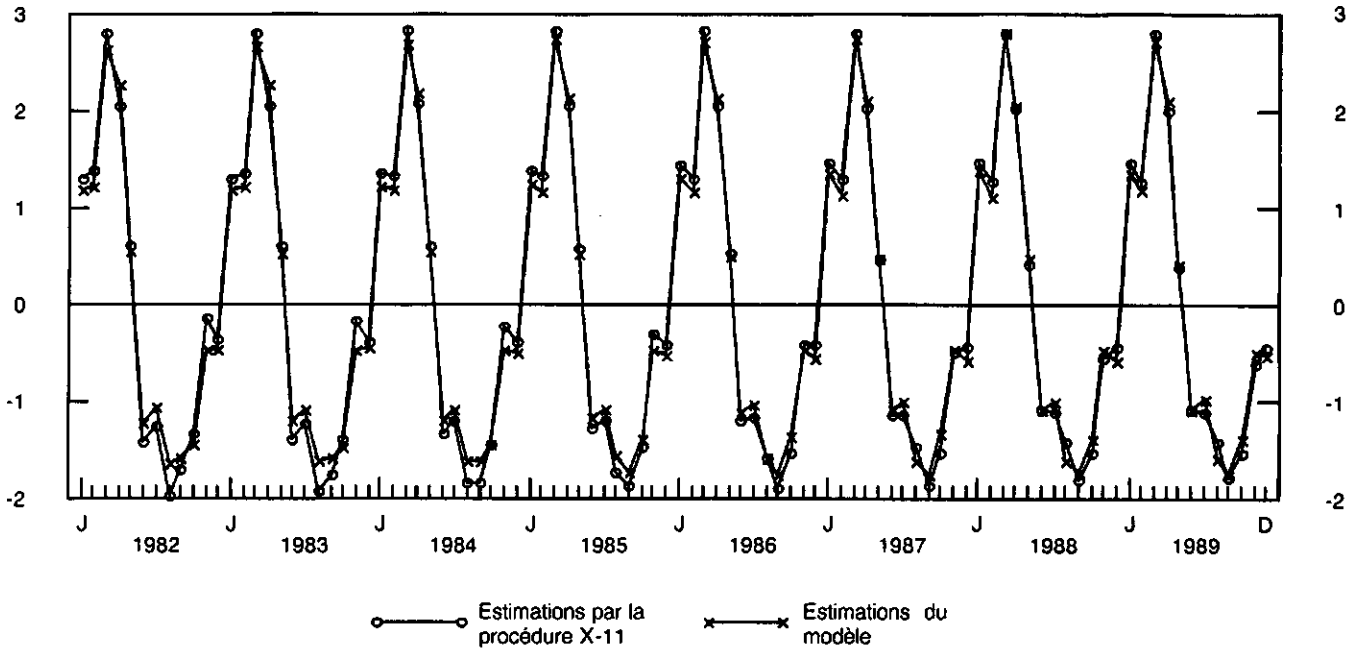


**Figure 4.** Moyennes pondérées des effets saisonniers par la procédure X-11 et selon le modèle (× 100)

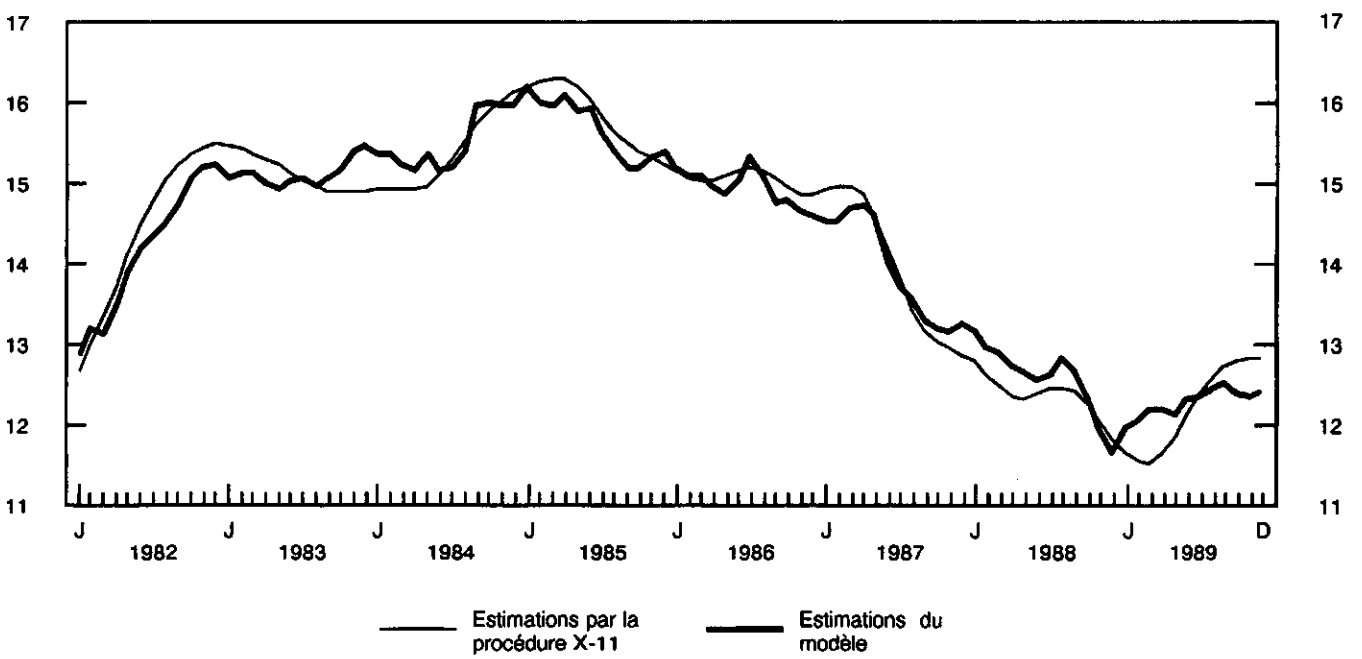


**Figure 5.** Moyennes pondérées des niveaux de la tendance obtenues par la prodédure X-11 et selon le modèle

Comme on peut le voir, les effets saisonniers produits par les deux méthodes sont très voisins. Les estimations des niveaux de la tendance sont aussi voisines, mais la courbe de la tendance X-11 est plus lisse que la courbe du modèle. Une correspondance étroite semblable entre la procédure X-11 et le modèle est obtenue pour chacune des quatre provinces séparément, y compris, notamment, l'Î.-P.-É., où les tailles des échantillons sont relativement faibles.

### 3.3 Comparaison entre les estimateurs fondés sur le plan et les estimateurs fondés sur le modèle

Il est mentionné, dans l'introduction, qu'une des principales raisons qui incitent à modéliser les estimateurs bruts de l'enquête est que le modèle produit des estimations des valeurs de la population qui, au moins pour les petites régions, sont plus exactes (lorsque le modèle tient) que celles de l'enquête. Nous avons calculé les deux ensembles d'estimations pour les quatre provinces et constaté que, comme prévu, les estimations produites par les deux méthodes se comportent de façon très semblable, mais que les estimateurs fondés sur le plan sont moins stables, affichant généralement des pointes et des creux plus prononcés. Un important aspect de la comparaison entre les deux ensembles d'estimations, c'est leur aptitude à estimer les variations annuelles des valeurs de la population. Une telle comparaison a l'avantage de ne pas être obscurcie par les effets saisonniers. La figure 6 montre les résultats obtenus pour l'Î.-P.-É. Les estimations fondées sur le modèle sont les valeurs lissées du modèle conjoint qui utilisent toutes les données de tous les mois. Comme on peut le voir, les estimations produites par le modèle sont beaucoup plus stables et ne varient que légèrement d'un mois à l'autre, comparativement aux estimations fondées sur le plan. La figure 7 montre les erreurs-types (E.-T.) des estimateurs des taux de chômage pour l'Î.-P.-É., calculées en vertu du plan (valeurs lissées, voir la figure 2) et en vertu du modèle conjoint. Sont également indiquées l'erreur-type résultant de l'ajustement du modèle séparé défini par (2.2), (2.5) et (2.6) et l'erreur-type correspondante après prise en considération du fait que des estimations des paramètres sont utilisées à la place des valeurs réelles inconnues. Voir la section 2.5 pour plus de détails. (Nous avons calculé cette dernière erreur-type uniquement pour le modèle séparé afin d'économiser du temps d'ordinateur.)

Il y a trois aspects importants qui ressortent des graphiques:

1) Les erreurs-types des estimateurs fondés sur le modèle, en vertu du modèle conjoint, ne sont que légèrement inférieures aux erreurs-types obtenues pour le modèle séparé, mais beaucoup plus faibles que celles des estimateurs de l'enquête.

2) Les erreurs-types des estimateurs fondés sur le modèle suivent le même profil que les erreurs-types des estimateurs de l'enquête, ce qui est une conséquence directe de la prise en considération, dans la définition du modèle, des changements des variances des erreurs de l'enquête au fil du temps. Voir la section 2.3 pour plus de détails.

3) La prise en considération du fait qu'on utilise des valeurs estimées des paramètres dans le calcul de l'erreur-type des estimateurs fondés sur le modèle n'a qu'un effet minime sur l'erreur-type calculée. Rappelons que l'Î.-P.-É. est la province ayant les plus faibles tailles d'échantillon. Pour les autres provinces, l'effet de la prise en considération de l'utilisation d'estimations des paramètres est encore plus faible.

## 4. RÉSUMÉ

Le présent article montre que des données recueillies en vertu d'un plan de sondage complexe, comprenant plusieurs degrés de sélection et des groupes de renouvellement, peuvent être modélisées avec succès par un modèle relativement simple. Le modèle comprend deux parties: le modèle du recensement, qui représente les valeurs de la population, et le modèle des erreurs de l'enquête, qui décrit les relations qui existent dans les séries chronologiques des erreurs de l'enquête. L'utilisation du modèle donne des estimateurs plus exacts des valeurs de la population et de leurs composantes comme la tendance et la saisonnalité, et elle permet d'estimer l'erreur-type de ces estimateurs d'une façon relativement simple. On peut modifier les équations du modèle de manière à assurer la robustesse des estimateurs fondés sur le modèle et ainsi se protéger contre des défaillances possibles du modèle.

Le modèle utilisé dans cet article peut être prolongé dans diverses directions. Le premier effort devrait consister à appliquer simultanément le modèle à un plus grand nombre de provinces ou à d'autres petites régions, pour s'assurer que les estimateurs globaux de l'échantillon $\sum_{a=1}^{A} w_{ta}\,\bar{y}_{ta}$ sont suffisamment proches des valeurs correspondantes de la population. Voir l'analyse présentée à la section 2.7. L'incorporation au modèle d'un mécanisme de détection des valeurs aberrantes, afin de mieux mesurer la performance et l'adéquation du modèle, est une autre addition valable.

Deux autres extensions pourraient consister à abandonner l'hypothèse d'une variance constante pour le terme d'erreur $\epsilon_t$ dans le modèle du recensement et à laisser varier dans le temps les biais liés aux groupes de renouvellement. La première extension découle de l'observation faite à la section 3.1 selon laquelle les variances des erreurs de l'enquête sont l'objet d'effets saisonniers, affichant un profil saisonnier semblable à celui des estimations brutes. L'ajustement des équations (2.4) dans les quatre provinces révèle également l'existence d'une légère tendance dans les variances qui, encore une fois, est semblable à celle des estimations brutes de l'enquête. Ainsi, les variances des erreurs de l'enquête semblent dépendre de la grandeur des estimateurs de l'enquête, ce qui laisse croire que les variances $\sigma_t^2 = V(\epsilon_t)$ changent avec le niveau des valeurs de la population. Comme première approximation, on pourrait supposer que $\sigma_t^2$ est proportionnel à la variance correspondante de l'erreur de l'enquête.

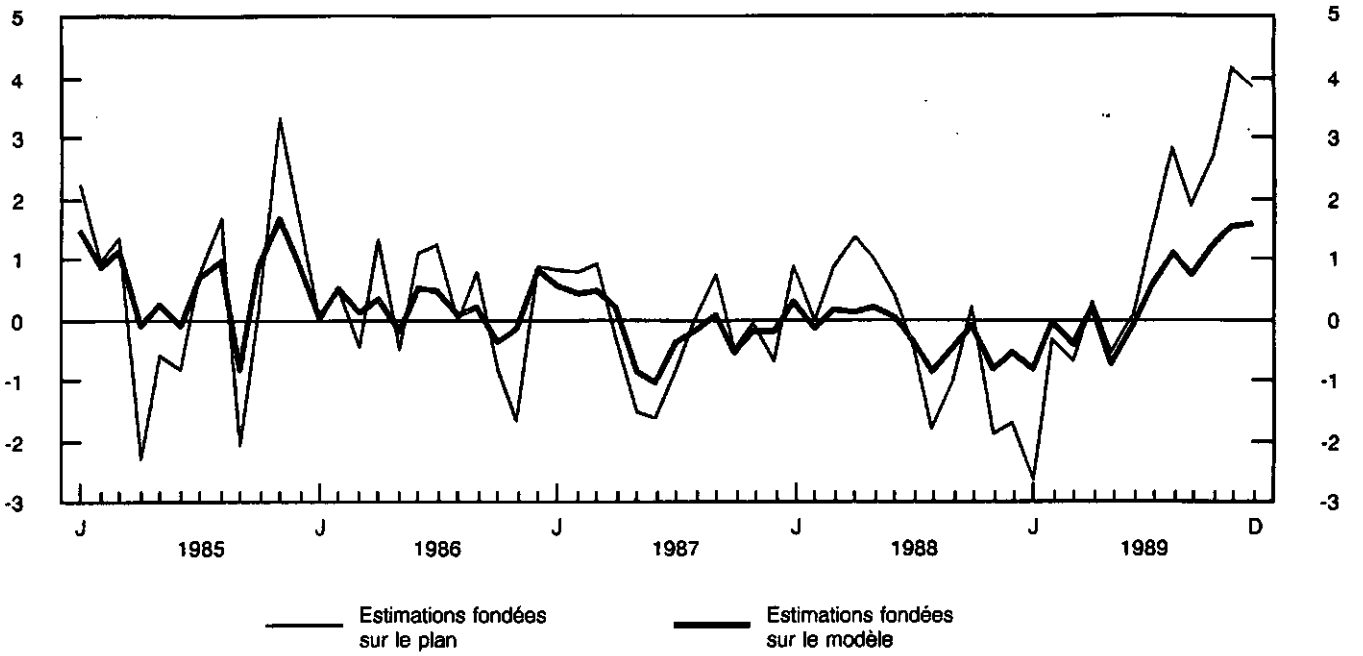


**Figure 6.** Variations annuelles des estimations fondées sur le plan et des estimations fondées sur le modèle des taux de chômage de l'Î.-P.-É. (× 100)

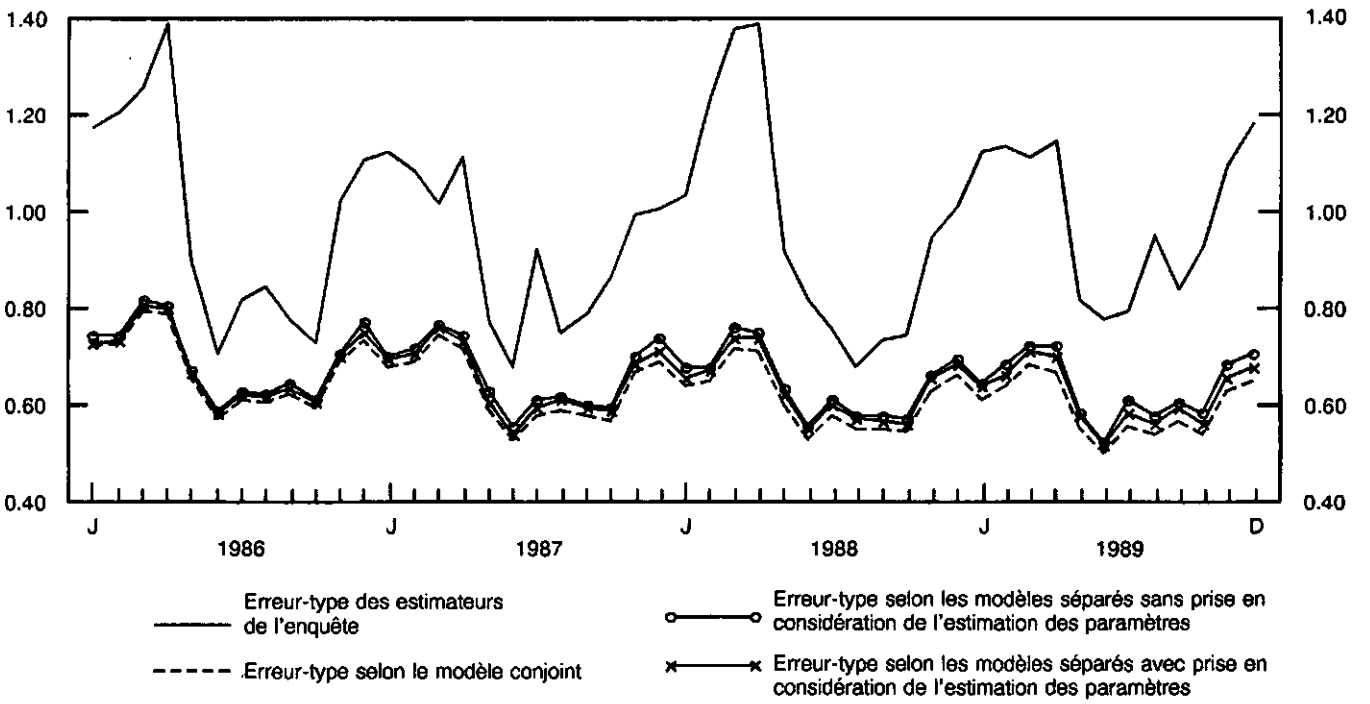


**Figure 7.** Erreurs-types des estimateurs de l'enquête et des estimateurs fondés sur le modèle avec et sans prise en considération de l'estimation des paramètres (× 100) - Île-du-Prince-Édouard

Le fait de laisser varier dans le temps les biais liés aux groupes de renouvellement est une extension naturelle du modèle, si l'on tient compte du fait que les moyennes des valeurs de la population varient en fonction du temps. Toutefois, la modélisation de l'évolution des biais liés aux groupes pourrait poser des difficultés, en raison d'éventuels problèmes d'identifiabilité touchant les modèles représentant la tendance et les effets saisonniers. Voir l'analyse présentée dans Pfeffermann (1991).

Les deux dernières extensions sont importantes et méritent d'être étudiées, mais notre expérience des données sur l'emploi nous porte à croire qu'elles auront un effet très minime sur les estimateurs du modèle.

## REMERCIEMENTS

Les auteurs désirent remercier les arbitres pour leurs commentaires et leurs suggestions utiles. Les travaux relatifs à cette étude ont été réalisés pendant que le premier auteur se trouvait à Statistique Canada dans le cadre de son programme de bourse de recherche.

## BIBLIOGRAPHIE

ANSLEY, C.F., et KOHN, R. (1986). Predicted mean square error for state-space models with estimated parameters. *Biometrika*, 73, 467-473.

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

CHANG, I., TIAO, G.C., et CHEN, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, 30, 193-204.

DAGUM, E.B. (1980). *La méthode de désaisonnalisation* X-11 *ARMMI*. N° 12-564 au catalogue, Statistique Canada, Ottawa, Ontario, K1A 0T6.

HAMILTON, J.D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 387-397.

HARRISON, P.J., et STEVENS, C.F. (1976). Bayesian forecasting (avec discussion). *Journal of the Royal Statistical Society*, Séries B, 38, 205-247.

HARVEY, A.C. (1984). A unified view of statistical forecasting procedures (avec discussion). *Journal of Forecasting*, 3, 245-275.

HARVEY, A.C., et TODD, P.H.J. (1983). Forecasting economic time series with structural and Box-Jenkins models (avec discussion). *Journal of Business and Economic Statistics*, 1, 299-315.

LEE, H. (1990). Estimation des coefficients de corrélation de panel pour l'Enquête sur la population active du Canada. *Techniques d'enquête*, 16, 297-306.

MARAVALL, A. (1985). On structural time series models and the characterization of components. *Journal of Business and Economic Statistics*, 3, 350-355.

MORRIS, N.D., et PFEFFERMANN, D. (1984). A Kalman filter approach to the forecasting of monthly time series affected by moving festivals. *Journal of Time Series*, 5, 255-268.

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

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

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

SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). Méthodologie de l'enquête sur la population active du Canada. N° 71-526 au catalogue, Statistique Canada, Ottawa, Ontario, K1A 0T6.

TILLER, R.B. (1992). Time series modeling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 175–182
Statistique Canada



# Estimation d'un modèle d'étalonnage à biais multiplicatif constant par la méthode du maximun de vraisemblance et application

IJAZ U.H. MIAN et NORMAND LANIEL[1]

RÉSUMÉ

Nous considérons le cas de l'estimation d'un modèle d'étalonnage non linéaire par la méthode du maximum de vraisemblance, tel que le présentent Laniel et Fyfe (1989; 1990). Ce modèle tient compte des biais et des erreurs d'échantillonnage rattachés à la série originale. Comme on ne peut exprimer les estimateurs du maximum de vraisemblance des paramètres du modèle sous une forme analytique fermée, nous examinons deux méthodes itératives permettant de calculer les estimations du maximum de vraisemblance. Nous donnons aussi les expressions en forme analytique fermée pour les variances et les covariances asymptotiques des séries étalonnées et des valeurs ajustées. Pour illustrer les méthodes, nous nous servons de données publiées sur le commerce de détail au Canada.

MOTS CLÉS: Autocorrélations; modèle à biais; moindres carrés généralisés; erreurs d'échantillonnage.

## 1. INTRODUCTION

Les méthodes d'étalonnage servent très souvent à améliorer les estimations tirées d'enquêtes infra-annuelles; on utilise à cette fin les estimations correspondantes tirées d'enquêtes annuelles et que l'on appelle des estimations repères. Cette opération se traduit généralement par une réduction du biais et de la variance des estimations infra-annuelles. Par exemple, les estimations tirées des enquêtes annuelles sur le commerce de détail peuvent servir à améliorer les estimations mensuelles du commerce de détail. Les estimations infra-annuelles renferment souvent un biais à cause des lacunes de la base de sondage au point de vue de la couverture. Le sous-dénombrement est attribuable au fait que les nouvelles entreprises tardent à être incluses dans la base de sondage et à l'absence, dans cette base, des entreprises sans salariés. En outre, la variance des estimations infra-annuelles est souvent plus élevée que celle des estimations annuelles correspondantes et on observe des covariances d'échantillonnage pour les estimations de périodes différentes à cause du chevauchement des échantillons. En revanche, on peut supposer les estimations annuelles non biaisées parce que, dans la pratique, les bases de sondage correspondantes présentent peu de lacunes au point de vue de la couverture. On peut trouver des discussions détaillées sur l'étalonnage dans Laniel et Fyfe (1989; 1990), Cholette (1987; 1988) et d'autres ouvrages.

La littérature statistique présente plusieurs méthodes d'étalonnage des séries chronologiques. Dans une perspective de minimisation quadratique, Denton (1971) propose plusieurs méthodes pour étalonner une série temporelle unique. Cholette (1984), à son tour, propose une variante de la méthode proportionnelle d'ordre un de Denton, par laquelle il supprime la condition de départ dans le but d'éviter les effets transitoires. Les hypothèses que posent les auteurs ont très peu de chances d'être vérifiées avec la plupart des séries économiques. En effet, les modèles que ces auteurs élaborent supposent que le biais rattaché aux estimations infra-annuelles suit une marche aléatoire et que les observations infra-annuelles et annuelles ne renferment aucune erreur d'échantillonnage. En règle générale, les estimations proviennent d'enquêtes par sondage et sont donc exposées à l'erreur d'échantillonnage.

Hillmer et Trabelsi (1987) ont proposé une autre approche pour l'étalonnage, qui repose sur un modèle ARMMI (voir, par exemple, Box et Jenkins, 1976). Bien que cette approche tienne compte des covariances d'échantillonnage des estimations infra-annuelles et annuelles, elle fait abstraction du biais dans les estimations infra-annuelles. Cholette et Dagum (1989) ont modifié l'approche de Hillmer et Trabelsi en remplaçant le modèle ARMMI par un modèle d'"intervention". Cette variante permet la modélisation d'effets systématiques dans les séries temporelles mais elle présente néanmoins les mêmes lacunes que le modèle de Hillmer et Trabelsi (Laniel et Fyfe 1990).

Afin de suppléer les lacunes mentionnées plus haut, Laniel et Fyfe (1989; 1990) ont proposé un modèle d'étalonnage non linéaire pour les mesures de niveau. Ils ont élaboré un algorithme complexe pour calculer les estimations par les moindres carrés généralisés (MCG) (et les covariances asymptotiques correspondantes) des paramètres du modèle. Ce modèle tient compte des covariances d'échantillonnage des estimations infra-annuelles et annuelles et peut être utilisé lorsque les données repères proviennent soit de recensements ou d'échantillons chevauchants annuels. Il suppose aussi l'existence d'un biais (relatif) multiplicatif constant dans les estimations infra-annuelles de niveau. D'autres modèles d'étalonnage à biais multiplicatif constant ont été proposés par Cholette (1992) et Laniel et Mian (1991). Cholette suppose un modèle dans

[1] Ijaz U.H. Mian et Normand Laniel, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6, Canada.

lequel non seulement le biais mais aussi les erreurs sont multiplicatives. L'auteur se sert de la théorie des MCG pour calculer les estimations des paramètres du modèle après avoir soumis ce modèle à une transformation logarithmique. Laniel et Mian (1991) ont défini un algorithme pour calculer les estimations du maximum de vraisemblance d'un modèle d'étalonnage à biais multiplicatif constant avec données-repères mixtes (c.-à-d. un mélange de repères fermes et de repères non fermes). Les données-repères fermes sont en l'occurrence des estimations tirées d'un recensement (c.-à-d. des estimations à variance nulle) tandis que les données-repères non fermes sont des estimations fondées sur un échantillon. L'hypothèse de l'existence d'un biais multiplicatif constant se vérifiera dans la pratique si la fréquence de mise à jour de la base de sondage est relativement stable, c'est-à-dire si le taux de sous-dénombrement pour cette base varie très peu d'une année à l'autre. Cette hypothèse implique aussi que le rapport moyen des valeurs de la variable étudiée d'une période à l'autre est le même pour les entreprises qui sont incluses dans la base de sondage et celles qui ne le sont pas. La nature du biais rattaché aux estimations infra-annuelles peut varier d'une série temporelle à l'autre. Cholette et Dagum (1991) ont proposé une méthode d'étalonnage qui suppose l'existence d'un biais additif constant dans les estimations infra-annuelles.

L'objet de cet article est d'estimer les paramètres du modèle de Laniel et Fyfe par la méthode du maximum de vraisemblance (MV); les résultats s'inspirent du rapport de Mian et Laniel (1991). Le modèle en question est décrit dans la section suivante. Ensuite, nous étudions deux processus itératifs qui permettent de calculer les estimations du maximum de vraisemblance (EMV) des paramètres du modèle. Dans la section 4, nous donnons les expressions en forme analytique fermée pour les covariances asymptotiques des estimateurs des paramètres du modèle et des valeurs ajustées. À titre d'exemple, nous nous servons des données publiées par Statistique Canada sur le commerce de détail au Canada.

## 2. MODÈLE A BIAIS MULTIPLICATIF CONSTANT (MBMC)

Afin de répondre aux exigences des enquêtes économiques en matière d'étalonnage, Laniel et Fyfe (1989; 1990) ont proposé le modèle à biais multiplicatif constant (MBMC) défini ci-dessous. Ce modèle suppose que les estimations infra-annuelles biaisées $y_t$ répondent à la relation

$$y_t = \beta\theta_t + a_t, \quad t = 1, 2, \ldots, n \tag{2.1}$$

et que les estimations annuelles non biaisées $z_T$ répondent à la relation

$$z_T = \sum_{t \in T} \theta_t + b_T, \quad T = 1, 2, \ldots, m, \tag{2.2}$$

où les indices $t$ et $T$ désignent respectivement les périodes infra-annuelles et les périodes annuelles, $\theta_t$ est le paramètre (inconnu) infra-annuel fixe, $\beta$ est un paramètre (inconnu) de biais constant rattaché à $y_t$, et $a_t$ et $b_T$ sont les erreurs d'échantillonnage rattachées à $y_t$ et à $z_T$ respectivement. Le modèle ci-dessus est un modèle de type hybride (mixte), dans lequel le biais est multiplicatif mais les erreurs, additives.

Avant d'aller plus loin, définissons les vecteurs colonnes $y = (y_1, y_2, \ldots, y_n)'$, $z = (z_1, z_2, \ldots, z_m)'$, $a = (a_1, a_2, \ldots, a_n)'$, $b = (b_1, b_2, \ldots, b_m)'$, et $\Theta = (\theta_1, \theta_2, \ldots, \theta_n)'$. Le MBMC, défini par les équations (2.1) et (2.2), peut être réexprimé sous la forme

$$\begin{aligned} w &= X_\beta \Theta + u \\ &= X_\Theta \beta + X_D \Theta + u, \end{aligned} \tag{2.3}$$

où

$$X_\beta = (\beta I_n : D')', \quad X_\Theta = (\Theta' : 0')', \quad X_D = (0' : D')',$$
$$w = (y' : z')', \qquad u = (a' : b')', \qquad D = (d_{Tt}), \tag{2.4}$$

$I_n$ étant une matrice unité d'ordre $n$, $\mathbf{0}$ étant un vecteur ou une matrice nuls d'ordre approprié et $d_{Tt}$ étant une fonction indicatrice égale à 1 pour $t \in T$ et égale à 0 dans le cas contraire. On suppose que les vecteurs d'erreurs d'échantillonnage $a$ et $b$ suivent une distribution normale multidimensionnelle telle que $a \sim MN(\mathbf{0}, V_{aa})$ et $b \sim MN(\mathbf{0}, V_{bb})$. De plus, dans le cas général, $a$ et $b$ sont corrélés, ce qui signifie que $\mathrm{Cov}(a,b) = V_{ab} = V'_{ba} \neq \mathbf{0}$. Nous allons voir dans la section suivante que, pour ce modèle, l'estimateur du maximum de vraisemblance (MV) de $\Theta$ et de $\beta$ est identique à l'estimateur des moindres carrés généralisé (MCG) des mêmes paramètres. Par conséquent, l'hypothèse concernant la normalité de $a$ et $b$ n'est nécessaire que pour obtenir la matrice d'information de Fisher (et, donc, les variances) des EMV.

## 3. ESTIMATION PAR LA MÉTHODE DU MAXIMUM DE VRAISEMBLANCE

Suivant le MBMC, la fonction de vraisemblance logarithmique peut s'écrire

$$\ln(L) = -\frac{(n+m)}{2}\ln(2\pi) - \frac{1}{2}\ln|V| - \frac{1}{2}Q, \tag{3.1}$$

où

$$Q = (w - X_\beta\Theta)'V^{-1}(w - X_\beta\Theta) \tag{3.2}$$

et

$$V = \begin{pmatrix} V_{aa} & V_{ab} \\ V_{ba} & V_{bb} \end{pmatrix}.$$

On peut obtenir les estimations du maximum de vraisemblance des paramètres du modèle $\Theta$ et $\beta$ (si l'on suppose $V$ connu) en maximisant la fonction de vraisemblance

logarithmique (3.1) ou, ce qui revient au même, en minimisant le terme quadratique $Q$ (3.2). Pour ce modèle en particulier, les estimateurs du MV et les estimateurs des MCG des paramètres du modèle sont les mêmes et nous n'en ferons la distinction que si le besoin s'en fait sentir. Si nous calculons les dérivées partielles d'ordre un de $\ln(L)$ par rapport à $\boldsymbol{\Theta}$ et à $\beta$ respectivement et si nous posons le résultat égal à zéro, nous avons

$$\frac{\partial \ln(L)}{\partial \boldsymbol{\Theta}} = X_\beta' V^{-1}(w - X_\beta \boldsymbol{\Theta}) = 0,$$
$$\frac{\partial \ln(L)}{\partial \beta} = X_\Theta' V^{-1}(w - X_\beta \boldsymbol{\Theta}) = 0. \tag{3.3}$$

Puisque $\mathrm{E}(w) = X_\beta \boldsymbol{\Theta}$ selon le modèle (2.3), les équations ci-dessus sont des équations d'estimation au sens de Godambe (1960) et elles sont non biaisées du point de vue de l'information. Il est intéressant de noter que $X_\beta' V^{-1}$ et $X_\Theta' V^{-1}$ ne dépendent pas de $w$, de sorte que les équations (3.3) convergent vers zéro et ont, par conséquent, des racines consistantes pourvu que $\mathrm{E}(w) = X_\beta \boldsymbol{\Theta}$. Autrement dit, même si $V$ dans les équations ci-dessus est remplacé par une de ses estimations consistantes, les équations donneront des estimations consistantes des vecteurs $\boldsymbol{\Theta}$ et $\beta$. Il convient aussi de noter que ces équations sont non linéaires par rapport aux paramètres à estimer et qu'on ne peut obtenir d'expression formelle pour les estimateurs de $\boldsymbol{\Theta}$ et de $\beta$. On peut donc recourir à une méthode itérative comme la méthode bien connue de Fisher-Newton-Raphson (aussi appelée méthode "de la fonction de caractérisation" de Fisher) pour calculer les estimations. La section 4 donne les éléments de la matrice d'information de Fisher espérée nécessaire à l'application de la méthode de Fisher-Newton-Raphson.

Une autre manière de calculer les EMV des paramètres du modèle est de résoudre successivement les équations d'estimation (3.3). Si nous résolvons la première expression de (3.3), l'estimateur de $\boldsymbol{\Theta}$, comme fonction de $\beta$, est défini

$$\hat{\boldsymbol{\Theta}} \equiv \hat{\boldsymbol{\Theta}}(\beta) = (X_\beta' V^{-1} X_\beta)^{-1} X_\beta' V^{-1} w. \tag{3.4}$$

De même, si nous résolvons la seconde expression de (3.3), l'estimateur de $\beta$, comme fonction de $\boldsymbol{\Theta}$, est défini

$$\hat{\beta} \equiv \hat{\beta}(\boldsymbol{\Theta}) = \left[\boldsymbol{\Theta}' V_{aa.b}^{-1}\left(y - V_{ab} V_{bb}^{-1}(z - D\boldsymbol{\Theta})\right)\right] / \left[\boldsymbol{\Theta}' V_{aa.b}^{-1} \boldsymbol{\Theta}\right], \tag{3.5}$$

où

$$V_{aa.b} = V_{aa} - V_{ab} V_{bb}^{-1} V_{ba}.$$

On obtient les estimations du MV de $\boldsymbol{\Theta}$ et de $\beta$ en calculant successivement les équations (3.4) et (3.5) jusqu'à convergence. L'avantage de cette méthode par rapport à la méthode de Fisher-Newton-Raphson est qu'elle est facile à appliquer. Cependant, la convergence est ordinairement très lente pour ce genre d'algorithme. Dans la section 6, nous comparerons ces deux méthodes dans le but de vérifier leur vitesse de convergence.

Une fois que l'on connaît les estimations du MV des paramètres du modèle, on peut déterminer les valeurs infra-annuelles ajustées $\hat{y} = \hat{\beta}\hat{\boldsymbol{\Theta}}$ et les valeurs annuelles ajustées $\hat{z} = D\hat{\boldsymbol{\Theta}}$.

**Estimation initiale de $\boldsymbol{\Theta}$ et $\beta$**

Afin d'établir une estimation initiale pour $\beta$, disons $\hat{\beta}_0$, réécrivons le modèle (2.3) comme suit:

$$w^* = X_\Theta^* \beta + u^*,$$

où $w^* = ((Dy)':(z - D\boldsymbol{\Theta})')'$, $X_\Theta^* = \left((D\boldsymbol{\Theta})':\mathbf{0}'\right)'$ et $u^* = \left((Da)':b'\right)'$. L'EMV de $\beta$ est donc défini par l'expression

$$\hat{\beta} = \left[X_\Theta^{*\prime}(V^*)^{-1} w^*\right] / \left[X_\Theta^{*\prime}(V^*)^{-1} X_\Theta^*\right], \tag{3.6}$$

où

$$V^* = \mathrm{Cov}(u^*) = \begin{pmatrix} DV_{aa}D' & DV_{ab} \\ V_{ba}D' & V_{bb} \end{pmatrix}.$$

Si l'on tient compte de ce que $\mathrm{E}(z) = D\boldsymbol{\Theta}$ et si on remplace $D\boldsymbol{\Theta}$ par $z$ dans (3.6), on peut établir une estimation initiale pour $\beta$ au moyen de la formule

$$\hat{\beta}_0 = \left[\begin{pmatrix} z \\ \mathbf{0} \end{pmatrix}'(V^*)^{-1} w^*\right] \Big/ \left[\begin{pmatrix} z \\ \mathbf{0} \end{pmatrix}'(V^*)^{-1}\begin{pmatrix} z \\ \mathbf{0} \end{pmatrix}\right]$$
$$= \left[z'(DV_{aa.b}D')^{-1} Dy\right] / \left[z'(DV_{aa.b}D')^{-1} z\right]. \tag{3.7}$$

On peut établir une estimation initiale pour à l'aide de l'équation (3.4) en remplaçant $\beta$ par $\hat{\beta}_0$.

## 4. COVARIANCES DES ESTIMATEURS

Dans cette section, nous établissons les formules des covariances asymptotiques des estimateurs du MV des paramètres du MBMC en inversant la matrice d'information de Fisher, dénotée par $\boldsymbol{\Omega}$. Les covariances asymptotiques des valeurs annuelles et infra-annuelles ajustées sont calculées au moyen de la méthode delta. Considérons tout d'abord le calcul des covariances des estimateurs du MV de $\boldsymbol{\Theta}$ et $\beta$. Les éléments de $\boldsymbol{\Omega}$ (c.-à-d. les espérances négatives des dérivées partielles d'ordre deux de $\ln(L)$) sont définis par les expressions

$$\boldsymbol{\Omega}_{11} = -\mathrm{E}\left[\frac{\partial^2 \ln(L)}{\partial \boldsymbol{\Theta}\, \partial \boldsymbol{\Theta}'}\right] = X_\beta' V^{-1} X_\beta,$$

$$\Omega_{22} = -\mathrm{E}\left[\frac{\partial^2 \ln(L)}{\partial \beta^2}\right] = \Theta' V_{aa.b}^{-1} \Theta$$

et

$$\Omega_{12} = \Omega_{21}' = -\mathrm{E}\left[\frac{\partial^2 \ln(L)}{\partial \Theta\, \partial \beta}\right] = X_\beta' V^{-1} X_\Theta.$$

Par conséquent, la matrice d'information de Fisher d'ordre $(n + 1) \times (n + 1)$ est définie

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}. \tag{4.1}$$

Si nous inversons $\Omega$ au moyen de l'algèbre de partition des matrices, nous avons

$$\begin{aligned} \mathrm{Cov}(\hat{\Theta}) &= \Omega_{11.2}^{-1}, \\ \mathrm{Var}(\hat{\beta}) &= \Omega_{22.1}^{-1}, \\ \mathrm{Cov}(\hat{\Theta}, \hat{\beta}) &= -\Omega_{11.2}^{-1} \Omega_{12} \Omega_{22}^{-1} \\ &= -\Omega_{11}^{-1} \Omega_{12} \Omega_{22.1}^{-1}, \end{aligned} \tag{4.2}$$

où

$$\begin{aligned} \Omega_{11.2} &= \Omega_{11} - \Omega_{12} \Omega_{22}^{-1} \Omega_{21}, \\ \Omega_{22.1} &= \Omega_{22} - \Omega_{21} \Omega_{11}^{-1} \Omega_{12}. \end{aligned} \tag{4.3}$$

Une fois que la matrice de covariances $\Omega^{-1}$ est obtenue, on peut calculer les covariances asymptotiques des valeurs infra-annuelles ajustées $\hat{y}$ à l'aide de la méthode delta (voir, par ex., Rao 1973). Soit $\Delta$ la matrice des dérivées partielles d'ordre un de $y$ par rapport aux éléments de $(\Theta' : \beta)'$. De toute évidence, la matrice $n \times (n + 1)$ est $\Delta = (\beta I_n : \Theta)$. Si on utilise la méthode delta, la matrice des covariances asymptotiques de $\hat{y}$ est définie par l'expression

$$\mathrm{Cov}(\hat{y}) = \Delta \Omega^{-1} \Delta'. \tag{4.4}$$

En outre, la matrice des covariances des valeurs annuelles ajustées $\hat{z}$, d'après la théorie de la distribution normale centrée réduite à plusieurs variables, est définie par l'expression

$$\mathrm{Cov}(\hat{z}) = D \Omega_{11.2}^{-1} D', \tag{4.5}$$

où $D$ et $\Omega_{11.2}$ sont définies en (2.4) et en (4.3) respectivement.

## 5. ESTIMATION PAR LA MÉTHODE DU MAXIMUM DE VRAISEMBLANCE LORSQUE $V_{ab} = 0$

Dans cette section, nous examinons l'estimation par le maximum de vraisemblance des paramètres du modèle dans le cas particulier où les vecteurs d'erreurs $\boldsymbol{a}$ et $\boldsymbol{b}$ sont non corrélés (c.-à-d., $\mathrm{Cov}(\boldsymbol{a}, \boldsymbol{b}) = V_{ab} = V_{ba}' = \mathbf{0}$). Cette situation est observée normalement dans les enquêtes par sondage lorsque les échantillons infra-annuels et annuels sont tirés de façon indépendante. On peut voir un changement dans les résultats des sections 3 et 4 en substituant $V_{ab} = V_{ba}' = \mathbf{0}$ dans les équations. Par exemple, dans ce cas particulier, les estimateurs du MV de $\Theta$ et de $\beta$, définis en (3.4) et en (3.5), se réduisent à

$$\hat{\Theta}^* \equiv \hat{\Theta}^*(\beta) = \left(\beta^2 V_{aa}^{-1} + D' V_{bb}^{-1} D\right)^{-1} \left(\beta V_{aa}^{-1} y + D' V_{bb}^{-1} z\right)$$

et

$$\hat{\beta}^* \equiv \hat{\beta}^*(\Theta) = \left[\Theta' V_{aa}^{-1} y\right] / \left[\Theta' V_{aa}^{-1} \Theta\right],$$

respectivement. On doit résoudre ces équations successivement pour obtenir les estimations voulues.

De la même manière, les éléments de la matrice d'information de Fisher se ramènent à

$$\begin{aligned} \Omega_{11}^* &= \beta^2 V_{aa}^{-1} + D' V_{bb}^{-1} D, \\ \Omega_{22}^* &= \Theta' V_{aa}^{-1} \Theta, \\ \Omega_{12}^* &= \Omega_{21}^{*\prime} = \beta V_{aa}^{-1} \Theta. \end{aligned}$$

## 6. APPLICATION

Dans cette section, nous présentons un exemple qui utilise des données publiées sur le commerce de détail au Canada; ces données viennent des enquêtes mensuelles et annuelles qu'effectue Statistique Canada sur le commerce de détail. Les estimations mensuelles sur le commerce de détail et les coefficients de variation (CV) correspondants se trouvent dans la publication de Statistique Canada intitulée Commerce de détail (n° 63-005 au catalogue, mensuel). Il existe deux catégories d'estimations mensuelles sur le commerce de détail, à savoir les estimations provisoires et les estimations révisées. Dans l'exemple qui suit, nous avons utilisé les estimations révisées mais non désaisonnalisées (brutes). Comme les CV des estimations révisées ne sont pas disponibles, nous nous sommes servis des CV des estimations provisoires pour calculer une approximation de la variance des estimations mensuelles révisées. Les données utilisées portent sur la période de janvier 1985 à décembre 1988. Une autre difficulté que comporte cet exemple est de déterminer les coefficients d'autocorrélation

entre les estimations mensuelles. Se fondant sur des données mensuelles sur le commerce de détail, Hidiroglou et Giroux (1986) ont calculé les valeurs estimées des coefficients d'autocorrélation aux décalages 1, 3, 6, 9 et 12 pour trois sortes de strate dans plusieurs provinces du Canada. La moyenne des estimations des coefficients d'autocorrélation entre les estimations mensuelles sur le commerce de détail pour les strates formées de la province de l'Ontario et du code 60 de la Classification type des industries (industries des aliments, boissons et médicaments) a servi d'approximation pour les coefficients d'autocorrélation entre les estimations mensuelles sur le commerce de détail. Ces coefficients (moyens) d'autocorrélation, c'est-à-dire $\rho(k)$, figurent dans le tableau 1.

**Tableau 1**

Coefficients d'autocorrélation approximatifs $\rho(k)$ pour les estimations mensuelles sur le commerce de détail

| Décalage $k$ | 1 | 3 | 6 | 9 | 12 |
|---|---|---|---|---|---|
| $\rho(k)$ | 0.970 | 0.940 | 0.918 | 0.914 | 0.962 |

La méthode des moindres carrés ordinaires ainsi qu'un algorithme de McLeod (1975) pour le calcul de coefficients d'autocorrélation théoriques pour des séries temporelles autorégressives à moyennes mobiles ont servi à réviser les coefficients d'autocorrélation observés. On a ajusté un modèle multiplicatif saisonnier ARMA $(1,0)(1,0)_{12}$ aux cinq coefficients observés en minimisant la somme des carrés des écarts entre le coefficient d'autocorrélation observé et le coefficient d'autocorrélation théorique. On a ensuite calculé les coefficients d'autocorrélation pour tous les autres décalages pertinents à l'aide des paramètres du modèle estimés et de l'algorithme de McLeod mentionné plus haut. Étant donné que le modèle ARMA est correct pour les coefficients d'autocorrélation théoriques, cette méthode produit une estimation consistante de la fonction d'autocorrélation. Le tableau 2 donne les coefficients d'autocorrélation approximatifs révisés pour 47 décalages; ces coefficients ont servi à calculer approximativement les covariances des estimations mensuelles sur le commerce de détail par multiplication avec les écarts-types.

Au moment de l'étude des estimations annuelles sur le commerce de détail existaient uniquement pour les années 1985 à 1988. On peut trouver ces estimations dans la publication de Statistique Canada intitulée Commerce de détail annuel (n° 63-223 au catalogue, annuel). Les variances correspondantes ne sont pas publiées; on a donc dû les calculer à l'aide des données de l'enquête. Les covariances entre les estimations mensuelles et annuelles sont nulles parce que les échantillons des enquêtes mensuelle et annuelle sur le commerce de détail ont été tirés indépendamment l'un de l'autre. Comme les estimations annuelles sur le commerce de détail reposent sur des échantillons qui ne sont pas indépendants, leur covariance est non nulle. Or, les opérations normales de traitement des données d'enquête ne permettent pas de connaître les estimations des covariances; il faudrait une étude pour les obtenir. Par conséquent, pour les besoins de notre exemple, nous avons supposé que les covariances des estimations annuelles sur le commerce de détail sont nulles.

**Tableau 2**

Coefficients d'autocorrélation approximatifs révisés $\rho^*(k)$ pour les estimations mensuelles sur le commerce de détail (calculés pour 47 décalages)

| Décalage $k$ | $\rho^*(k)$ | Décalage $k$ | $\rho^*(k)$ | Décalage $k$ | $\rho^*(k)$ | Décalage $k$ | $\rho^*(k)$ |
|---|---|---|---|---|---|---|---|
| 0 | 1.0000 | 12 | 0.9602 | 24 | 0.8896 | 36 | 0.8100 |
| 1 | 0.9758 | 13 | 0.9345 | 25 | 0.8647 | 37 | 0.7869 |
| 2 | 0.9555 | 14 | 0.9126 | 26 | 0.8433 | 38 | 0.7669 |
| 3 | 0.9391 | 15 | 0.8943 | 27 | 0.8253 | 39 | 0.7501 |
| 4 | 0.9266 | 16 | 0.8798 | 28 | 0.8107 | 40 | 0.7363 |
| 5 | 0.9177 | 17 | 0.8687 | 29 | 0.7994 | 41 | 0.7254 |
| 6 | 0.9126 | 18 | 0.8612 | 30 | 0.7913 | 42 | 0.7176 |
| 7 | 0.9113 | 19 | 0.8572 | 31 | 0.7864 | 43 | 0.7126 |
| 8 | 0.9136 | 20 | 0.8567 | 32 | 0.7843 | 44 | 0.7106 |
| 9 | 0.9196 | 21 | 0.8595 | 33 | 0.7862 | 45 | 0.7114 |
| 10 | 0.9293 | 22 | 0.8661 | 34 | 0.7909 | 46 | 0.7151 |
| 11 | 0.9429 | 23 | 0.8760 | 35 | 0.7989 | 47 | 0.7217 |

Un des arbitres a soulevé une question intéressante: qu'arrive-t-il lorsque les variances et les covariances des estimations de l'enquête sont inconnues? C'est une question difficile, à laquelle on ne peut répondre sans une réflexion approfondie. Cependant, le modèle présenté ici suppose que ces variances et covariances sont connues. En règle générale, il suffit que les équations d'estimation utilisées pour calculer les estimations du MV soient des estimateurs consistants des variances et des covariances. Il est courant, en étalonnage, d'estimer ces variances et ces covariances à l'aide des données de l'enquête puisqu'on ne connaît jamais les valeurs théoriques (voir, par exemple, Hillmer et Trabelsi 1987).

Les calculs requis pour notre exemple sont exécutés par un algorithme rédigé dans le langage de programmation GAUSS pour micro-ordinateurs. L'estimation initiale de $\beta$ pour le processus itératif, d'après (3.7), est $\hat{\beta}_0 = 0.9162$. L'estimation initiale du vecteur de paramètres $\boldsymbol{\theta}$ est calculée à l'aide de (3.4), après que l'on a remplacé $\beta$ par $\hat{\beta}_0$. La méthode de Fisher-Newton-Raphson et la méthode d'itération successive, dont il a été question dans la section 3, servent toutes deux à calculer les estimations du MV des paramètres du modèle. L'estimation du MV finale de $\beta$ se rapproche sensiblement de l'estimation initiale et $\hat{\beta}$ est égale à 0.9016, avec un CV de 0.0065. Il est intéressant de constater que, dans cet exemple, la méthode de Fisher-

**Tableau 3**

Estimations mensuelles sur le commerce de détail, estimations du MV de $\Theta_t$'s et des valeurs ajustées (en millions de dollars), et CV correspondants

| Année | Mois | $y_t$* | CV($y_t$)* | $\hat{\Theta}_t$ | CV($\hat{\Theta}_t$) | $\hat{y}_t$ | CV($\hat{y}_t$) |
|---|---|---|---|---|---|---|---|
| 1985 | 1 | 8,689.668 | 0.008 | 9,686.630 | 0.00210 | 8,733.384 | 0.00667 |
| | 2 | 8,390.380 | 0.008 | 9,350.078 | 0.00210 | 8,429.951 | 0.00665 |
| | 3 | 10,107.485 | 0.006 | 11,248.048 | 0.00233 | 10,141.146 | 0.00496 |
| | 4 | 10,541.145 | 0.008 | 11,741.785 | 0.00200 | 10,586.294 | 0.00656 |
| | 5 | 11,763.659 | 0.007 | 13,094.151 | 0.00198 | 11,805.576 | 0.00570 |
| | 6 | 11,067.487 | 0.008 | 12,321.326 | 0.00189 | 11,108.803 | 0.00647 |
| | 7 | 10,810.755 | 0.008 | 12,029.467 | 0.00184 | 10,845.666 | 0.00643 |
| | 8 | 11,289.656 | 0.009 | 12,554.808 | 0.00206 | 11,319.309 | 0.00726 |
| | 9 | 10,336.540 | 0.009 | 11,484.216 | 0.00205 | 10,354.073 | 0.00728 |
| | 10 | 11,213.751 | 0.010 | 12,447.696 | 0.00256 | 11,222.737 | 0.00809 |
| | 11 | 11,935.495 | 0.010 | 13,234.412 | 0.00258 | 11,932.034 | 0.00808 |
| | 12 | 13,300.288 | 0.008 | 14,734.891 | 0.00188 | 13,284.853 | 0.00643 |
| 1986 | 1 | 9,753.373 | 0.009 | 10,794.009 | 0.00221 | 9,731.787 | 0.00716 |
| | 2 | 9,249.279 | 0.009 | 10,227.777 | 0.00224 | 9,221.277 | 0.00709 |
| | 3 | 10,609.952 | 0.008 | 11,729.293 | 0.00207 | 10,575.031 | 0.00622 |
| | 4 | 11,637.936 | 0.008 | 12,860.626 | 0.00206 | 11,595.032 | 0.00614 |
| | 5 | 12,695.108 | 0.008 | 14,024.139 | 0.00205 | 12,644.046 | 0.00605 |
| | 6 | 11,826.254 | 0.008 | 13,059.556 | 0.00202 | 11,774.385 | 0.00598 |
| | 7 | 11,940.908 | 0.010 | 13,164.500 | 0.00233 | 11,869.002 | 0.00740 |
| | 8 | 11,866.547 | 0.010 | 13,070.205 | 0.00232 | 11,783.987 | 0.00743 |
| | 9 | 11,540.397 | 0.009 | 12,712.283 | 0.00202 | 11,461.287 | 0.00670 |
| | 10 | 12,208.845 | 0.010 | 13,430.932 | 0.00235 | 12,109.215 | 0.00747 |
| | 11 | 12,201.498 | 0.010 | 13,418.219 | 0.00240 | 12,097.753 | 0.00747 |
| | 12 | 14,479.170 | 0.009 | 15,933.951 | 0.00215 | 14,365.916 | 0.00670 |
| 1987 | 1 | 10,271.723 | 0.012 | 11,276.676 | 0.00357 | 10,166.956 | 0.00891 |
| | 2 | 9,951.105 | 0.010 | 10,945.319 | 0.00261 | 9,868.208 | 0.00737 |
| | 3 | 11,492.162 | 0.008 | 12,663.849 | 0.00230 | 11,417.620 | 0.00584 |
| | 4 | 12,867.443 | 0.009 | 14,172.605 | 0.00235 | 12,777.901 | 0.00652 |
| | 5 | 13,508.434 | 0.012 | 14,850.145 | 0.00343 | 13,388.765 | 0.00862 |
| | 6 | 13,608.274 | 0.011 | 14,973.985 | 0.00287 | 13,500.418 | 0.00786 |
| | 7 | 13,278.474 | 0.023 | 14,483.340 | 0.01066 | 13,058.057 | 0.00165 |
| | 8 | 12,728.196 | 0.008 | 14,028.998 | 0.00227 | 12,648.426 | 0.00577 |
| | 9 | 12,616.239 | 0.009 | 13,888.982 | 0.00233 | 12,522.188 | 0.00659 |
| | 10 | 13,760.829 | 0.008 | 15,156.409 | 0.00227 | 13,664.890 | 0.00592 |
| | 11 | 13,380.142 | 0.008 | 14,733.240 | 0.00227 | 13,283.365 | 0.00597 |
| | 12 | 16,269.757 | 0.007 | 17,928.148 | 0.00241 | 16,163.867 | 0.00525 |
| 1988 | 1 | 11,134.013 | 0.010 | 12,234.529 | 0.00274 | 11,030.548 | 0.00753 |
| | 2 | 10,959.374 | 0.010 | 12,042.761 | 0.00276 | 10,857.651 | 0.00754 |
| | 3 | 13,177.788 | 0.008 | 14,508.565 | 0.00233 | 13,080.800 | 0.00602 |
| | 4 | 13,666.311 | 0.009 | 15,035.737 | 0.00243 | 13,556.094 | 0.00676 |
| | 5 | 14,267.530 | 0.006 | 15,742.039 | 0.00379 | 14,192.890 | 0.00448 |
| | 6 | 14,432.944 | 0.009 | 15,884.130 | 0.00240 | 14,320.997 | 0.00673 |
| | 7 | 13,960.825 | 0.009 | 15,363.957 | 0.00240 | 13,852.014 | 0.00673 |
| | 8 | 13,691.315 | 0.008 | 15,073.691 | 0.00233 | 13,590.312 | 0.00606 |
| | 9 | 13,773.109 | 0.008 | 15,159.075 | 0.00235 | 13,667.294 | 0.00613 |
| | 10 | 13,900.743 | 0.009 | 15,279.950 | 0.00255 | 13,776.282 | 0.00696 |
| | 11 | 14,453.461 | 0.009 | 15,884.279 | 0.00260 | 14,321.132 | 0.00700 |
| | 12 | 17,772.990 | 0.009 | 19,529.791 | 0.00267 | 17,607.895 | 0.00702 |

***Source**: Commerce de détail, publication de Statistique Canada (n° 63-005 au catalogue, mensuel).

Newton-Raphson a convergé très rapidement vers une solution. De fait, elle a convergé en 6 itérations seulement (environ 1 minute) - avec une précision de dix chiffres - tandis que la méthode des calculs successifs a convergé, pour le même degré de précision, au bout de 500 itérations et plus (plus de 45 minutes) sur un ordinateur personnel 386DX-25Mhz. Cependant, comme c'était prévu, les deux méthodes ont convergé vers la même solution finale. On obtient la matrice des covariances du vecteur estimé $(\hat{\boldsymbol{\theta}}':\hat{\beta})'$ en inversant la matrice d'information de Fisher $\boldsymbol{\Omega}$, définie en (4.1), après avoir remplacé les paramètres par les estimations du MV correspondantes. Le tableau 3 donne la série originale des estimations mensuelles sur le commerce de détail de même que la série étalonnée des estimations du MV avec les CV correspondants. On y trouve aussi la série des valeurs infra-annuelles ajustées et les CV correspondants (deux dernières colonnes). La figure 1 représente graphiquement la série originale et la série étalonnée. Ce graphique permet de constater que l'étalonnage ne modifie pas la tendance initiale de la série et qu'il s'opère une très forte réduction du CV des estimations infra-annuelles. Le tableau 4 donne la série originale des estimations annuelles sur le commerce de détail de même que la série des valeurs annuelles ajustées avec les CV correspondants. On calcule les variances des valeurs ajustées des tableaux 3 et 4 au moyen des expressions (4.4) et (4.5) respectivement, après avoir remplacé les paramètres par les estimations du MV correspondantes. Les résultats relatifs aux valeurs ajustées témoignent aussi d'une forte réduction du CV des estimations originales. Autrement dit, l'étalonnage a pour effet d'accroître la précision des séries mensuelles et annuelles.

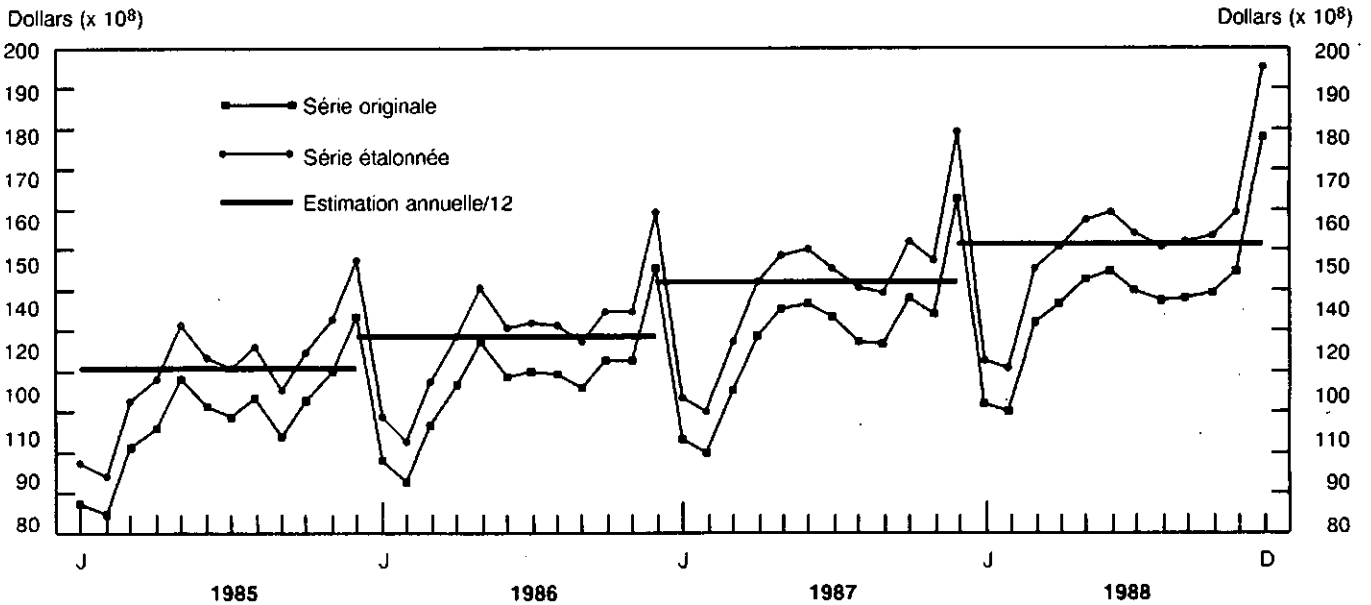


**Figure 1.** Estimations mensuelles sur le commerce de détail Canada, tous les magasins

**Tableau 4**

Estimations annuelles sur le commerce de détail et valeurs annuelles ajustées (en millions de dollars) et CV correspondants

| Année | $z_T^*$ | CV($z_T$) | $\hat{z}_T$ | CV($\hat{z}_T$) |
|---|---|---|---|---|
| 1985 | 143,965.400 | 0.00033 | 143,927.507 | 0.00032 |
| 1986 | 154,377.100 | 0.00031 | 154,425.491 | 0.00030 |
| 1987 | 169,944.600 | 0.00193 | 169,101.697 | 0.00128 |
| 1988 | 181,594.000 | 0.00137 | 181,738.512 | 0.00127 |

***Source**: Commerce de détail annuel, publication de Statistique Canada (n° 63-223 au catalogue, annuel).

## 7. CONCLUSIONS

Le modèle non linéaire traité dans cet article semble se prêter très bien à l'étalonnage de séries économiques construites à partir de grandes enquêtes par sondage. Les méthodes itératives proposées pour calculer les estimations du maximum de vraisemblance des paramètres du modèle sont très faciles à appliquer. Cependant, la convergence se fait beaucoup plus lentement avec la méthode des calculs successifs des équations d'estimation qu'avec la méthode de Fisher-Newton-Raphson. Nous avons présenté dans cet article les expressions en forme analytique fermée pour les covariances des estimateurs du MV. Ces estimateurs et les covariances correspondantes peuvent servir à faire des inférences sur les paramètres du modèle. Nous avons aussi

défini les expressions pour les valeurs infra-annuelles et annuelles ajustées et les covariances asymptotiques correspondantes. La méthode décrite dans cet article semble offrir un ajustement précis pour les données sur le commerce de détail au Canada. Cependant, il est nécessaire d'élaborer des tests de validité de l'ajustement pour ce modèle et pour d'autres modèles d'étalonnage.

## REMERCIEMENTS

Nous tenons à remercier le rédacteur en chef, le rédacteur associé et deux arbitres pour leurs précieux conseils qui ont contribué à hausser grandement la qualité de cet article. Nous voulons aussi exprimer notre gratitude à messieurs J. Gambino et M. Kovačević, de Statistique Canada, pour les commentaires qu'ils ont faits sur cette étude.

## BIBLIOGRAPHIE

BOX, G.E.P., et JENKINS, G.M. (1976). *Time Series Analysis, Forecasting and Control*. New York: Holden-Day.

CHOLETTE, P.A. (1984). L'ajustement des séries infra-annuelles aux repères annuels. *Techniques d'enquête*, 10, 39-53.

CHOLETTE, P.A. (1987). Benchmarking and interpolation of time series. Direction de la méthodologie, document de travail, Statistique Canada.

CHOLETTE, P.A. (1988). Benchmarking systems of socio-economic time series. Direction de la méthodologie, document de travail, Statistique Canada.

CHOLETTE, P.A. (1992). Users' manual of programmes BENCH and CALEND to benchmark, interpolate and Calendarize time series data on micro computers. Direction de la méthodologie, document de travail, Statistique Canada.

CHOLETTE, P.A., et DAGUM, E.B. (1989). Benchmarking socio-economic time series data: a unified approach. Direction de la méthodologie, document de travail, Statistique Canada.

CHOLETTE, P.A., et DAGUM, E.B. (1991). Benchmarking time series with autocorrelated sampling errors. Direction de la méthodologie, document de travail, Statistique Canada.

DENTON, F.T. (1971). Adjustment on monthly or quarterly series to annual totals: An approach based on quadratic minimization, *Journal of the American Statistical Association*, 66, 99-102.

GODAMBE, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 13, 1208-1211.

HIDIROGLOU, M.A., et GIROUX, S. (1986). Composite estimation for the Retail Trade Survey. Direction de la méthodologie, document de travail, Statistique Canada.

HILLMER, S.C., et TRABELSI, A. (1987). Benchmarking of economic time series. *Journal of American Statistical Association*, 82, 1064-1071.

LANIEL, N., et FYFE, K. (1989). Benchmarking of economic time series. Direction de la méthodologie, document de travail, Statistique Canada.

LANIEL, N., et FYFE, K. (1990). Étallonnage des séries économiques. *Techniques d'enquête*, 16, 283-289.

LANIEL, N., et MIAN, I.U.H. (1991). Maximum likelihood estimation for the constant bias model with mixed benchmarks. Direction de la méthodologie, document de travail, Statistique Canada.

McLEOD, I. (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Applied Statistics*, 24, 255-256.

MIAN, I.U.H., et LANIEL, N. (1991). Maximum likelihood estimation for the constant bias benchmarking model. Direction de la méthodologie, document de travail, Statistique Canada.

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

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 183–192
Statistique Canada



# Plans de sondage à deux degrés optimaux pour des estimateurs de ratios: application au contrôle de qualité du recensement français de 1990

JEAN-CLAUDE DEVILLE[1]

RÉSUMÉ

Cette étude est basée sur l'utilisation de modèles de superpopulation pour anticiper la variance d'une mesure par sondage de ratios *avant* l'enquête. On arrive, en utilisant des modèles simples qu'on voudrait néanmoins assez réalistes, à des expressions plus ou moins complexes qu'on parvient à optimiser, parfois rigoureusement, quelquefois de façon approximative. La solution du dernier des problèmes évoqués fait apparaître un facteur assez peu étudié en matière d'optimisation de plan de sondage: le coût lié à la mobilisation d'une information individuelle.

MOTS CLÉS: Contrôle de qualité du recensement; modèle de superpopulation; optimisation d'un plan à deux degrés; enquête à objectifs multiples.

## 1. INTRODUCTION

Le contrôle par sondage de la qualité des données du recensement Français a posé quelques problèmes à la fois intéressants et nouveaux. Trois d'entre eux sont traités dans cet article. Nous les étudierons en termes généraux et nous décrirons ensuite leur application au cas du recensement.

Dans tous les cas, le problème est celui de l'optimisation d'un sondage à deux degrés où les unités primaires sont des districts de collecte du recensement. Ceux-ci sont repérés par un indice $k$ variant dans une population $U$ de districts, qui est concrètement une unité de traitement des bulletins collectés.

Le premier problème consiste à estimer la fréquence d'un caractère dans la population de bulletins (le fait de comporter une erreur). Désirant avoir une précision donnée pour cette estimation on cherche à minimiser le coût du sondage avec une fonction de coût de la forme:

$$C_T = mC_o + nC_1, \tag{1.1}$$

où $m$ est le nombre d'unités primaires (districts) échantillonnées, $C_o$ le coût de traitement d'une U.P., $n$ le nombre d'unités finales (bulletins) échantillonnées, et $C_1$ le coût de traitement d'une unité finale. Le problème est assez habituel dans le cas de l'estimation d'une moyenne (voir par exemple W. Cochran (1977)). Il reçoit ici une solution plus complète tenant compte de la grande variabilité de taille des unités primaires.

Le second problème est plus original et possède une portée plus générale. La population finale (ici les bulletins) est composée de $G$ groupes ($g = 1$ à $G$) distincts. On désire avoir une estimation de la fréquence du caractère dans chacun des groupes, avec une précision donnée pour chacun d'eux. La contrainte réside dans le fait que les unités primaires seront communes à tous les groupes, l'échantillonnage au sein de chaque U.P. portant alors sur chaque groupe.

L'objectif est alors de minimiser le coût du sondage celui-ci ayant la forme:

$$C_T = mC_o + \sum_{g=1}^{G} n_g C_g, \tag{1.2}$$

où $n_g$ est le nombre total d'unités finales du groupe $g$ et $C_g$ le coût de traitement d'une unité finale du groupe $g$. En pratique les groupes sont constitués par les différents types de bulletins utilisés dans le recensement.

Le troisième problème a sa source dans le contrôle de la codification. On possède, *a priori*, une mesure de la difficulté de codification de chaque bulletin. Formellement donc, on dispose au niveau de chaque individu $i$ de la population, d'une variable quantitative $X_i$, telle que la probabilité – en un sens à préciser – pour que l'individu possède le caractère à mesurer soit à peu près proportionnelle à $X_i$. On cherche à utiliser cette information pour minimiser le coût du contrôle (mesure de la fréquence du caractère "codification erronée") sous la requête d'une précision donnée du sondage.

Dans chacun des cas, on utilise des modèles de superpopulation plausibles et simples qui permettent d'évaluer la variance anticipée du sondage. On a, en quelque sorte, une illustration presque typique d'un "échantillonnage assisté par un modèle" dans l'esprit du livre de Särndal, Swensson, Wretman (1992).

[1] Jean-Claude Deville, Chef de la Division des Méthodes Statistiques et Sondages, Institut National de la Statistique et des Études Économiques, 18, boul. Adolphe Pinard, 75675 Paris, CEDEX 14.

## 2. ESTIMATION OPTIMALE DE LA PROPORTION D'ENREGISTREMENTS ERRONNÉS À L'AIDE D'UN PLAN DE SONDAGE À DEUX DEGRÉS

Les unités primaires $k$ (districts) comportent chacune un nombre connu $N_k$ d'individus (des bulletins). Parmi ceux-ci $D_k$ possèdent le caractère (être erronné). Le but est donc d'estimer:

$$P = \sum_U D_k \Big/ \sum_U N_k .$$

Le sondage consistera à tirer un échantillon $s$ d'unités primaires (U.P.) avec des probabilités d'inclusion $\pi_k$ au premier ordre et $\pi_{k\ell}$ au second ordre à déterminer. Ensuite, si l'unité primaire $k$ est tirée dans $s$ on vérifiera $n_k$ individus tirés par sondage aléatoire simple sans remise. Soit $d_k$ le nombre de bulletins erronnés qu'on y relèvera.

L'estimateur $\hat{P}_k$ de $P_k = D_k/N_k$ sera $\hat{P}_k = d_k/n_k$ et $\hat{D}_k = N_k\hat{P}_k$ estimera $D_k$ sans biais. L'estimateur de $P$ sera:

$$\hat{P} = \frac{\sum_s \frac{\hat{D}_k}{\pi_k}}{\sum_s \frac{\hat{N}_k}{\pi_k}} . \tag{2.1}$$

C'est le ratio des estimateurs sans biais de $D$ et de $N$, le nombre total de bulletins. Bien que ce nombre soit connu, il est bien évident que l'estimateur (4.1) est plus précis que $1/N \sum_s \hat{D}_k/\pi_k$.

On a:

$$\mathrm{Var}(\hat{P}) = E\,\mathrm{Var}(\hat{P} \mid s) + \mathrm{Var}E(\hat{P} \mid s). \tag{2.2}$$

Or:

$$\mathrm{Var}(\hat{P} \mid s) = \hat{N}^{-2} \sum_s \frac{N_k^2}{\pi_k^2} \frac{P_k(1-P_k)N_k}{N_k - 1}\left(\frac{1}{n_k} - \frac{1}{N_k}\right)$$

avec $$\hat{N} = \sum_s \frac{N_k}{\pi_k}.$$

D'où:

$$E\,\mathrm{Var}(\hat{P} \mid s) \simeq N^{-2} \sum_U \frac{N_k^2}{\pi_k} \frac{P_k(1-P_k)N_k}{N_k - 1}\left(\frac{1}{n_k} - \frac{1}{N_k}\right). \tag{2.3}$$

Par ailleurs,

$$E(\hat{P} \mid s) = \frac{\sum_s \frac{D_k}{\pi_k}}{\sum_s \frac{N_k}{\pi_k}} .$$

La variance de cette quantité s'obtient par linéarisation en introduisant la variable $Z_k = D_k - PN_k = N_k(P_k - P)$.

On obtient:

$$\mathrm{Var}\, E(\hat{P} \mid s) \simeq N^{-2}\, \mathrm{Var}\left(\sum_s \frac{z_k}{\pi_k}\right).$$

Soit, compte tenu de ce que $\sum_U Z_k = 0$:

$$\mathrm{Var}E(\hat{P} \mid s) = N^{-2}\left(\sum_k \frac{Z_k^2}{\pi_k} + \sum\sum_{k \neq l} \frac{Z_k Z_\ell}{\pi_k \pi_\ell}\pi_{k\ell}\right). \tag{2.4}$$

La somme des quantités (2.3) et (2.4) nous donne la variance de l'estimateur (2.1).

### 2.1 Introduction d'un modèle

La variance de $\hat{P}$ est difficile à manipuler et, de plus, contient des paramètres inconnus. On se tire de la difficulté en faisant de nécessaires hypothèses qui se traduisent par un modèle de superpopulation. On supposera plus loin que les paramètres de ce modèle sont susceptibles d'être estimés à partir d'un essai préliminaire portant sur une toute petite partie de la population. On note $E_\xi$ l'espérance sous le modèle (resp $\mathrm{Var}_\xi$ pour la variance) dont tous les aléas sont supposés indépendants du processus d'échantillonnage.

Le modèle suit les spécifications suivantes:

a) $D_k$ suit une loi binomiale $(N_k, p_k)$. $P_k$ est donc, *sous le modèle*, un estimateur de $p_k$.

b) $p_k$ est lui même aléatoire. On suppose les $p_k$ indépendantes et de même loi avec:

$$E_\xi\, p_k = P,$$

$$\mathrm{Var}_\xi\, p_k = \sigma^2$$

pour tout $k$, quelle que soit, en particulier, la valeur de $N_k$.

En conditionnant, dans le modèle, par les $p_k$ on a évidemment:

$$E_\xi(D_k \mid p_k) = N_k\, p_k,$$

$$\mathrm{Var}_\xi(D_k \mid p_k) = N_k\, p_k(1 - p_k).$$

La *variance anticipée* de $\hat{P}$ est la quantité $E_\xi\, \mathrm{Var}\hat{P}$. C'est à elle que nous allons nous intéresser désormais. Pour l'évaluer on remarque que:

a) $$E_\xi(P_k - P)^2 = E_\xi(E_\xi(P_k - p_k + p_k - P)^2 \mid p_k)$$

$$= \frac{P(1-P) - \sigma^2}{N_k} + \sigma^2,$$

b) $E_\xi P_k(1 - P_k) = E_\xi(E_\xi((P_k - P_k^2) \mid p_k))$

$$= E_\xi p_k(1 - p_k)\frac{N_k - 1}{N_k}$$

$$= (P(1 - P) - \sigma^2)\frac{N_k - 1}{N_k},$$

c) $E_\xi Z_k Z_\ell = 0$ à cause de l'indépendance des $Z_k$ et des $Z_\ell$, ce qui nous débarrasse d'un terme bien encombrant en même temps que des $\pi_{k\ell}$.

En récollant tous les morceaux de (2.3) et (2.4) un petit miracle algébrique se produit et nous avons l'expression:

$$\boxed{\begin{array}{c} E_\xi \mathrm{Var}\,\hat{P} \simeq N^{-2} \sum_U \frac{N_k^2}{\pi_k}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \\ \text{avec } \tau^2 = P\,(1 - P) - \sigma^2 \\ \text{(quantité positive par nature)} \end{array}} \quad . \qquad (2.1.1)$$

**Remarque:**

Le miracle algébrique s'explique bien si on ne cherche pas à obtenir la variance sous le plan de sondage uniquement. Elle est d'ailleurs la conséquence d'un modèle un peu plus général que celui que nous avons posé.

Supposons que nous voulions estimer le total $N\bar{Y} = \sum_U Y_i$ d'une variable $Y$ et que pour cela nous réalisions un tirage à deux degrés: un premier degré où des unités primaires $k$ sont tirées avec des probabilités $\pi_k$, un second où $n_k$ unités finales sont tirées par sondage aléatoire simple.

Nous posons un modèle où:

$$Y_i = \bar{Y} + \alpha_k + \epsilon_i,$$

avec $\alpha_k$ variable liée à l'U.P. d'indice $k$. Les $\alpha_k$ sont indépendantes de même loi d'espérance nulle de variance $\sigma^2$. Les $\epsilon_i$ sont également indépendantes centrées de variance égale à $\tau^2$. Avec $\pi_i^* = \pi_k n_k / N_k$ ($N_k$ taille de l'U.P. numéro $k$), l'estimateur de Horvitz-Thompson du *total* vaut $\hat{Y} = \sum Y_i/\pi_i^*$ la somme étant étendue à l'échantillon. Sous le modèle, et conditionnellement à l'échantillon on a:

$$\mathrm{Var}_\xi(\hat{Y} \mid s) = \sum_s \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right).$$

L'espérance sous le plan de cette expression redonne la formule (2.1.1).

## 2.2 Recherche d'un plan de sondage optimal

La variance maximum de $\hat{P}$ est fixée par les critères retenus pour le contrôle de qualité. Le sondage étant répété pour chacune des unités de traitement il est tout à fait naturel de chercher à minimiser l'espérance du coût du sondage donné en (2.1.1) soit:

$$E \sum_s (C_o + n_k C_1) = \sum_U \pi_k(C_o + n_k C_1). \qquad (2.2.1)$$

Le problème d'optimisation s'écrit donc:

$$\text{Minimiser} \sum_U \pi_k (C_o + n_k C_1)$$

sous les contraintes:

$$N^{-2} \sum_U \frac{N_k^2}{\pi_k}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \leq V_o$$

et pour tout $k$, $n_k \leq N_k$.

Associons un multiplicateur de Lagrange $\lambda$ à la première contrainte - qui sera évidemment saturée - et des multiplicateurs $\mu_k$ aux autres. On obtient les solutions:

$$C_o + n_k C_1 = \lambda \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \qquad (2.2.2)$$

et, pour tout $k$:

$$C_1 \pi_k = \lambda \frac{N_k^2}{\pi_k} \cdot \frac{\tau^2}{n_k^2} + \mu_k \qquad (2.2.3)$$

avec

$$\mu_k = 0 \quad \text{si} \quad n_k < N_k \quad \text{et} \quad \mu_k > 0 \quad \text{si} \quad n_k = N_k.$$

Pour l'utilisation des multiplicateur de Lagrange, voir par exemple Luenberger (1973).

Pour toutes les unités primaires où $\mu_k = 0$ (les plus grosses) on obtient:

$$n_k = \frac{\tau}{\sigma}\left(\frac{C_o}{C_1}\right)^{1/2} = n^*. \qquad (2.2.4)$$

Chaque unité primaire reçoit donc la même allocation, ce qui correspond à l'idée qu'on a besoin de la même précision dans chacune d'elle. Retournons à l'équation (2.2.3). On constate alors que, toujours pour ces unités primaires, les probabilités d'inclusion $\pi_k$ doivent être proportionnelles aux tailles $N_k$ soit:

$$\pi_k = \lambda^{1/2} C_1^{-1/2} \frac{\tau}{n^*} N_k. \qquad (2.2.5)$$

C'est la justification habituelle d'un sondage autopondéré avec un premier degré tiré avec des probabilités proportionnelles à une mesure de taille (Voir par exemple Cochran 1977).

Comme $n_k$ ne dépend pas de $N_k$ on ne pourra avoir $n_k = N_k$ et $\mu_k > 0$ que si $N_k \leq n^*$. L'équation (2.2.2) nous permet alors d'obtenir les probabilités d'inclusion à un facteur près:

$$\pi_k = \lambda^{1/2} N_k \left(\frac{\sigma^2 + \tau^2/N_k}{C_o + C_1 N_k}\right)^{1/2} = \lambda^{1/2} N_k^{1/2} \left(\frac{N_k \sigma^2 + \tau^2}{N_k C_1 + C_o}\right)^{1/2}. \tag{2.2.6}$$

Les relations (2.2.5), valide si $N_k \geq n^*$ et (2.2.6) valide si $N_k \leq n^*$ établissent que $\pi_k$ est proportionnelle à une variable connue $T_k = f(N_k)$ dont le graphique est donné à la figure 1.

Pour spécifier entièrement le sondage il reste à trouver le nombre $m$ d'unités primaires à tirer. Or, $T = \sum_U T_k$ est aussi une quantité connue.

En se restreignant à un échantillonnage de taille fixe on aura donc $\pi_k = m\, T_k/T$. On trouve $m$ en portant cette valeur dans la contrainte de variance soit:

$$N^2 V_o m = T \sum_U \frac{N_k^2}{T_k}(\sigma^2 + \tau^2/n_k).$$

Si, en première approximation, on prend $T_k = N_k$, on obtient la formule simplifiée:

$$m V_o = \sigma^2 + \tau^2/n^*.$$

On a ainsi obtenu une solution complète au problème.

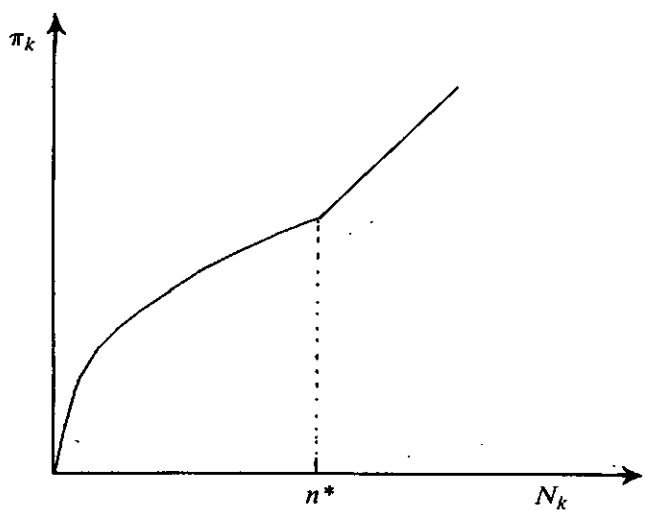


**Figure 1.** Graphique de $\pi_k$ en fonction de $N_k$

## 3. ESTIMATION OPTIMALE DANS LE CAS D'UN SONDAGE À DEUX DEGRÉS OU LES UNITÉS PRIMAIRES SONT STRATIFIÉES

La dure réalité des choses nous amène à compliquer un peu le problème car on doit, en fait, contrôler indépendamment plusieurs types de bulletins. Ceci amène à poser un problème assez général qui est le suivant:

Pour chaque unité primaire (ici les districts d'une unité de traitement) on connaît les effectifs $N_{kg}$, d'unités secondaires appartenant à $G$ groupes. La "population" de l'UP numéro $k$ vaut $N_{k+} = \sum_g N_{kg}$; celle du groupe $g$ vaut $N_{+g} = \sum_k N_{kg}$. Comme dans ce qui précède on cherche avec quelle probabilité d'inclusion $\pi_k$ échantillonner l'UP numéro $k$, le nombre d'UP à tirer et l'allocation $n_{kg}$ de l'échantillon parmi les différents groupes dans l'UP $k$, sachant que ces $n_{kg}$ unités sont tirées par un sondage aléatoire simple parmi les $N_{kg}$ unités tirables.

### 3.1 Recherche d'un plan optimal à l'aide d'un modèle

On postule, dans chacun des groupes, un modèle identique à celui formulé à la section (2.1) (ou sous une forme plus générale dans la remarque qui la termine).

Pour $g = 1$ à $G$ on aura donc:

$$v_g = E_\xi \operatorname{Var}(\hat{P}_g) = N_{+g}^{-2} \sum_U \frac{N_{kg}^2}{\pi_k}(\sigma_g^2 + \tau_g^2/n_{kg}). \tag{3.1.1}$$

La fonction de côut est donnée par la forme générale (1.2). On va chercher à minimiser l'espérance du coût de sondage:

$$C_T = \sum_U \pi_k \left(C_o + \sum_g n_{kg} C_g\right), \tag{3.1.2}$$

sous les contraintes $V_g \leq \mathcal{V}_g$, où les quantités $\mathcal{V}_g$, sont fixées de façon extérieure, par exemple par la qualité des données qu'on veut obtenir et la rigueur du contrôle.

Sous cette forme, le problème peut s'avérer assez complexe. Nous allons écrire un "Lagrangien" général:

$$L = \lambda C_T + \sum_g \lambda_g V_g.$$

Le problème posé fixe $\lambda = 1$ et les $\lambda_g$ sont des multiplicateurs à déterminer. Une variante simple consiste à fixer les $\lambda_g$: on désire alors minimiser une combinaison linéaire donnée des variances sous une contrainte de coût. Dans toutes les hypothèses, on obtient par dérivation par rapport aux $n_{kg}$ (considérées comme des variables réelles):

$$\lambda \pi_k^2 C_g = \lambda_g N_{+g}^{-2} N_{kg}^2 \tau_g^2/n_{kg}^2. \tag{3.1.3}$$

Les $\pi_k$ étant, pour l'instant, destinées à être connues à un facteur près, on peut écrire:

$$\pi_k n_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g \frac{N_{kg}}{N_{+g}}. \qquad (3.1.4)$$

Par sommation sur $k$ on en déduit que:

$$E\, n_{+g} = \sum_U \pi_k n_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g. \qquad (3.1.5)$$

La taille totale de l'échantillon dans chaque groupe est donc directement liée au multiplicateur $\lambda_g$.

La dérivation du Lagrangien par rapport aux $\pi_k$ nous donne de nouvelles relations qui se simplifient miraculeusement si on utilise aussi (3.1.4). On obtient:

$$C_o = \sum_g C_g \left(\frac{\sigma_g}{\tau_g}\right)^2 n_{kg}^2, \qquad (3.1.6)$$

où encore, si on introduit les nombres

$$n_g^* = \left(\frac{C_o}{C_g}\right)^{1/2} \frac{\tau_g}{\sigma_g},$$

on écrit:

$$\sum_g \left(\frac{n_{kg}}{n_g^*}\right)^2 = 1. \qquad (3.1.7)$$

Comme on s'en rend compte en jetant un oeil à la formule (2.2.4), les $n_g^*$ sont les nombres d'unités secondaires à tirer par UP s'il n'y a qu'un seul groupe; $n_{kg}$ sera toujours inférieur à $n_g^*$.

De (2.1.4), (3.1.5) et (3.1.7) on tire les relations:

$$\pi_k^2 = \frac{1}{C_o} \sum_g \lambda_g \sigma_g^2 \left(\frac{N_{kg}}{N_{+g}}\right)^2. \qquad (3.1.8)$$

Ainsi, les $\pi_k$ sont proportionnelles aux quantités $T_k$ telles que $T_k^2 = \sum_g \lambda_g \sigma_g^2 N_{kg}^2 / N_{+g}^2$ qui apparaissent comme la mesure de taille adéquate. Les relations (3.1.4) montrent que, à $k$ fixé, les $n_{kg}$ sont proportionnelles à $n_g^* \lambda_g^{1/2} \sigma_g N_{kg}/N_{+g}$, ce qui compte tenu de (3.1.7) conduit à:

$$n_{kg} = n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}} T_k^{-1}. \qquad (3.1.9)$$

### 3.2 Solutions explicites dans deux cas particuliers

a) Si les $\lambda_g$ étaient connus, c'est-à-dire si on minimisait $\sum_g \lambda_g v_g$ sous une contrainte de coût, alors (3.1.2) et (3.1.9) nous permettraient de calculer les $T_k$. En reportant:

$$\pi_k = m\, T_k / T \left( T = \sum_U T_k,\ m \ \text{nombre d'unités primaires à tirer} \right)$$

dans la contrainte de budget $C_T \leq C_T^*$, on trouve:

$$C_T^* = \frac{m}{T} \sum_U \left( C_o T_k + \sum_g C_g n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}} \right)$$

soit:

$$m = C_T^* \Big/ \left( C_o + \sum_g C_g n_g^* \cdot \frac{\lambda_g^{1/2} \sigma_g}{T} \right).$$

Si un seul des $\lambda_g$ est différent de zéro, on peut vérifier assez facilement qu'on retrouve le résultat donné à la fin de la section (2.2).

b) Le problème initial (min $C_T$ sous $V_g \leq \mathcal{V}_g$) se résoud assez facilement dans deux cas particuliers:

b1 - *Dispersion maximale* des groupes. Pour toute UP $k$, on a $N_{kg} = N_{k+}$ pour un certain $k$. Le problème est décomposé en $G$ problèmes distincts, chacun d'eux étant du type étudié dans la section 2.

b2 - *Dispersion minimale*: La répartition est la même dans toutes les UP; autrement dit on a pour tout $k$ et $g$

$$N_{kg} = N_{k+} \frac{N_{+g}}{N} \quad \text{avec} \quad \left( N = \sum_g N_{+g} \right),$$

$T_k$ est alors proportionnelle à $N_{k+}$, et les $n_{kg}$ sont des quantités $n_g^* u_g$ indépendantes de $k$.

Avec $\pi_k = m N_{k+}/N$, on obtient en écrivant $V_g = \mathcal{V}_g$:

$$m \mathcal{V}_g = \sigma_g^2 + \tau_g^2 / n_g^* u_g$$

soit:

$$m = \frac{\sigma_g^2}{\mathcal{V}_g} + u_g^{-1} \frac{\tau_g^2}{n_g^* \mathcal{V}_g}.$$

On obtient ainsi G-1 relations linéaires entre les $u_g^{-1}$ ce qui permet, en principe, de résoudre complètement le problème sachant que la somme des $u_g^2$ vaut 1.

### 3.3 Un algorithme numérique permettant de trouver la solution optimale dans le cas général

Une résolution numérique itérative du problème peut se faire de la façon suivante:

Étape 1: On fixe une allocation approximative de l'échantillon dans chaque groupe, soit $n_{+g}$ unités dans le groupe $g$. Pour y arriver on peut, par exemple, se servir de la solution approximative avec les hypothèses du point a) ou du point b).

Étape 2: La valeur des $\lambda_g$ est déterminée par les relations (3.1.5):

$$\lambda_g = C_g n_{+g}^2 / \tau_g^2.$$

Étape 3: Les $\pi_k$ sont déterminées par les relations (3.1.8). La somme des $\pi_k$ fixe, en particulier, le nombre d'UP à tirer.

Étape 4: Les $n_{kg}$ sont déterminés par les relations (3.1.4). On peut ensuite itérer par retour à l'étape 2 en espérant que cet algorithme converge vers la solution d'optimisation.

**Remarque:** La probabilité de tirer une unité de type $g$ vaut

$$\pi_k n_{kg} \Big/ N_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g / N_{+g}.$$

Elle ne dépend pas de l'Unité Primaire $k$ et est donc la même pour chaque unité d'un groupe $g$ donné (sondage à probabilités égales). On en déduit la taille $n_{+g}$ de l'échantillon dans le groupe $g$, ou du moins, son espérance mathématique. Pratiquement, il arrive que l'on fixe "autoritairement" les tailles des échantillons. Ceci revient à déterminer les $\lambda_g$, ou, implicitement, des variances $\mathcal{V}_g$. Ce résultat est assez naturel lui aussi.

## 4. ESTIMATION OPTIMALE À L'AIDE D'UNE MESURE DE LA DIFFICULTÉ À CODER UN ENREGISTREMENT

Il s'agit d'estimer la proportion de bulletins présentant une erreur de codification dans l'"univers" $U$ de tous les bulletins codifiés une semaine donnée dans une Direction Régionale. Le caractère particulier du problème est le suivant: tous les bulletins $i$ sont déjà précodifiés ce qui permet, grâce à des informations tirées de l'essai de recensement, d'attribuer à chacun d'eux une variable numérique positive $X_i$ qui traduit sa "difficulté". Cette variable a été calibrée de façon à ce que $Y_i$ (qui vaut 1 en cas d'erreur et 0 sinon) ait une "espérance" proportionnelle à $X_i$.

Toujours pour les mêmes raisons de coût du contrôle, on est amené à envisager un sondage à deux degrés:

- au premier degré de sondage on tirera un échantillon $s_1$ de districts $k$ (les Unités Primaires) à probabilités inégales $\pi_k$ à déterminer. On notera $\pi_{k\ell}$ les probabilités d'inclusion double pour cet échantillonnage.

- au second degré de sondage, on tirera un échantillon $s_k$ d'unités finales (les bulletins) dans l'unité primaire échantillon $k$. On notera $\pi_{i|k}$ la probabilité d'inclusion de l'unité dans l'unité primaire $k$, $\pi_{ij|k}$ la probabilité d'inclusion du couple $(i,j)$ dans les unités primaires et $s = U_{k\in s_1} s_k$ l'échantillon d'unités finales.

On notera $X_k = \sum_{i\in k} X_i$ le total des $X_i$ dans l'unité primaire $k$, $X = \sum_{k\in U_o} X_k = \sum_U X_i$ et on adoptera des notations analogues pour toutes les variables. ($U_o$ désigne la population des Unités Primaires - districts, $U$ la population des Unités finales - bulletins).

Le but est d'estimer une quantité de la forme $R = \sum_U Y_i / \sum_U W_i$ où $W_i$ est une variable connue pour chaque bulletin. Cela pourra être $W_i = 1$ ou $W_i = X_i$ selon la mesure qui semble la plus adéquate du taux d'erreur.

### 4.1 Choix d'estimateur et variance

a) Au niveau de l'Unité Primaire numéro $k$ il est naturel d'estimer le total $Y_k$ des $Y_i$ pour $i\in k$ par le ratio:

$$\hat{Y}_k = X_k\left(\sum_{s_k} Y_i/\pi_{i|k}\right)\Big/\left(\sum_{s_k} X_i/\pi_{i|k}\right) = X_k \hat{a}_k.$$

Ici $\hat{a}_k$ estime $a_k = Y_k/X_k$ avec un faible biais.

b) Pour estimer le ratio $Y/X$ on utilisera:

$$\hat{a} = \frac{\sum_{s_1} \frac{\hat{Y}_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}} = \frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}}.$$

c) Si on veut estimer $R$, on remarquera que:

$$R = \frac{Y}{X} \cdot \frac{X}{W},$$

où $X$ et $W$ sont des totaux connus (difficulté totale et nombre total de bulletins, par exemple). Comme la variable $X_i$ a été choisie pour sa bonne corrélation avec $Y_i$, un estimateur *a priori* intéressant de $R$ sera:

$$\hat{R} = \hat{a}\frac{X}{W}$$

de sorte que la seule véritable question porte sur l'estimation de $a = \sum_k a_k X_k / X$.

d) On aura:

$$\mathrm{Var}(\hat{a}) = \mathrm{Var}E(\hat{a} \mid s_1) + E\,\mathrm{Var}(\hat{a} \mid s_1).$$

Pour le premier terme, compte tenu du fait que $\hat{a}_k$ estime (à peu près) sans biais $a_k$, on peut écrire:

$$\mathrm{Var}E(\hat{a} \mid s_1) \simeq \frac{1}{X^2}\mathrm{Var}\left(\sum_{s_1} \frac{(a_k - a)X_k}{\pi_k}\right)$$

$$= \frac{1}{X^2}\left(\sum_k \frac{(a_k - a)^2 X_k^2}{\pi_k^2}\right.$$

$$\left. + \sum\sum_{k\neq l} (a_k - a)(a_\ell - a)\frac{X_k X_\ell \pi_{k\ell}}{\pi_k\,\pi_\ell}\right). \quad (4.1.1)$$

Pour le second terme, on a, *conditionnellement à* $s_1$:

$$\mathrm{Var}\left(\frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}}\right) = \left(\sum_{s_1} \frac{X_k}{\pi_k}\right)^{-2} \cdot \sum_{s_1} \mathrm{Var}(\hat{a}_k) \frac{X_k^2}{\pi_k^2}.$$

L'espérance de cette quantité vaut approximativement

$$X^{-2} \sum_k E\ \mathrm{Var}(\hat{a}_k \mid s_1) \frac{X_k^2}{\pi_k}, \qquad (4.1.2)$$

avec:

$$\mathrm{Var}(\hat{a}_k \mid s_1) = \mathrm{Var} \frac{\sum_{s_k} \frac{Y_i}{\pi_{i|k}}}{\sum_{s_k} \frac{X_i}{\pi_{i|k}}} = \frac{1}{X_k^2} \mathrm{Var} \sum_{s_k} \frac{Y_i - a_k X_i}{\pi_{i|k}}$$

$$= \frac{1}{X_k^2}\left(\sum_{i \in k} \frac{(Y_i - a_k X_i)^2}{\pi_{i|k}}\right.$$

$$\left. + \sum_{k \neq l}\sum \frac{(Y_i - a_k X_i)(Y_j - a_k X_j)\pi_{ij|k}}{\pi_{i|k}\ \pi_{j|k}}\right).$$

Comme dans les parties précédentes, nous arrivons à des formules complexes et, finalement, inutilisables. Un modèle va nous simplifier un peu l'existence.

### 4.2 Intervention d'un modèle

Il aura la même structure que ceux qui ont déjà servi antérieurement:

a) Les $a_k$ seront des variables aléatoires indépendantes de même espérance et de même variance:

$$E_\xi\, a_k = a \qquad \mathrm{Var}_\xi\, a_k = \sigma^2.$$

La variance prend en compte l'influence de l'opérateur ou de l'opératrice, qu'on renonce à isoler, mais aussi celle du jour de la semaine, de l'heure dans la journée, de certains jours du mois *etc.*...

b) Conditionnellement à $a_k$, les $Y_i$ de l'Unité Primaire $k$ sont des variables de Bernoulli indépendantes avec $E_\xi(Y_i \mid k) = a_k X_i$

$$\mathrm{Var}_\xi(Y_i \mid k) = a_k X_i - a_k^2 X_i^2.$$

**Remarque:**
La variable $X_i$ n'a pas de véritable sens concret et n'est d'ailleurs définie qu'à un facteur d'échelle près. En revanche $aX_i$ et $\sigma X_i$ ont une interprétation physique invariante, car ce sont des probabilités. Dans tout ce qui suit il faudra toujours garder en tête que les résultats devront être invariants si les $X_i$ sont multipliés par un facteur arbitraire à condition que $a$ et $\sigma$ soient divisés par le même facteur. En particulier $\mathrm{Var}(\hat{a})$ n'a pas de sens "concret". Seule $\mathrm{Var}(\hat{a}X)$ en a un.

Comme précédemment nous allons étudier la variance anticipée, espérance sous modèle de la somme de (4.1.1) et (4.1.2).

Pour le premier terme, l'espérance des produits croisés est nulle, comme de bien entendu. L'espérance sous modèle de ce terme est donc:

$$X^{-2}\sigma^2 \sum_k \frac{X_k^2}{\pi_k}.$$

Pour le second terme on trouve: (vu les définitions données au 4.2.a et 4.2.b):

$$X^{-2} \sum_k \frac{X_k^2}{\pi_k} \cdot \frac{1}{X_k^2} \sum_i E_\xi \frac{(a_k X_i - a_k^2 X_i^2)}{\pi_{i|k}}$$

$$= X^{-2} \sum_k \frac{1}{\pi_k} \sum_i \frac{aX_i - (a^2 + \sigma^2)X_i^2}{\pi_{i|k}}.$$

Globalement donc:

$$E_\xi\, \mathrm{Var}(\hat{a}X) = \sigma^2 \sum_{k \in U_o} \frac{X_k^2}{\pi_k}$$

$$+ \sum_{k \in U_o} \frac{1}{\pi_k} \sum_{i \in k} \frac{aX_i - (a^2 + \sigma^2)X_i^2}{\pi_{i|k}}.$$

Ici pas de miracle algébrique. *Pour simplifier* nous admettons que $(a^2 + \sigma^2)X_i^2$ est négligeable devant $aX_i$. Numériquement on peut attendre $aX_i = 2$ à $5 \times 10^{-2}$ et $(a^2 + \sigma^2)X_i^2 = 3$ à $30 \times 10^{-4}$. D'où notre approximation

$$E_\xi\, \mathrm{Var}(\hat{a}X) \simeq \sigma^2 \sum_{k \in U_o} \frac{X_k^2}{\pi_k} + a \sum_{k \in U_o} \frac{1}{\pi_k} \sum_{i \in k} \frac{X_i}{\pi_{i|k}}.$$

### 4.3 Optimisation du plan de sondage

Nous utiliserons la fonction de coût suivante:

$$C = \sum_{s_1} (C_o + C_1 n_k).$$

Ici, $n_k = \sum_{i \in k} \pi_{i|k}$ est la taille de l'échantillon tiré dans le district $k$ (supposé de taille fixe à $s_1$ fixé). Son espérance vaut:

$$C_T = \sum_{k \in U_o} \pi_k (C_o + C_1 n_k).$$

Posons

$$\pi_{i|k} = n_k P_i \left( \text{avec} \sum_{i \in k} P_i = 1 \right) \quad \text{et} \quad Q_k = \pi_k n_k.$$

Le problème d'optimisation est maintenant:

$$\text{Min: } C_o \sum_k \pi_k + C_1 \sum_k Q_k$$

$$\text{sous: } \sigma^2 \sum_k \frac{X_k^2}{\pi_k} + a \sum_k \frac{1}{Q_k} \sum_{i \in k} \frac{X_i}{P_i} \leq \mathcal{V}_o.$$

Sous cette forme, on constate avec plaisir qu'on peut minimiser les termes en $\sum_i X_i / P_i$ indépendamment du reste. Autrement dit, $n_k$ n'a pas d'incidence sur ce terme. Laissons l'optimisation du second degré de tirage pour plus tard et notons seulement $S_k^{*2}$ la valeur optimisée de $\sum_i X_i / P_i$. Avec un multiplicateur de Lagrange $\lambda$ on obtient par dérivation par rapport aux $\pi_k$ puis au $Q_k$:

$$*C_o = \lambda \sigma^2 \frac{X_k^2}{\pi_k^2} \quad \text{soit} \quad \boxed{\pi_k \text{ proportionnel à } X_k} \quad (4.3.1)$$

$$*C_1 = \lambda a \frac{S_k^{*2}}{Q_k^2} \quad \text{d'où} \quad \boxed{n_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{S_k^*}{X_k}} \quad . (4.3.2)$$

En particulier on tirera les Unités Primaires avec des probabilités proportionnelles à leur difficulté totale, ce qui est un résultat standard (voir par exemple Särndal, Swensson, Wretman 1992, chapitre 12).

Passons maintenant au tirage infra-district (deuxième degré de sondage).

Commençons par un cas simple et naïf: on tire les bulletins individuellement. La minimisation conduit à $P_i$ proportionnelle à $\sqrt{X_i}$. Un calcul simple nous montre qu'alors $S_k^* = \sum_{i \in k} \sqrt{X_i}$. Ceci nous permet de calculer $n_k$ grâce à (4.3.2) et notre problème est entièrement résolu.

En fait les choses sont plus compliquées. Pour des raisons assez naturelles, on ne sélectionnera que les bulletins que de ménages entiers. Autrement dit le sondage au second degré est un sondage en *grappes*. Les valeurs de $P_i$ seront les mêmes, soit $P_m$, pour tous les membres d'une même grappe (ménage) $m$.

Notons par $X_m$ la somme des $X_i$ des individus $i$ du ménage $m$. Le problème est donc de minimiser $\sum X_m / P_m$ sous $\sum n_m P_m = 1$ avec $n_m$ taille du ménage $m$. On trouve facilement la solution:

$$P_m = \sqrt{\bar{X}_m} \Big/ \sum n_m \sqrt{\bar{X}_m},$$

avec $\bar{X}_m = X_m / n_m$, *difficulté moyenne des bulletins du ménage m*. Par suite on trouve $S_k^* = \sum n_m \sqrt{\bar{X}_m}$.

Cette solution nous permet de déterminer le nombre $n_k$ *d'unités finales* à tirer grâce à (4.3.2). Le nombre de grappes (*ménages*), en revanche n'est pas déterminé. Cette difficulté était prévisible. La fonction de coût, en effet, ne fait pas intervenir cette contrainte. Pour obtenir le nombre $m_k$ de grappes à tirer, on s'arrangera de façon à ce que l'espérance du nombre d'unités finales soit égale à $n_k$. Elle vaut:

$$m_k \left( \sum n_m \sqrt{\bar{X}_m} \right) \Big/ \sum \sqrt{\bar{X}_m}$$

d'où:

$$m_k = n_k \frac{\sum \sqrt{\bar{X}_m}}{\sum n_m \sqrt{\bar{X}_m}}.$$

Compte tenu de (4.3.2) on a aussi:

$$m_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{\sum \sqrt{\bar{X}_m}}{X_k}$$

et la probabilité de tirer un ménage vaut alors:

$$\frac{m_k \sqrt{\bar{X}_m}}{\sum \sqrt{\bar{X}_m}}.$$

Après quelques manipulations algébriques, on trouve la valeur de la variance optimale:

$$E_\xi \operatorname{Var}(\hat{a}X)_{\text{OPT}} = \frac{(\sigma X)^2}{m} \left( 1 + \frac{a}{\sigma} \frac{a^{-1/2} S^*}{X} \left(\frac{C_1}{C_o}\right)^{1/2} \right).$$

Cette forme respecte le caractère homogène des différents facteurs. On a, en particulier $a^{-1/2} S^* / X = a^{1/2} S^* / aX$; le dénominateur est interprétable comme un nombre total d'erreurs dans un lot, tandis que le numérateur est homogène à une taille.

Nous avons obtenu une solution complète du problème.

**Remarque 1:**
Dans les deux cas qui ont été traités $S_k^*$ est multiplié par $C^{1/2}$ si les $X_i$ sont multipliés par $C$. La formule qui donne $n_k$ est donc bien invariante à l'échelle de mesure.

**Remarque 2:**
La solution dans le cas du tirage en grappes privilégie le tirage de petites grappes dont les unités finales on un fort indice de difficulté.

**Remarque 3:**
Ici comme dans les parties précédentes nous déterminons les probabilités d'inclusion simple mais pas les probabilités d'inclusion double. L'algorithme de tirage, qui fixe ces dernières, est donc sans influence. Ceci est relativement naturel si nous nous disons que l'information auxiliaire utilisée pour optimiser le tirage déterminera les $\pi_k$ et $\pi_{i|k}$ mais ne peut pas avoir d'influence sur les probabilités doubles.

## 5. APPLICATIONS AUX CONTRÔLES PAR SONDAGE DE LA QUALITÉ DU RECENSEMENT DE 1990 EN FRANCE

### 5.1 Problème de contrôle de la saisie

Les techniques d'échantillonnage décrites aux paragraphes 2 et 3 ont été motivées par la nécessité de contrôler la saisie du recensement de 1990. Pour comprendre la nature des problèmes statistiques une description des principes d'exploitation est utile.

L'unité de base pour la collecte est le district, correspondant, en ville, a un pâté de maisons et, à la campagne, à un village ou une réunion de hameaux. Il peut comporter une population variant de zéro à environ deux mille habitants. La moyenne est de 150 logements pour 350 habitants environ.

À mesure de l'achèvement de la collecte et de sa vérification, les différents bulletins du recensement, notamment les bulletins individuels (BI) et les feuilles de logement (FL), sont soigneusement comptés pour chaque district. Les données récapitulatives des districts sont saisies sur support informatique, tandis que les bulletins groupés, dans une chemise de district, partent pour la saisie.

Des ensembles de districts groupant environ 100,000 logements sont constitués. Ce sont les unités de traitement (UT). Chaque UT est saisie par une entreprise à façon pour le compte de l'INSEE.

L'INSEE, le "client" en termes de théorie du contrôle, vérifie la qualité du travail de chaque façonnier en contrôlant par sondage un certain nombre de bulletins dans chaque UT.

Le but du sondage décrit au paragraphe 2 est d'estimer la proportion de bulletins erronnés dans chaque UT avec une précision (écart-type) de un point. La proportion maximum de bulletins erronnés ne saurait excéder 4%. Un test de recensement portant sur environ 400 districts permet d'estimer les valeurs des deux paramètres du modèle. On trouve:

$$\sigma^2 \simeq P^2 \simeq 14.10^{-4}$$

$$\tau^2 \simeq P \simeq 4.10^{-2}.$$

La fonction de coût (1.1) a pu être évaluée en temps de travail. Les mesures faites dans les ateliers ont permis d'estimer à 5 minutes le temps de manipulation d'une chemise de district (du moment où on va la chercher dans une étagère au moment où on la range) et à 30 secondes le temps de saisie d'un BI. Avec des données numériques, l'optimisation du plan avec les hypothèses du paragraphes 1 conduit à contrôler 40 districts par lot de traitement et 16 bulletins à l'intérieur de chaque district.

Après discussion de cette solution avec l'équipe responsable du recensement, il est apparu qu'il fallait, en fait, contrôler deux types de documents: les bulletins individuels (BI) et les feuilles de logement (FL). On avait été conduit à négliger ces dernières, en première approximation, parce qu'elles sont moins susceptibles de recéler des erreurs et que leur temps de codification est plus court (la moitié environ) que celui nécessaire pour un BI. Toutefois, dans certains districts, par exemple dans les communes très touristiques, on trouve une forte majorité de résidences secondaires et donc beaucoup de FL pour très peu de BI. Cette situation demande une étude particulière, dont la théorie a été faite au paragraphe 3.

Dans le cas du recensement le nombre $G$ de groupes vaut 2 ($g = 1$ pour les BI et $g = 2$ pour les FL). Les données numériques relatives aux deux groupes étaient les suivantes:

$$\cdot\ P_1 = 0{,}04 \qquad \sigma_1 = P_1 \qquad \tau_1^2 = P_1(1 - P_1) - \sigma_1^2 = P_1 - 2P_1^2,$$

$$\cdot\ P_2 = 0{,}01 \qquad \sigma_2 = P_2 \qquad \tau_2^2 = P_2 - P_2^2,$$

$$\cdot\ \mathcal{V}_1 = (0{,}0075)^2 \qquad \mathcal{V}_2 = (0{,}0150)^2.$$

Pour la fonction de coût on a pris $C_o = 5$ minutes, $C_1 = 0{,}5$ minute et $C_2 = 0{,}25$ minute. L'optimisation du problème selon les hypothèses de la partie 3.2.b a conduit à examiner 73 districts par unité de traitement. La solution pratique à consisté à traiter 15 bulletins individuels par distcrit ainsi que les FL associées. Pour les districts comportant moins de 15 BI, l'intégralité des BI était traitée. Pour les districts vides de BI on traitait 4 FL (si ce nombre était inférieur au nombre de FL du district).

**Remarque:**
La technique de la partie 2 semble avoir un domaine d'application assez fréquent. Elle a, en particulier, été utilisée pour l'échantillonnage de l'enquête française de 1992 sur les migrations en ce qui concerne la population de nationalité étrangère. Pour les agglomérations de moins de 20,000 habitants, l'échantillon était à deux degrés, le premier degré de sondage étant constitué par les 90 départements où ce type d'agglomération existe. La population étrangère (sur la base du recensement) était divisée en 8 groupes de nationalités pour lesquels on devrait obtenir des indicateurs ayant la même précision.

### 5.2 Problèmes liés à la codification

La seconde étape d'élaboration des données est dite opération COLIBRI (pour Codification en Ligne des Bulletins du Recensement des Individus). Recevant des bulletins toujours groupés en districts, les opérateurs et opératrices des Directions Régionales de l'INSEE procédaient à leur codification pour constituer le sondage au quart.

Physiquement, chaque opérateur (trice) travaille devant un écran qui lui indique l'identifiant du prochain logement à inclure dans l'échantillon au quart dont il doit codifier tous les BI.

Le contrôle de la qualité de la codification est également réalisé par sondage. L'unité de contrôle est l'ensemble du travail réalisé en une semaine dans une Direction Régionale. L'opération dure un peu plus d'un an dans les 22 Directions Régionales soit plus de mille sondages. L'unité à contrôler est le ménage (c'est-à-dire l'ensemble des BI d'un ménage tiré pour figurer dans l'échantillon de contrôle). L'objectif est d'estimer la proportion de bulletins comportant une erreur. Pour cela, on détecte automatiquement ceux pour lesquels apparaît une divergence entre les deux codifications. Une opération de réconciliation permet de chiffrer le nombre d'erreurs. La théorie de ce contrôle a fait l'objet de la partie 4 de cet article. L'indice de difficulté des bulletins a été élaboré à partir des données déjà saisies pour une étude faite à partir du précédent recensement et d'un test. Les modalités pratiques et les enseignements tirés de ces contrôles sont détaillés dans G. Badeyan (1992).

L'application pratique et numérique de la théorie repose sur des hypothèses concernant les ordres de grandeurs des différents paramètres (ce qui demande qu'on puisse les raccrocher à une interprétation physique simple). Dans la phase de préparation du recensement, sans mesures préalables très précises, on a utilisé les valeurs $\sigma/a = 0{,}5$ et $C_1/C_o = 0{,}1$.

À la suite de diverses hypothèses sur les autres paramètres et de discussions entre experts, il a été décidé un contrôle portant sur 50 districts chacun d'eux étant contrôlé pour environ 20 BI (par région et par semaine). Cet ordre de grandeur initial pouvait, évidemment, être modulé dans la suite du contrôle, les paramètres du modèle pouvant être réestimés après chacun d'eux.

**Remarque finale:**

Ce problème fait apparaître des résultats un peu surprenants sur lesquels il est utile de réfléchir un peu.

Dans un premier cas, nous avons supposé qu'on pouvait isoler chaque bulletin. On tirait alors ceux-ci avec des probabilités proportionnelles à leur difficulté individuelle. On supposait, dans une certaine mesure, que le coût d'utilisation de l'information individuelle était nul.

Dans le second cas, la réalité du contrôle, ce coût était considéré comme infini et la seule information ayant un coût négligeable était celle relative à l'ensemble du ménage. La solution fait alors apparaître des probabilités de tirage des individus (BI) qui est fonction de la difficulté moyenne de codification des bulletins de l'ensemble du ménage auquel appartient cet individu.

Il en irait de même en ce qui concerne le tirage des districts. Si on sait y distinguer les BI, on les tire avec des probabilités proportionnelles à la difficulté totale; à l'intérieur des districts, on tirera des BI difficiles avec une plus grosse probabilité. Supposons, au contraire, qu'on ne sache pas distinguer les BI à l'intérieur des districts. Ce serait le cas, par exemple, si la désignation des BI à contrôler ne pouvait pas se faire en temps réel par suite d'une organisation du traitement inadéquate. On tirerait alors les districts proportionnellement à leur difficulté moyenne: à l'intérieur des districts, on serait obligé de réaliser des sondages aléatoires simples.

Dans le premier cas le sondage privilégiera les gros districts à l'intérieur desquels on tirera plutôt les BI difficiles. Dans le second cas, on privilégiera les petits districts difficiles à l'intérieur desquels on tirera des bulletins à probabilités égales. *Dans les deux cas*, on cherchera à augmenter la probabilité de sonder des BI difficiles. La différence réside simplement dans la possibilité (c'est-à-dire le coût) de mobiliser l'information au moment où on en a besoin.

## REMERCIEMENTS

L'auteur tient à remercier de leurs commentaires très positifs le rédacteur en chef, le rédacteur associé et l'arbitre qui ont examiné ce travail. Il en va de même de Claude Thelot dont les remarques sont à l'origine d'un certain nombre de développements et de Gérard Badeyan qui a mis en place à l'INSEE les techniques ici préconisées. Il remercie Françoise Hitier sans qui ce texte n'aurait jamais pu exister.

## BIBLIOGRAPHIE

BADEYAN, G. (1992). Communication aux secondes Journées de Méthodologie Statistique, 17 et 18 juin 1992, INSEE, Paris.

COCHRAN, W. (1977). *Sampling Techniques*, (3ième édition). New York: Wiley.

DESABIE, J. (1965). *Théorie et Pratique des Sondages*. Paris: Dunod.

LUENBERGER, D.G. (1973) *Introduction To linear and Non-linear Programming*. New York: Addison-Wesley.

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

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 193–203
Statistique Canada



# Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale

ROBERT J. CASADY et RICHARD VALLIANT[1]

RÉSUMÉ

La stratification a posteriori est une technique courante d'amélioration de la précision des estimateurs, qui consiste à utiliser des éléments d'information qui n'étaient pas disponibles au moment de la préparation du plan de l'enquête. Pour des échantillons vastes et complexes, le vecteur des estimateurs de Horvitz-Thompson des variables d'intérêt de l'enquête et des tailles de la population des strates a posteriori suivra approximativement, dans des conditions appropriées, une distribution normale multidimensionnelle. Cette normalité pour de grands échantillons amène à définir un nouvel estimateur de régression fondé sur une stratification a posteriori, analogue à l'estimateur de régression linéaire dans le cas de l'échantillonnage aléatoire simple. Nous calculons, pour de grands échantillons, le biais et les erreurs quadratiques moyennes selon le plan de ce nouvel estimateur, de l'estimateur de stratification a posteriori courant, de l'estimateur de Horvitz-Thompson et d'un estimateur par quotient. Nous utilisons des populations réelles et une population artificielle pour étudier empiriquement les propriétés conditionnelles et non conditionnelles des estimateurs en vertu d'un échantillonnage à plusieurs degrés.

MOTS CLÉS: Normalité asymptotique; estimateur de régression; bases de sondage défectueuses; estimateur par quotient; estimateur de Horvitz-Thompson.

## 1. INTRODUCTION

### 1.1 Généralités

Les études sur la théorie de l'échantillonnage au cours des vingt dernières années ont été axées dans une large mesure sur la recherche de moyens de restreindre l'ensemble d'échantillons servant de base à l'inférence. Dans une approche fondée purement sur le plan de sondage, comme le décrivent Hansen, Madow et Tepping (1983), aucune restriction de ce genre n'est imposée. Les propriétés statistiques sont calculées sous forme de moyennes sur l'ensemble des échantillons qui auraient pu être sélectionnés en vertu d'un plan particulier. Bien qu'on admette généralement qu'une certaine forme d'inférence conditionnelle, fondée sur le plan, soit souhaitable (Fuller 1981, Rao 1985, Hidiroglou et Särndal 1989), aucune théorie satisfaisante n'a encore été élaborée, sauf dans des cas relativement simples. Il existe d'autres solutions comme la théorie de la prédiction, mise au point par Royall (1971) et de nombreux autres, ainsi que l'approche bayesienne, décrite par Ericson (1969), qui évite d'établir des moyennes sur des échantillons répétés, grâce à l'utilisation de modèles de superpopulation. Une méthode d'inférence conditionnelle fondée sur le plan a été présentée par Robinson (1987) pour le cas particulier des estimations par quotient dans les enquêtes par sondage. Robinson a utilisé la théorie des grands échantillons et la normalité approximative de certaines statistiques pour produire une théorie d'inférence conditionnelle, fondée sur le plan, pour l'estimateur par quotient.

Dans la présente communication, nous étendons ce genre de raisonnement au problème de la stratification a posteriori. Durbin (1969), Holt et Smith (1979) et Yates (1960) ont fait valoir de façon convaincante que des échantillons de stratification a posteriori devraient faire l'objet d'une analyse conditionnelle, c.-à-d. en fonction de la distribution d'échantillonnage des unités parmi les strates a posteriori. Toutefois, comme l'a signalé Rao (1985), les difficultés de formuler une théorie exacte, fondée sur le plan, visant des échantillons finis, à l'égard de la stratification a posteriori dans les plans d'échantillonnage généraux pourraient être insurmontables. Des analyses conditionnelles, fondées sur le modèle, d'échantillons de stratification a posteriori sont présentées dans Little (1991) et Valliant (1993). La voie empruntée ici est fondée sur le plan et utilise les grands échantillons et la normalité approximative d'une manière semblable à celle de Robinson (1987) afin de permettre l'étude des propriétés conditionnelles des estimateurs.

### 1.2 Définitions de base et notation

La **population cible** est un ensemble bien défini d'unités élémentaires (ou analytiques). Dans de nombreuses applications, les unités élémentaires sont des personnes ou des établissements. Nous supposons que la population cible a été divisée en **unités d'échantillonnage du premier degré** (UPD). Pour les enquêtes auprès des personnes, les UPD sont en général des ménages, des groupes de ménages ou même des comtés, tandis que pour les enquêtes auprès des établissements, il n'est pas rare que l'établissement individuel soit une UPD. Quoi qu'il en soit, l'ensemble des UPD sera désigné comme la **base de sondage du premier degré** (ou simplement **la base de sondage**). Nous supposons

[1] Robert J. Casady et Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington D.C., 20212-0001.

qu'il y a $M$ UPD dans la base de sondage et qu'elles sont dénotées 1, 2, ... $M$. Nous supposons aussi que les unités de la population peuvent être réparties en $K$ "strates a posteriori" qui peuvent servir à des fins d'estimation.

Soit $y$ la valeur de la caractéristique d'intérêt (p. ex. revenu hebdomadaire, nombre d'heures travaillées la semaine précédente, jours d'activité restreinte au cours des deux dernières semaines, *etc.*) pour une unité élémentaire. À la $i$-ième UPD sont associés $2K$ nombres réels:

$y_{ik}$ = somme des valeurs $y$ pour les unités élémentaires de la $i$-ième UPD se trouvant dans la $k$-ième strate a posteriori,

$N_{ik}$ = nombre d'unités élémentaires de la $i$-ième UPD se trouvant dans la $k$-ième strate a posteriori.

Pour chaque strate a posteriori, nous définissons ensuite:

$Y_{.k} = \sum_{i=1}^{M} y_{ik}$ = somme des valeurs $y$ pour l'ensemble des unités élémentaires dans la $k$-ième strate a posteriori,

$N_{.k} = \sum_{i=1}^{M} N_{ik}$ = nombre total d'unités élémentaires dans la $k$-ième strate a posteriori.

Nous supposons, dans ce qui suit, que les $N_{.k}$ sont des valeurs fixes connues. Dans certaines enquêtes, les $N_{.k}$ peuvent en fait être eux-mêmes des estimations, mais notre analyse considère comme donné l'ensemble de $N_{.k}$ utilisés dans l'estimation. Dans la Current Population Survey des États-Unis, par exemple, chaque $N_{.k}$ est un nombre de personnes projeté d'après le recensement décennal précédent au moyen de méthodes démographiques. La somme pour la population des valeurs $y$ est donnée par $Y_{..} = \sum_{k=1}^{K} Y_{.k}$ et la taille totale de la population, par $N_{..} = \sum_{k=1}^{K} N_{.k}$. Dans les sections 1 à 3, nous supposons que la base de sondage "couvre" la totalité de la population cible. À la section 4, nous examinons le problème posé par une base défectueuse, c.-à-d. une base dont la couverture ne correspond pas à l'ensemble de la population cible.

### 1.3 Plan de sondage et estimation de base

Supposons que la base de sondage du premier degré soit divisée en $L$ strates et qu'un plan stratifié à plusieurs degrés soit utilisé, avec un échantillon total de $m$ UPD. Dans ce qui suit, nous avons supprimé l'indice représentant les strates du plan, de manière à simplifier la notation. Pour les besoins de l'exposé théorique subséquent, il est inutile de définir explicitement les méthodes d'échantillonnage et d'estimation utilisées au deuxième degré et aux degrés suivants du plan. Toutefois, pour chaque UPD de l'échantillon, nous avons besoin d'estimateurs $\hat{y}_{ik}$ et $\hat{N}_{ik}$ tels que $E_{2+}[\hat{y}_{ik}] = y_{ik}$ et $E_{2+}[\hat{N}_{ik}] = N_{ik}$, où la notation $E_{2+}$ indique l'espérance selon le plan pour le deuxième degré et les degrés supérieurs. Soit $\pi_i$ la probabilité que la $i$-ième UPD soit incluse dans l'échantillon et $w_i = 1/\pi_i$; il s'ensuit que l'estimateur $\hat{Y}_{.k} = \sum_{i=1}^{m} w_i \hat{y}_{ik}$ est non biaisé pour $Y_{.k}$ et que l'estimateur $\hat{N}_{.k} = \sum_{i=1}^{m} w_i \hat{N}_{ik}$ est non biaisé pour $N_{.k}$.

### 1.4 Résultat asymptotique comme celui de Robinson

Robinson (1987) a étudié l'estimateur par quotient $(\bar{X}/\bar{x}_s)\bar{y}_s$ en vertu d'un échantillonnage aléatoire simple, où $\bar{y}_s$ est la moyenne de l'échantillon d'une variable d'intérêt $y$, $\bar{x}_s$ est la moyenne de l'échantillon d'une variable auxiliaire $x$, et $\bar{X}$ est la moyenne de $x$ pour la population totale. Dans certaines conditions, $(\bar{y}_s, \bar{x}_s)$ sera asymptotiquement une distribution normale à deux variables, pour de grands échantillons aléatoires simples. Il découle des résultats de Robinson que l'estimateur de régression linéaire $\bar{y}_s + \beta(\bar{X} - \bar{x}_s)$ est asymptotiquement non biaisé selon le plan, pour $\bar{x}_s$ donné. Ce résultat est étendu dans la présente section à des échantillons complexes.

Conformément à l'approche de Krewski et Rao (1981), nous pouvons établir nos résultats asymptotiques quand $L \to \infty$ dans un cadre formé d'une séquence de populations finies $\{\Pi_L\}$ avec $L$ strates dans $\Pi_L$. Il est à souligner que nous supposons implicitement (sans énoncé formel) l'existence des conditions précisées dans Krewski et Rao, et développées davantage dans Rao et Wu (1985), en ce qui concerne le plan de sondage et la régularité. Les détails des preuves apportent peu de nouveaux éléments par rapport à ce qu'on trouve dans la littérature, et nous les avons omis.

En notation matricielle, nous posons $Y = [Y_{.1} \ldots Y_{.k}]'$, $N = [N_{.1} \ldots N_{.k}]'$, $\hat{Y} = [\hat{Y}_{.1} \ldots \hat{Y}_{.k}]'$, $\hat{N} = [\hat{N}_{.1} \ldots \hat{N}_{.k}]'$ et $V = \mathrm{var}\{[\hat{\bar{Y}}\ \hat{\bar{N}}]'\}$, où $\hat{\bar{Y}} = (1/N_{..})\hat{Y}$ et $\hat{\bar{N}} = (1/N_{..})\hat{N}$. Notons que $\hat{\bar{Y}}$, qui comprend $N_{..}$ au dénominateur, est utilisé pour faciliter la notation, et n'est pas une estimation des moyennes pour les strates a posteriori. Comme le faisaient Krewski et Rao (1981) dans leurs conditions C4 et C5, nous supposons que:

$$\lim_{L\to\infty} \frac{Y_{.k}}{N_{.k}} = \mu_k, \quad \text{pour} \quad k = 1, 2, \ldots, K, \tag{1}$$

$$\lim_{L\to\infty} \frac{N_{.k}}{N_{..}} = \phi_k > 0 \quad \text{pour} \quad k = 1, 2, \ldots, K, \text{ et} \tag{2}$$

$$\lim_{L\to\infty} mV = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \text{ (définie positive)}, \tag{3}$$

où $\Sigma$ est partitionnée de la manière évidente. Signalons qu'encore une fois, nous avons supprimé l'indice représentant les strates du plan. Les hypothèses (1) à (3) exigent simplement que certaines quantités clés se stabilisent dans de grandes populations. La condition (2), en particulier, assure qu'aucune strate a posteriori n'est vide à mesure que s'accroît la taille de la population. Nous énonçons maintenant ce qui suit:

**Résultat**: Sous réserve des conditions précisées par Krewski et Rao concernant le plan de sondage et la régularité et de l'existence de $\Sigma_{22}^{-1}$, alors, pour $\hat{\bar{N}}$ donné, la distribution conditionnelle de $\hat{\bar{Y}}$ est asymptotiquement $\mathfrak{N}(M_1 + \Sigma_{12}\Sigma_{22}^{-1}(\hat{\bar{N}} - M_2), m^{-1}V_c)$, où $V_c = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $M_1 = \lim_{L\to\infty} \hat{\bar{Y}} = [\phi_1\ \mu_1\ \ldots\ \phi_K\ \mu_K]'$ et $M_2 = \lim_{L\to\infty} \hat{\bar{N}} = [\phi_1 \ldots \phi_K]'$.

**Preuve**: Ce résultat est analogue au résultat pour $K = 1$ présenté par Robinson (1987) et découle directement du fait que le vecteur aléatoire

$$m^{1/2}\begin{bmatrix} \hat{\bar{Y}} - M_1 - \Sigma_{12}\Sigma_{22}^{-1}(\hat{\bar{N}} - M_2) \\ \hat{\bar{N}} - M_2 \end{bmatrix}$$

tend, en distribution, vers

$$\mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} V_c & 0 \\ 0 & \Sigma_{22} \end{bmatrix}\right).$$

En termes stricts, comme dans Robinson, nous considérons la distribution conditionnelle de $\hat{\bar{Y}}$ pour $\hat{\bar{N}}$ dans une cellule de taille $\epsilon m^{-1/2}$ pour un petit $\epsilon$. Notons que dans certains plans d'échantillonnage $\mathbf{1}'\hat{N} = N_{..}$ (par exemple ceux dans lesquels un nombre fixe d'unités élémentaires sont sélectionnées avec probabilités égales), auquel cas $\Sigma_{22}^{-1}$ n'existe pas; on ne considère alors, pour l'énoncé conditionnel, que les $K - 1$ premières strates a posteriori.

Dans la section qui suit, la moyenne asymptotique de $\hat{\bar{Y}}$ est utilisée comme source d'un estimateur de régression linéaire de la moyenne des $y$ pour la population.

## 2. PROPRIÉTÉS CONDITIONNELLES D'ESTIMATEURS DE LA MOYENNE DE LA POPULATION

### 2.1 Estimateurs de la moyenne de la population

La **moyenne de la population** est, par définition,

$$\mu = \lim_{L\to\infty}(Y_{..}/N_{..}) = \lim_{L\to\infty}(\mathbf{1}'Y/\mathbf{1}'N) = \sum_{k=1}^{K}\phi_k\mu_k$$

où $\mathbf{1}'$ est un vecteur ligne comprenant $K$ fois le chiffre un. Notons que la moyenne $\mu$ n'est pas un paramètre de population finie, mais plutôt une valeur limite. Dans les grandes populations ($L\to\infty$), $\mu$ et la moyenne réelle de population finie seront arbitrairement voisines l'une de l'autre. Quatre estimateurs de la moyenne de la population seront examinés. Les trois premiers sont des estimateurs courants qu'on trouve dans la littérature, tandis que le quatrième est un nouvel estimateur dont la construction est suggérée par la normalité conjointe asymptotique de $\hat{\bar{Y}}$ et $\hat{\bar{N}}$:

1) Estimateur de Horvitz-Thompson

$$\hat{\bar{Y}}_{HT} = \mathbf{1}'\hat{Y}/\mathbf{1}'N = \mathbf{1}'\hat{\bar{Y}}.$$

2) Estimateur par quotient

$$\hat{\bar{Y}}_R = \mathbf{1}'\hat{Y}/\mathbf{1}'\hat{N} = \mathbf{1}'\hat{\bar{Y}}/\mathbf{1}'\hat{\bar{N}}.$$

3) Estimateur de stratification a posteriori

$$\hat{\bar{Y}}_{PS} = N_{..}^{-1}\sum_{k=1}^{K}\left(\frac{N_{.k}}{\hat{N}_{.k}}\right)\hat{Y}_{.k} = r'\hat{\bar{Y}}$$

où

$$r' = [N_{.1}/\hat{N}_{.1}, \ldots, N_{.K}/\hat{N}_{.K}].$$

4) Estimateur de régression linéaire

$$\hat{\bar{Y}}_{LR} = [\mathbf{1}'(\hat{\bar{Y}} - \Sigma_{12}\Sigma_{22}^{-1}(\hat{\bar{N}} - M_2))].$$

L'estimateur de régression linéaire s'appuie sur la forme de la moyenne pour grands échantillons de la variable aléatoire conditionnelle $\hat{\bar{Y}} \mid \hat{\bar{N}}$ présentée à la fin de la section 1.4, et est très semblable à l'estimateur de régression généralisée examiné par Särndal, Swensson et Wretman (1992). L'estimateur de régression linéaire (4) a également été examiné par Rao (1992) dans le contexte de l'estimation d'étalonnage. Il est à remarquer que l'estimateur par quotient n'exige pas que les $N_{.k}$ ou leur somme $N_{..}$ soient connus. L'estimateur de Horvitz-Thompson exige seulement que $N_{..}$ soit connu, tandis que les estimateurs de stratification a posteriori et de régression linéaire exigent que $\{N_{.k} | k = 1, \ldots, K\}$ soit connu. En pratique, l'estimateur de régression linéaire présente une complication additionnelle du fait que les matrices de covariance $\Sigma_{12}$ et $\Sigma_{22}$ sont inconnues et doivent être estimées à partir de l'échantillon. Dans la mise en application de $\hat{\bar{Y}}_{LR}$ à la section 3, les quantités connues des populations finies $(1/N_{..})N$ seront utilisées en remplacement du vecteur limite $M_2$.

### 2.2 Espérances et variances conditionnelles des estimateurs

À l'aide des expressions asymptotiques données plus haut, les espérances et les variances conditionnelles, pour $\hat{\bar{N}}$ donné, des quatre estimateurs peuvent être calculées. Dans le cas de la stratification a posteriori, le choix de $\hat{\bar{N}}$ comme condition donnée dans un plan complexe est un prolongement naturel du choix des tailles d'échantillon réalisées des strates a posteriori en vertu d'un échantillonnage aléatoire simple. Dans d'autres situations, toutefois, il peut se révéler difficile de déterminer quoi choisir comme condition donnée, et la réponse peut ne pas être unique (voir, p. ex., Kiefer 1977). Premièrement, définissons les trois matrices suivantes:

$$H = \Sigma_{12}\Sigma_{22}^{-1},$$

$$R = H - D(\mu), \quad \text{et}$$

$$P = H - D(\mu_k),$$

où $D(\mu) = \mathrm{diag}(\mu, \ldots, \mu)$ et $D(\mu_k) = \mathrm{diag}(\mu_1, \ldots, \mu_k)$ sont des matrices diagonales $K \times K$. Nous présentons ci-dessous la moyenne et la variance des quatre estimateurs

sans fournir aucun détail des calculs. Quand l'échantillon des unités du premier degré est grand, chacun des estimateurs a essentiellement la même variance conditionnelle. Les estimateurs de Horvitz-Thompson, par quotient et de stratification a posteriori sont, toutefois, conditionnellement biaisés, tandis que l'estimateur de régression linéaire ne l'est pas. Ainsi, l'estimateur de régression linéaire est celui qui affiche l'erreur quadratique moyenne asymptotique la plus faible parmi les quatre estimateurs qui nous intéressent. Rao (1992) a également signalé le caractère optimal de l'estimateur de régression à l'intérieur d'une certaine classe d'estimateurs de différences, ainsi que son biais négligeable pour de grands échantillons.

1) Estimateur de Horvitz-Thompson:

$$\mathrm{E}[\hat{Y}_{HT} \mid \hat{\bar{N}}] = \mu + [\mathbf{1}'\boldsymbol{H}(\hat{\bar{N}} - \boldsymbol{M}_2)]$$

$$\mathrm{var}[\hat{Y}_{HT} \mid \hat{\bar{N}}] = m^{-1}[\mathbf{1}'(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\mathbf{1}]$$

$$= m^{-1}[\mathbf{1}'\boldsymbol{V}_c\mathbf{1}] = V_{HT(c)}.$$

2) Estimateur par quotient:

$$\mathrm{E}[\hat{Y}_R \mid \hat{\bar{N}}] = \mu + \left(\frac{N_{..}}{\hat{N}_{..}}\right)[\mathbf{1}'\boldsymbol{R}(\hat{\bar{N}} - \boldsymbol{M}_2)]$$

$$= \mu + [\mathbf{1}'\boldsymbol{R}(\hat{\bar{N}} - \boldsymbol{M}_2)] + o(m^{-1})$$

$$\mathrm{var}[\hat{Y}_R \mid \hat{\bar{N}}] = (N_{..}/\hat{N}_{..})^2 V_{HT(c)}$$

$$= V_{HT(c)} + o(m^{-(3/2)}).$$

3) Estimateur de stratification a posteriori:

$$\mathrm{E}[\hat{Y}_{PS} \mid \hat{\bar{N}}] = \mu + [\boldsymbol{r}'\boldsymbol{P}(\hat{\bar{N}} - \boldsymbol{M}_2)]$$

$$= \mu + [\mathbf{1}'\boldsymbol{P}(\hat{\bar{N}} - \boldsymbol{M}_2)] + o(m^{-1})$$

$$\mathrm{var}[\hat{Y}_{PS} \mid \hat{\bar{N}}] = m^{-1}[\boldsymbol{r}'\boldsymbol{V}_c\boldsymbol{r}]$$

$$= V_{HT(c)} + o(m^{-(3/2)}).$$

4) Estimateur de régression linéaire:

$$\mathrm{E}[\hat{Y}_{LR} \mid \hat{\bar{N}}] = \mu$$

$$\mathrm{var}[\hat{Y}_{LR} \mid \hat{\bar{N}}] = V_{HT(c)}.$$

Comme il a été indiqué à la section 1, de légères modifications doivent être apportées aux formules ci-dessus pour les plans, comme l'échantillonnage aléatoire simple, dans lesquels $\mathbf{1}'\hat{N}_{..} = N_{..}$. La détermination des modifications requises est directe, et nous n'en fournissons pas les détails ici.

Les biais pour de grands échantillons des trois premiers estimateurs dépendent de $\hat{\bar{N}} - \boldsymbol{M}_2$. Autrement dit, les biais de ces estimateurs sont déterminés par l'exactitude avec laquelle l'échantillon permet d'estimer la distribution de la population parmi les strates a posteriori. Dans certains cas spéciaux, chacun des trois premiers estimateurs peut être conditionnellement non biaisé. L'estimateur de stratification a posteriori, par exemple, sera approximativement non biaisé si $\mathbf{1}'(\boldsymbol{H} - \boldsymbol{D}(\mu_k)) = \mathbf{0}'$. Cette situation existe dans l'échantillonnage aléatoire simple et est possible, bien que certainement pas de façon générale, dans des plans plus complexes. La matrice $\boldsymbol{H}$ peut être interprétée comme la pente d'une régression multidimensionnelle de $\hat{\bar{Y}}$ en $\hat{\bar{N}}$, ou de $\bar{Y}$ en $\bar{N}$ lorsque les estimations de l'échantillon sont voisines des valeurs de la population. Heuristiquement, en termes de superpopulation, si $\mathrm{E}_\xi(y_{ik}) = \mu_k N_{ik}$, comme dans Valliant (1993), et si $\mathrm{E}_\xi$ dénote une espérance à l'égard du modèle, alors $\mathrm{E}_\xi(Y_{.k}) = \mu_k N_{.k}$. La pente de la régression de $Y_{.k}$ en $N_{.k}$ est alors $\mu_k$ et, dans le cas habituel où les $\hat{Y}_{.k}'$ sont indépendants, $\boldsymbol{H}$ est diagonale. En fait, $\boldsymbol{H} = \boldsymbol{D}(\mu_k)$, de sorte que le biais conditionnel selon le plan de l'estimateur de stratification a posteriori serait nul. Si, par contre, le modèle a une ordonnée à l'origine, c.-à-d. si $\mathrm{E}_\xi(Y_{.k}) = \alpha_k + \mu_k N_{.k}$, l'estimateur de stratification a posteriori peut avoir un important biais conditionnel selon le plan. Nous raisonnerons de cette façon dans l'étude empirique de la section 3 afin de déterminer une population pour laquelle $\hat{Y}_{ps}$ est conditionnellement biaisé.

Un raisonnement semblable axé sur le modèle peut être appliqué aux estimateurs de Horvitz-Thompson et par quotient, afin de déterminer des populations pour lesquelles les biais conditionnels selon le plan seront vraisemblablement faibles pour de grands échantillons. Supposons, comme ci-dessus, que les $\hat{Y}_{.k}$ soient indépendants. Si chaque total de strate a posteriori n'est pas relié au nombre d'unités dans la strate a posteriori, c.-à-d. une situation particulière où $\mathrm{E}_\xi(Y_{.k})$ ne dépend pas de $N_{.k}$, alors $\hat{Y}_{HT}$ est conditionnellement non biaisé selon le plan. Si $\mathrm{E}_\xi(Y_{.k}) = \mu N_{.k}$, ce qui signifie que toutes les unités élémentaires de la population ont la même moyenne peu importe la strate a posteriori, $\hat{Y}_R$ est conditionnellement non biaisé selon le plan.

### 2.3 Espérances et variances non conditionnelles des estimateurs

En l'absence de condition, tous les estimateurs sont approximativement non biaisés selon le plan, comme il est indiqué ci-dessous. Les tailles relatives des variances dépendent des valeurs de $\Sigma_{12}$, $\Sigma_{22}$, $\mu$ et $\boldsymbol{D}(\mu_k)$. La situation est semblable à celle d'un échantillonnage aléatoire simple visant une variable d'intérêt $y$ et une variable auxiliaire $x$. Dans ce cas, la possibilité que l'estimateur par quotient, $\bar{y}_s\bar{X}/\bar{x}_s$, ou l'estimateur de régression, $\bar{y}_s + b(\bar{X} - \bar{x}_s)$, ait une moindre variance selon le plan dépend également des valeurs de certains paramètres de la population.

1) Estimateur de Horvitz-Thompson:

$$\mathrm{E}[\hat{Y}_{HT}] = \mu$$

$$\mathrm{var}[\hat{Y}_{HT}] = m^{-1}[\mathbf{1}'\Sigma_{11}\mathbf{1}].$$

2) Estimateur par quotient:

$$E[\hat{\bar{Y}}_R] = \mu + o(m^{-1})$$

$$\mathrm{var}[\hat{\bar{Y}}_R] = m^{-1}[1'[\Sigma_{11} - 2\mu\Sigma_{21} + \mu^2\Sigma_{22}]1] + o(m^{-(3/2)}).$$

3) Estimateur de stratification a posteriori:

$$E[\hat{\bar{Y}}_{PS}] = \mu + o(m^{-1})$$

$$\mathrm{var}[\hat{\bar{Y}}_{PS}] = m^{-1}[1'[\Sigma_{11} - 2D(\mu_k)\Sigma_{21} + D(\mu_k)\Sigma_{22}D(\mu_k)]1] + o(m^{-(3/2)}).$$

4) Estimateur de régression linéaire:

L'espérance et la variance non conditionnelles sont les mêmes que l'espérance et la variance conditionnelles.

## 3. RÉSULTATS DES SIMULATIONS

La théorie élaborée dans les sections précédentes a été vérifiée dans le cadre d'un ensemble d'études de simulation fondées sur trois populations distinctes. Les tailles des populations et les paramètres de base du plan de sondage des trois études sont présentés au tableau 1. La première population est formée d'un sous-ensemble des personnes incluses dans l'échantillon du premier trimestre de la National Health Interview Survey (NHIS) de 1985, tandis que la deuxième population est constituée d'un sous-ensemble de personnes incluses dans l'échantillon de septembre 1988 de la Current Population Survey (CPS). La NHIS et la CPS sont deux enquêtes par sondage menées par le gouvernement des États-Unis. La variable d'intérêt pour la population de la NHIS est le nombre de jours d'activité restreinte au cours des deux semaines ayant précédé l'interview, tandis que la variable d'intérêt pour la population de la CPS est le salaire hebdomadaire par personne.

**Tableau 1**

Tailles des populations et paramètres de base du plan de sondage pour trois études de simulation

| Population | Taille de la population $N$ | Nombre d'UPD $M$ | Nombre d'UPD dans l'échantillon $m$ |
|---|---|---|---|
| HIS | 2,934 | 1,100 | 115 |
| CPS | 10,841 | 2,826 | 200 |
| Artificielle | 22,001 | 2,000 | 200 |

Des strates a posteriori, parmi les populations de la NHIS et de la CPS, ont été formées d'après des caractéristiques démographiques (comme c'est généralement le cas dans les enquêtes auprès des ménages), de façon à créer des sous-groupes de la population qui étaient homogènes à l'égard de la variable d'intérêt. Pour la population de la NHIS, les variables âge et sexe ont servi à définir 4 strates a posteriori, et pour la population de la CPS, les variables âge, race et sexe ont servi à définir 8 strates a posteriori.

La troisième population est artificielle; elle a été créée dans l'intention d'introduire un biais conditionnel important dans l'estimateur de stratification a posteriori de la moyenne. Comme il a été indiqué à la section 2.2, $\hat{\bar{Y}}_{PS}$ sera conditionnellement biaisé si les totaux des strates a posteriori d'UPD pour la variable d'intérêt, la condition donnée étant le nombre d'unités dans chaque UPD/strate a posteriori, se conforment à un modèle ayant une ordonnée à l'origine non nulle. Avec ce résultat en tête, nous avons généré la population de telle façon que

$$E(y_{ik} \mid N_{ik}) = \alpha_k + \beta N_{ik} + \gamma N_{ik}^2, \qquad (4)$$

où $N_{ik}$ est le nombre d'unités dans la $k$-ième strate a posteriori pour la $i$-ième UPD et $\alpha_k$, $\beta$ et $\gamma$ sont des constantes. Plus précisément, cinq strates a posteriori ont été utilisées, avec $\alpha_k = 100k$ ($k = 1, \ldots, 5$), $\beta = 10$ et $\gamma = -.05$. En tout, deux mille UPD ont été générées, et le nombre total d'unités dans la $i$-ième UPD, disons $N_{i.}$, était une variable aléatoire de Poisson de moyenne 10. Ensuite, pour $N_{i.}$, donné, les nombres d'unités dans les cinq strates a posteriori (c.-à-d. $N_{i1}, N_{i2}, \ldots, N_{i5}$) pour la $i$-ième UPD ont été déterminés à l'aide d'une distribution multinomiale de paramètres $N_{i.}$ et $p_k = .20$ pour $k = 1, 2, \ldots 5$.

Pour les UPD où $N_{ik} \geq 1$, la valeur de la variable d'intérêt pour la $j$-ième unité de la $k$-ième strate a posteriori dans la $i$-ième UPD était une réalisation de la variable aléatoire

$$y_{ijk} = \alpha_k/N_{ik} + \beta + \gamma N_{ik} + \epsilon_{1i} + \epsilon_{2ik} + \epsilon_{3ijk}N_{i.}$$

$$(j = 1, \ldots, N_{ik}; N_{ik} \geq 1),$$

où $\epsilon_{1i}$, $\epsilon_{2ik}$ et $\epsilon_{3ijk}$ sont trois variables aléatoires chi carré (à 6 degrés de liberté) réduites indépendantes. Il découle de cette structure que $E(y_{ik} \mid N_{ik})$ est donnée par (4). De plus, les valeurs de la variable d'intérêt pour les unités faisant partie d'une UPD sont corrélées, et la corrélation varie selon que les unités appartiennent ou non à la même strate a posteriori. Ce même algorithme a été utilisé pour chacune des 100 strates du plan. Vingt UPD ont été générées dans chaque strate du plan, ce qui a donné un total de 2,000 UPD.

Un plan stratifié à un seul degré a été utilisé pour la population de la NHIS, et les "ménages" ont été pris comme UPD. Dix strates du plan ont été utilisées et l'on a prélevé dans chaque strate, sans remplacement, un échantillon aléatoire simple d'environ 10% des ménages. Chaque échantillon comprenait 115 ménages et chaque ménage de l'échantillon était entièrement recensé. Un total de 5,000 échantillons de ce genre ont été prélevés pour les besoins de l'étude de simulation.

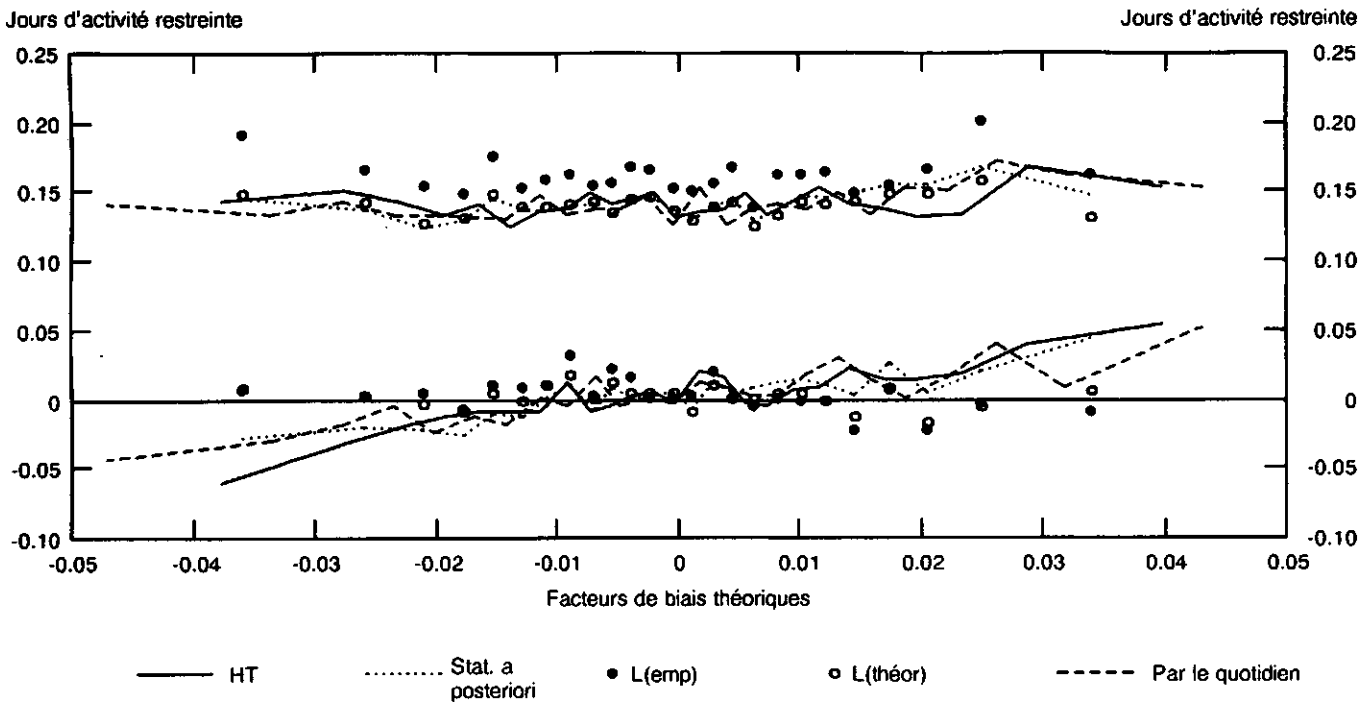


**Figure 1.** Simulation pour la HIS, $m = 115$

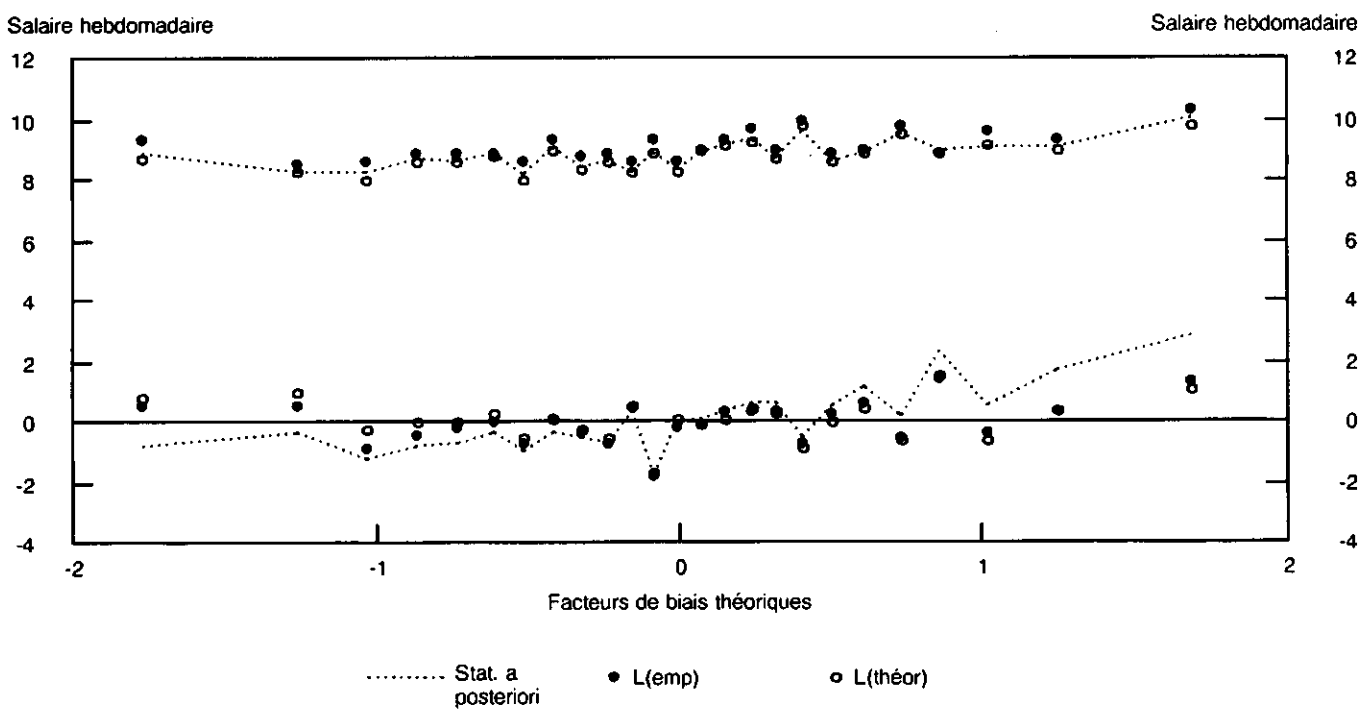


**Figure 2.** Simulation pour la CPS, $m = 200$

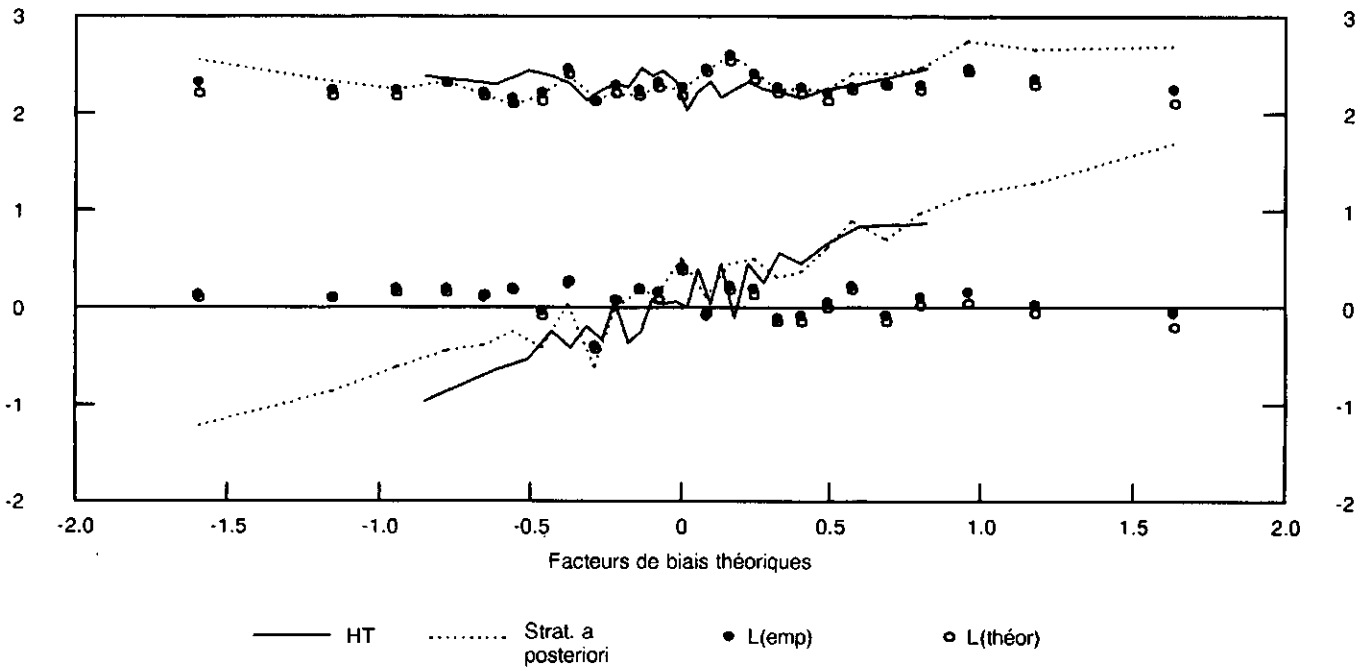


**Figure 3.** Simulation pour la population artificielle, $m = 200$

Tant pour la population de la CPS que pour la population artificielle, un plan d'échantillonnage stratifié à deux degrés a été utilisé. Dans le cas de la population de la CPS, des segments géographiques, employés dans l'enquête originale et composés d'environ quatre ménages voisins, ont été utilisés comme UPD, tandis que les personnes étaient les unités du deuxième degré. Dans les deux populations, 100 strates du plan ont été créées, chaque strate ayant environ le même nombre d'UPD, et un échantillon de $m = 2$ UPD a été prélevé avec probabilité proportionnelle à la taille dans chaque strate au moyen de la méthode d'échantillonnage systématique décrite par Hansen, Hurwitz et Madow (1953, p. 343). Ainsi, 200 UPD ont été sélectionnées pour les deux populations. La sélection du deuxième degré a également été semblable pour les deux populations. Dans le cas de la population de la CPS, un échantillon aléatoire simple de 4 personnes a été prélevé sans remplacement dans chaque UPD de l'échantillon pour laquelle $N_{i.} > 4$, et toutes les personnes ont été sélectionnées dans chaque UPD de l'échantillon pour laquelle $N_{i.} \leq 4$. Dans le cas de la population artificielle, la taille de l'échantillon à l'intérieur des UPD a été fixée à 15 au lieu de 4, ce qui a entraîné le recensement complet de la plupart des UPD de l'échantillon. Un total de 5,000 échantillons ont été prélevés dans chacune des populations pour les besoins de l'étude de simulation.

Pour chaque échantillon, nous avons calculé $\hat{\bar{Y}}_{HT}$, $\hat{\bar{Y}}_R$, $\hat{\bar{Y}}_{PS}$ et deux versions de $\hat{\bar{Y}}_{LR}$. Pour la première version de l'estimateur de régression, dénotée $\hat{\bar{Y}}_{LR}$(emp) dans les tableaux, $\boldsymbol{H}$ a été estimée séparément à partir de chaque échantillon, comme cela s'imposerait en pratique. Chaque composante de $\Sigma_{12}$ et de $\Sigma_{22}$ a été évaluée au moyen de l'estimateur de covariance de la grappe extrême, dans une forme convenant au plan, selon la définition de Hansen et coll. (1953, p. 419). La deuxième version, dénotée $\hat{\bar{Y}}_{LR}$(théor), utilisait la même valeur de $\boldsymbol{H}$ dans chaque échantillon, laquelle était une estimation davantage voisine de la valeur théorique de la matrice $\boldsymbol{H}$. Dans le cas de la population de la CPS et de la population artificielle, la matrice $\boldsymbol{H}$ théorique a été estimée d'après les covariances empiriques obtenues à partir de traitements de simulation distincts de 5,000 échantillons. Dans le cas de la population de la NHIS, le plan était suffisamment simple pour permettre le calcul théorique direct de $\boldsymbol{H}$. Quand l'échantillon d'UPD devient grand, la performance de $\hat{\bar{Y}}_{LR}$(emp) devrait s'approcher de celle de $\hat{\bar{Y}}_{LR}$(théor). La performance de $\hat{\bar{Y}}_{LR}$(théor) est, par conséquent, représentative du mieux qu'on puisse espérer de la version empirique de l'estimateur de régression pour une taille d'échantillon donnée.

Le Tableau 2 présente les résultats de l'analyse non conditionnelle pour l'ensemble des 5,000 échantillons de chaque population. Les erreurs quadratiques moyennes (rmse) empiriques ont été calculées selon l'expression $\text{rmse}(\hat{\bar{Y}}) = [\sum_{s=1}^{S}(\hat{\bar{Y}}_s - \bar{Y})^2/S]^{1/2}$, où $S = 5{,}000$ et $\hat{\bar{Y}}_s$ est l'une des estimations de la moyenne de la population,

venant de l'échantillon $s$. Dans le cas de la population de la CPS et de la population artificielle, les résultats donnés par l'estimateur de Horvitz-Thompson et l'estimateur par quotient étaient presque identiques, de sorte que nous présentons seulement les premiers. Sur l'ensemble des échantillons, le biais de chacun des estimateurs était négligeable. Comme la théorie le laissait prévoir, $\hat{\bar{Y}}_{LR}$(théor) était le plus précis des choix en présence, bien que le gain le plus élevé comparativement à $\hat{\bar{Y}}_{PS}$ n'ait été que de 4.7%, pour la population artificielle. La nécessité d'estimer $H$ déstabilise l'estimateur de régression, comme le montrent les résultats relatifs à $\hat{\bar{Y}}_{LR}$(emp). Pour les populations de la NHIS et de la CPS, $\hat{\bar{Y}}_{LR}$(emp) a une erreur quadratique moyenne supérieure à la fois à celle de $\hat{\bar{Y}}_{LR}$(théor) et à celle de $\hat{\bar{Y}}_{PS}$. La perte la plus prononcée concerne la population de la NHIS, où l'erreur quadratique moyenne de $\hat{\bar{Y}}_{LR}$(emp) est d'environ 15% supérieure aussi bien à celle de $\hat{\bar{Y}}_{LR}$(théor) qu'à celle de $\hat{\bar{Y}}_{PS}$. Ce résultat était prévisible en raison de la taille plus faible des échantillons d'UPD de la population de la NHIS, et donc de la moins grande stabilité de l'estimation de $H$ pour cette population.

**Tableau 2**

Résultats des simulations pour trois populations.
5,000 échantillons ont été prélevés dans chaque population

| Estimateur | Biais rel. $\hat{\bar{Y}}$ (%) | rmse($\hat{\bar{Y}}$) | $100*\left[\frac{\text{rmse}(\hat{\bar{Y}})}{\text{rmse}(\hat{\bar{Y}}_{PS})} - 1\right]$ |
|---|---|---|---|
| Population de la HIS | | | |
| $\hat{\bar{Y}}_{HT}$ | .12 | .141 | .05 |
| $\hat{\bar{Y}}_{R}$ | .10 | .141 | .02 |
| $\hat{\bar{Y}}_{PS}$ | .11 | .141 | 0 |
| $\hat{\bar{Y}}_{LR}$(emp) | .19 | .162 | 14.71 |
| $\hat{\bar{Y}}_{LR}$(théor) | .08 | .140 | −.96 |
| Population de la CPS | | | |
| $\hat{\bar{Y}}_{HT}$ | −.01 | 10.25 | 15.8 |
| $\hat{\bar{Y}}_{PS}$ | 0 | 8.85 | 0 |
| $\hat{\bar{Y}}_{LR}$(emp) | −.03 | 9.11 | 3.0 |
| $\hat{\bar{Y}}_{LR}$(théor) | −.01 | 8.79 | −.6 |
| Population artificielle | | | |
| $\hat{\bar{Y}}_{HT}$ | .02 | 2.30 | −2.93 |
| $\hat{\bar{Y}}_{PS}$ | .12 | 2.37 | 0 |
| $\hat{\bar{Y}}_{LR}$(emp) | .04 | 2.31 | −2.41 |
| $\hat{\bar{Y}}_{LR}$(théor) | .02 | 2.26 | −4.70 |

Les figures 1 à 3 présentent les résultats des simulations de l'analyse conditionnelle. Les 5,000 échantillons ont été triés selon les facteurs de biais théoriques présentés à la section 2.2. Le tri a été fait séparément pour chacun des estimateurs de la moyenne de la population. Dans le cas des deux estimateurs de régression, qui sont théoriquement non biaisés pour de grands échantillons, le facteur de biais pour $\hat{\bar{Y}}_{PS}$ a servi de base de tri. Les échantillons triés ont alors été répartis en 25 groupes de 200 échantillons chacun, et les biais et les erreurs quadratiques moyennes empiriques ont été calculés pour chaque groupe. On a ensuite tracé, comme le montrent les figures, le graphique des résultats des groupes par rapport aux facteurs de biais théoriques. Les ensembles de points supérieurs de chaque figure sont les erreurs quadratiques moyennes empiriques des groupes, tandis que les ensembles inférieurs sont les biais empiriques. Les deux estimateurs de régression sont, comme prévu, conditionnellement non biaisés. Les autres estimateurs, toutefois, affichent des biais conditionnels importants, lesquels, dans les ensembles d'échantillons les plus extrêmes, constituent une part importante des erreurs quadratiques moyennes. Pour la population de la CPS, l'intervalle des facteurs de biais pour $\hat{\bar{Y}}_{HT}$ est tellement plus vaste (−10 à 10) que celui des autres estimateurs que nous avons omis $\hat{\bar{Y}}_{HT}$ dans le graphique à des fins de clarté. Dans le voisinage du point d'équilibre, $\hat{\bar{N}} = \bar{N}$, tous les estimateurs ont une performance à peu près identique, mais en raison d'un manque de données au stade de la préparation du plan, nous ne pouvons agir sur la mesure dans laquelle un échantillon peut être proche du point d'équilibre. Le choix le plus sûr pour le contrôle du biais conditionnel est donc $\hat{\bar{Y}}_{LR}$(emp). Ce résultat est semblable à celui de Valliant (1990), qui a fait observer que dans l'échantillonnage stratifié à un seul degré, qu'il soit aléatoire ou systématique, l'estimateur de régression linéaire distinct est un bon choix pour le contrôle du biais, lorsqu'est donnée la moyenne de l'échantillon d'une variable auxiliaire.

## 4. BASES DÉFECTUEUSES

Les biais conditionnels examinés dans les sections précédentes étaient de nature technique et mathématique. Un problème pratique important, présent dans de nombreuses enquêtes et pouvant aussi être source de biais, est causé par la piètre couverture de la population cible. Nous nous penchons sur ce problème dans la présente section.

### 4.1 Le problème fondamental des bases défectueuses

Dans la plupart des applications réelles, les unités élémentaires de la population ne sont pas toutes incluses dans la base de sondage. Il n'est pas rare que dans les enquêtes auprès des ménages, certains sous-groupes démographiques, notamment les minorités, soient couverts de façon incomplète par la base de sondage. Bailar (1989), par exemple, signale qu'en 1985, l'estimation tirée de l'échantillon de la CPS du nombre total de Noirs de sexe masculin âgés de 22 à 24 ans ne correspondait qu'à 73% d'une estimation indépendante de la population totale de ce groupe. Les proportions correspondantes pour les Noirs de sexe masculin des groupes d'âge 25-29 et 60-61 étaient de 80% et de 76%.

Pour exprimer ce genre de problème de couverture en termes formels, supposons que $N_{.k}$ désigne maintenant le nombre d'unités élémentaires de la base de sondage et que $\dot{N}_{.k}$ soit le nombre *réel* d'unités de la population dans la $k$-ième strate a posteriori. Dans l'exposé qui suit, les termes surmontés d'un point sont des valeurs pour la population, tandis que les termes non surmontés d'un point sont des valeurs pour la base de sondage. Si $\dot{Y}_{.k}$ est la somme des valeurs $y$ sur l'ensemble des unités de la population de la $k$-ième strate a posteriori, il s'ensuit que la *moyenne vraie de la population* est donnée par:

$$\dot{\mu} = \lim_{L\to\infty} \frac{\sum_{k=1}^{K} \dot{Y}_{.k}}{\sum_{k=1}^{K} \dot{N}_{.k}} = \lim_{L\to\infty} \sum_{k=1}^{K} \frac{\dot{N}_{.k}}{\dot{N}_{..}} \frac{\dot{Y}_{.k}}{\dot{N}_{.k}} = \sum_{k=1}^{K} \dot{\phi}_k \dot{\mu}_k .$$

De toute évidence, les quatre estimateurs de la moyenne présentés à la section 2 sont tous biaisés (tant conditionnellement que non conditionnellement) pour $\dot{\mu}$, le terme de biais additionnel étant donné par $\mu - \dot{\mu}$ pour chacun des estimateurs. Il est à remarquer que ce terme de biais est $o(1)$, de sorte qu'il dominera les autres termes de biais indiqués à la section 2.2 à mesure qu'augmentera le nombre d'UPD. Autre problème encore plus fondamental: dans la plupart des cas, les valeurs individuelles $N_{.k}$ de la base sont inconnues, de sorte que l'estimateur par quotient est le seul qui soit bien défini. Par exemple, l'estimateur de Horvitz-Thompson de la moyenne, selon la définition de la section 2, exige la connaissance de $N_{..}$, soit le nombre total d'unités dans la base, mais $N_{..}$ peut être inconnu. En revanche, les $\dot{N}_{.k}$ (ou au moins les proportions $\dot{\phi}_k$) peuvent être connus grâce à des sources indépendantes et donc être disponibles pour la construction de l'estimateur. Dans les enquêtes auprès des ménages, par exemple, les $\dot{N}_{.k}$ peuvent être tirés de projections intercensitaires des tailles de population.

Avant de tenter de construire des estimateurs non biaisés pour $\dot{\mu}$, il importe de noter que

$$\mu - \dot{\mu} = \sum_{k=1}^{K} (\phi_k - \dot{\phi}_k)(\mu_k - \dot{\mu}_k)$$

$$+ \sum_{k=1}^{K} (\phi_k - \dot{\phi}_k)\dot{\mu}_k + \sum_{k=1}^{K} \dot{\phi}_k(\mu_k - \dot{\mu}_k).$$

Par conséquent, si nous supposons que pour chaque strate a posteriori la moyenne des unités de la base est égale à la moyenne vraie de la population (c.-à-d. $\mu_k = \dot{\mu}_k$ pour chaque $k$), le terme de biais se réduit à:

$$\mu - \dot{\mu} = \sum_{k=1}^{K} (\phi_k - \dot{\phi}_k)\mu_k = \sum_{k=1}^{K} (\phi_k - \dot{\phi}_k)\dot{\mu}_k .$$

Il s'agit d'une hypothèse très contraignante (et aussi très "commode"); toutefois, il est pratiquement impossible de s'attaquer au problème des bases de sondage défectueuses sans recourir à une telle hypothèse.

### 4.2 Estimateurs de remplacement

La stratégie de base consiste à élaborer un estimateur pour le biais de la base défectueuse, $\mu - \dot{\mu}$, puis à soustraire cet estimateur des estimateurs étudiés plus tôt. Deux cas doivent être examinés:

Cas 1. Les paramètres de la base $\{\phi_k, 1 \le k \le K\}$ sont inconnus, et

Cas 2. Les paramètres de la base $\{\phi_k, 1 \le k \le K\}$ sont connus.

**Cas 1.** Dans une telle situation, seul l'estimateur par quotient est bien défini et le seul candidat évident à titre d'estimateur du biais est

$$\hat{B}_1 = \sum_{k=1}^{K} \left(\frac{\hat{N}_{.k}}{\hat{N}_{..}} - \dot{\phi}_k\right) \frac{\hat{Y}_{.k}}{\hat{N}_{.k}} = \hat{Y}_R - \sum_{k=1}^{K} \dot{\phi}_k \frac{\hat{Y}_{.k}}{\hat{N}_{.k}}.$$

Selon la stratégie énoncée ci-dessus, l'estimateur résultant pour $\dot{\mu}$ est:

$$\hat{\dot{Y}}_1 = \hat{Y}_R - \hat{B}_1 = \sum_{k=1}^{K} \dot{\phi}_k \frac{\hat{Y}_{.k}}{\hat{N}_{.k}}.$$

Il s'agit de l'estimateur "de stratification a posteriori" qu'on trouve habituellement en pratique. On peut vérifier immédiatement les propriétés suivantes de $\hat{\dot{Y}}_1$:

$$\mathrm{E}[\hat{\dot{Y}}_1 \mid \hat{\dot{N}}] = \dot{\mu} + [p'P(\hat{\dot{N}} - M_1)] + o(m^{-1})$$

où

$$p' = \left[\frac{\dot{\phi}_1}{\phi_1}, \frac{\dot{\phi}_2}{\phi_2}, \ldots, \frac{\dot{\phi}_K}{\phi_K}\right]$$

$$\mathrm{var}[\hat{\dot{Y}}_1 \mid \hat{\dot{N}}] = m^{-1}[p'V_c p] + o(m^{-(3/2)})$$

$$\mathrm{E}[\hat{\dot{Y}}_1] = \dot{\mu} + o(m^{-1})$$

$$\mathrm{var}[\hat{\dot{Y}}_1] = m^{-1}[p'[\Sigma_{11} - 2D(\mu_k)\Sigma_{21}$$

$$+ D(\mu_k)\Sigma_{22}D(\mu_k)]p] + o(m^{-(3/2)}).$$

La tentative de compenser pour le biais de la base défectueuse porte fruit, en ce sens que $\hat{\dot{Y}}_1$ est non biaisé, de façon non conditionnelle, pour $\dot{\mu}$. Toutefois, le biais conditionnel est toujours présent.

**Cas 2.** Dans cette situation, on peut vérifier que l'estimateur

$$\hat{B}_2 = (1 - p)' \left[ \hat{\bar{Y}} - \Sigma_{12} \Sigma_{22}^{-1} \left( \frac{\hat{N}}{\hat{N}_{..}} - M_2 \right) \right],$$

est, de façon approximative, conditionnellement non biaisé pour $\mu - \dot{\mu}$ et que, comme $\hat{Y}_{LR}$ est conditionnellement non biaisé pour $\mu$, il s'ensuit directement que l'estimateur

$$\hat{Y}_2 = \hat{Y}_{LR} - \hat{B}_2 = p' \left[ \hat{\bar{Y}} - \Sigma_{12} \Sigma_{22}^{-1} \left( \frac{\hat{N}}{\hat{N}_{..}} - M_2 \right) \right].$$

est approximativement non biaisé, de façon tant conditionnelle que non conditionnelle, pour $\dot{\mu}$. On peut également vérifier que:

$$\mathrm{var}[\hat{Y}_2 \mid \hat{\bar{N}}] = \mathrm{var}[\hat{Y}_2] = m^{-1}[p' V_c p].$$

Outre les problèmes de l'estimateur de régression linéaire mentionnés plus haut, notons que cet estimateur, en général, n'est même pas bien défini, car les situations où les paramètres de la base $\{\phi_k, 1 \leq k \leq K\}$ sont connus quand la base est défectueuse sont rares, voire inexistantes.

## 5. CONCLUSION

La présente étude a généralisé les techniques asymptotiques suggérées par Robinson (1987) pour étudier le problème de la stratification *a posteriori* selon une approche d'inférence conditionnelle, fondée sur le plan. Un article important sur l'étude conditionnelle de la stratification *a posteriori* a été publié par Holt et Smith (1979); dans cet article, l'une des hypothèses de base était que $\hat{Y}_{PS}$ est conditionnellement non biaisé. Ce sera vrai (du moins asymptotiquement) seulement si $\mathbf{1}'(H - D(\mu_k)) = \mathbf{0}'$; par conséquent, de façon générale, cette hypothèse est fausse. En fait, l'échantillonnage aléatoire simple d'unités élémentaires pourrait être l'un des rares cas où, en situation réelle, cette hypothèse de base est vérifiée.

Selon notre analyse conditionnelle, l'estimateur de régression linéaire est le meilleur choix parmi les quatre étudiés. C'est le seul qui soit conditionnellement non biaisé. L'estimateur de stratification *a posteriori* n'est pas meilleur (ni pire) que l'estimateur de Horvitz-Thompson ou l'estimateur par quotient; tous ont des termes de biais conditionnel d'ordre $m^{-(1/2)}$. Tous les estimateurs ont la même variance conditionnelle jusqu'aux termes d'ordre $m^{-1}$; en outre, la variance conditionnelle ne dépend *pas* de $\hat{\bar{N}}$, le vecteur des proportions estimées dans les strates *a posteriori*. Par conséquent, du fait qu'il est conditionnellement non biaisé, l'estimateur de régression présente la plus faible erreur quadratique moyenne conditionnelle.

Les estimateurs de Horvitz-Thompson, par quotient et de stratification *a posteriori* sont non biaisés de façon non conditionnelle. Bien que cela soit quelque peu illogique, on pourrait tenter de défendre ces estimateurs en comparant leurs propriétés non conditionnelles aux propriétés conditionnelles de l'estimateur de régression linéaire. Mais même dans cette perspective mixte, l'estimateur $\hat{Y}_{LR}$(théor) est nettement supérieur aux autres. Non seulement il est conditionnellement non biaisé, mais la variance conditionnelle de l'estimateur de régression linéaire ne peut être supérieure à la variance non conditionnelle de n'importe quel des autres estimateurs. Pour les grands échantillons d'UPD, la version empirique de l'estimateur de régression héritera de ces propriétés intéressantes de $\hat{Y}_{LR}$(théor) et affichera elle aussi une bonne performance.

Le problème posé par une base de sondage défectueuse amène des complications qui n'existeraient pas autrement. Chacun des estimateurs de la moyenne étudiés ici est biaisé, tant de façon conditionnelle que de façon non conditionnelle. Des rajustements tenant compte de ce biais ne sont possibles qu'en vertu de l'hypothèse contraignante selon laquelle la moyenne des unités de chaque strate *a posteriori* est la même pour toutes les unités de la population, que celles-ci soient incluses dans la base ou qu'elles en soient exclues.

La question de l'estimation de la variance est un aspect que nous n'avons pas examiné. Un estimateur de la variance fondé sur le plan pour l'estimateur de régression peut être obtenu à l'aide des méthodes de Särndal, Swensson et Wretman (1989).

## REMERCIEMENTS

Les opinions exprimées sont celles des auteurs et n'engagent en rien le U.S. Bureau of Labor Statistics. Les auteurs tiennent également à remercier le rédacteur associé et l'arbitre pour leurs commentaires constructifs; nous croyons que leurs observations ont constitué un apport important à notre communication.

## BIBLIOGRAPHIE

BAILAR, B. (1989). Information needs, surveys, and measurement errors. Dans *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton, et M.P. Singh). New York: Wiley.

DURBIN, J. (1969). Inferential aspects of randomness of sample size in survey sampling. Dans *New Developments in Survey Sampling*, (Éds. N.L. Johnson et H. Smith). New York: Wiley.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society* B, 31, 195-233.

FULLER, W.A. (1981). Comment on an empirical study of the ratio estimator and estimators of its variance by R.M. Royall and W.G. Cumberland. *Journal of the American Statistical Association*, 76, 78-80.

HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory*, Vol. 1. New York: John Wiley and Sons.

HANSEN, M.H., MADOW, W.G., et TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-796.

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

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

KIEFER, J. (1977). Conditional confidence statements and confidence estimators (avec discussion). *Journal of the American Statistical Association*, 72, 789-827.

KREWSKI, D., et RAO, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

LITTLE, R.J.A. (1991). Post-Stratification: A modeler's perspective. *Proceeding of the Section on Survey Research Methods, American Statistical Association*, à paraître.

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

RAO, J.N.K. (1992). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. Présenté au Workshop on Uses of Auxiliary Information in Surveys, Statistics Sweden.

RAO, J.N.K., et WU, C.F.J. (1985). Inference from stratified samples: Second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

ROBINSON, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82, 826-831.

ROYALL, R.M. (1971). Linear regression models in finite population sampling theory. Dans *Foundations of Statistical Inference*, (Éds. V.P. Godambe et D.A. Sprott). Toronto: Holt, Rinehart et Winston.

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

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

VALLIANT, R. (1990). Comparisons of variance estimators in stratified random and systematic sampling. *Journal of Official Statistics*, 6, 115-131.

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

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3ième Éd.). London: Griffin.

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 205–209
Statistique Canada



# Échantillonnage dans des bases imparfaites contenant un nombre inconnu d'enregistrements répétés

SHIBDAS BANDYOPADHYAY et A.K. ADHIKARI[1]

RÉSUMÉ

Dans cette étude, nous nous intéressons à des bases de sondage imparfaites desquelles on n'a retiré aucune unité de population mais dans lesquelles un nombre indéterminé d'unités peuvent avoir été ajoutées un nombre indéterminé de fois sous des identités différentes. Lorsqu'on ne pose pas l'hypothèse de l'existence d'information supplémentaire concernant des unités de la base imparfaite, il est établi qu'en ce qui a trait à l'estimation d'un ratio ou d'une moyenne de population, l'erreur quadratique moyenne des estimateurs fondés sur la base imparfaite est inférieure à celle des estimateurs fondés sur la base parfaite pour l'échantillonnage aléatoire simple, lorsque les fractions de sondage des deux bases sont les mêmes. Cependant, cette relation n'est pas toujours vraie en ce qui concerne l'estimation d'un total de population. Il peut aussi arriver que l'estimateur d'un ratio, d'une moyenne ou d'un total ait une erreur quadratique moyenne moins élevée même si la fraction de sondage est plus faible dans la base imparfaite que celle utiliser dans la base parfaite.

MOTS CLÉS: Base de sondage imparfaite; efficience.

## 1. INTRODUCTION

L'absence de bases de sondage complètes est un problème fréquent dans la planification d'enquêtes. L'Institut international de statistique a reconnu l'importance de s'intéresser au problème de l'échantillonnage dans des bases imparfaites et a, de fait, inscrit cette question au programme de sa 34$^{\text{ième}}$ session, tenue à Ottawa (Canada). À cette occasion, Hansen et coll. (1963) et Szameitat et Schaffer (1963) avaient présenté des communications. Nous renvoyons aussi le lecteur aux ouvrages de Singh (1977, 1983). Wright et Tsao (1983) ont rédigé une bibliographie sur les bases de sondage afin de mettre en lumière les problèmes qui se posent lorsqu'on fait des sondages à l'aide de bases imparfaites.

Ces dernières années, l'Indian Statistical Institute a réalisé deux enquêtes indépendantes pour évaluer l'effet de programmes gouvernementaux sur la condition économique de la communauté de pêcheurs du Bengale-Occidental, en Inde. Dans la première enquête (1988), les ménages étaient choisis à même les listes de membres des sociétés coopératives de pêcheurs (SCP). Dans la seconde enquête, plus récente, on s'est servi de la liste des pêcheurs qui recevaient de l'aide de la Fish Farmer's Development Agency (FFDA). On savait que les membres des SCP ou les bénéficiaires de la FFDA n'appartenaient pas tous à des ménages différents, mais il n'était pas possible de déterminer quels membres ou quels bénéficiaires appartenaient au même ménage sans communiquer avec les ménages. Par conséquent, lorsqu'on s'est servi des listes de membres des SCP ou des listes de bénéficiaires de la FFDA pour le choix des ménages, les bases de sondage contenaient un nombre inconnu de doubles enregistrements. Comme les données sur le ménage étaient recueillies au moyen d'une interview sur place, on pouvait détecter les doubles enregistrements uniquement parmi les ménages échantillonnés. On a divisé les valeurs des variables observées pour les ménages de l'échantillon par le nombre correspondant d'enregistrements répétés présents dans la base de sondage tandis qu'on a conservé dans l'échantillon, sous leurs diverses identités, les ménages qui se répétaient.

L'exemple de bases imparfaites que nous étudions ici est un cas particulier de Rao (1968). Un des arbitres a souligné que la situation que nous décrivons dans cet article existe à Statistique Canada pour certaines bases utilisées dans les enquêtes-entreprises.

Le genre de bases imparfaites que nous étudions ici sont celles desquelles on n'a retiré aucune unité de population mais dans lesquelles un nombre indéterminé d'unités peuvent avoir été ajoutées un nombre indéterminé de fois sous des identités différentes. On suppose qu'il est possible d'évaluer, à l'étape de la collecte des données, le nombre d'enregistrements répétés présents dans la base de sondage pour chaque unité échantillonnée. On n'exclut pas la possibilité de tirer plusieurs enregistrements de la même unité de population. On ne suppose pas l'existence d'information supplémentaire concernant les unités de la base imparfaite et on s'en tient uniquement aux plans d'échantillonnage aléatoire simple sans remise (ÉASSR).

Comme les bases imparfaites que nous étudierons ici ne nous permettront pas de connaître le nombre total d'unités de population, l'estimation de la moyenne et l'estimation du total pour un caractère de la population ne posent pas les mêmes problèmes.

La question fondamentale à laquelle nous allons tenter de répondre dans cet article est la suivante. Quelle est la meilleure voie à suivre: mettre à jour la base de sondage imparfaite et tirer un échantillon, ou utiliser la base imparfaite telle quelle?

[1] Shibdas Bandyopadhyay et A.K. Adhikari, Indian Statistical Institute, Calcutta, Inde 700 035.

Dans les deux enquêtes sur les ménages de pêcheurs, on pensait que la plupart des variables économiques étudiées étaient étroitement liées au nombre de membres de SCP ou de bénéficiaires de la FFDA au sein d'un ménage, en ce sens que la variabilité de ces caractéristiques par membre de SCP ou par bénéficiaire de la FFDA était inférieure à la variabilité des mêmes caractéristiques par ménage. On estimait donc que l'utilisation d'une base imparfaite était acceptable dans ces conditions.

Nous verrons que dans des cas comme ci-dessus, l'estimateur d'un ratio, d'une moyenne ou d'un total peut avoir une erreur quadratique moyenne moins élevée même si la fraction de sondage est plus faible dans la base imparfaite que celle utiliser dans la base parfaite.

Même si la variabilité n'a pas de rapport avec le nombre d'enregistrements répétés, nous verrons que pour l'estimation d'un ratio ou d'une moyenne, la base imparfaite doit être préférée à la base parfaite, du point de vue de l'EQM, lorsque les fractions de sondage des deux bases sont les mêmes.

## 2. NOTATION ET RELATIONS

Considérons une population finie constituée de $N$ unités $U_1, U_2, \ldots, U_N$. Désignons par $U_1^*, U_2^*, \ldots, U_M^*$ les unités listées dans une base imparfaite. Pour $k = 1, 2, \ldots, r$, désignons par $A_k$ une sous-population des $N$ unités formée de $N_k$ unités distinctes. Chacune des unités de $A_k$ est listée exactement $k$ fois dans la base imparfaite sous des identités différentes. Supposons que

a) chaque $U_i$ appartient à une sous-population $A_k$, pour une valeur $k$ quelconque (c'est-à-dire que chaque $U_i$ est incluse dans la base imparfaite au moins une fois) et que

b) si $U_j^*$ est échantillonnée à l'aide de la base imparfaite, il sera possible de déterminer, à l'étape de la collecte de données, l'unité $U_i$ correspondante et la valeur de $k$ (c.-à-d. le nombre de fois que $U_i$ figure dans la base incomplète sous différentes identités, l'une d'elles correspondant à l'unité échantillonnée $U_j^*$) pour laquelle $U_i$ appartient à $A_k$.

Les relations suivantes sont valides:

$$N_1 + N_2 + \ldots + N_r = N;$$

$$N_k \geq 0, k = 1, 2, \ldots, r,$$

$$N_1 + 2N_2 + \ldots + rN_r = M,$$

où $r, N_1, N_2, \ldots, N_r$ et $N$ sont tous inconnus, seul $M$ étant connu $M \geq N$; $M$ peut s'écrire, pour une valeur $\alpha$ inconnue,

$$M = N(1 + \alpha), \quad \alpha \geq 0. \tag{2.1}$$

Désignons les valeurs $X$ et $Y$ pour l'unité $U_i$ par $X_i$ et $Y_i$ respectivement, $(i = 1, 2, \ldots, N)$. Puisque chaque $U_j^*$, $(j = 1, 2, \ldots, M)$, peut être identifiée à une $U_i$, pour un $i$ quelconque, $(i = 1, 2, \ldots, N)$ et puisque $U_i$ appartient à $A_k$, pour une valeur $k$ quelconque, $(k = 1, 2, \ldots, r)$, définissons les valeurs $X$, $Y$ et $C$ pour l'unité $U_j^*$ par les équations suivantes:

$$X_j^* = X_i/k, \quad Y_j^* = Y_i/k, \quad C_j^* = 1/k.$$

En raison des hypothèses a) et b), les valeurs $X^*$, $Y^*$ et $C^*$ sont observables pour les unités échantillonnées qui proviennent de la base imparfaite.

Les identités suivantes mettent en relation les paramètres tirées de la base imparfaite et celles tirées de la base parfaite:

$$\sum_{j=1}^{M} Y_j^* = M\bar{Y}^* = \sum_{i=1}^{N} Y_i = N\bar{Y};$$

$$\sum_{j=1}^{M} C_j^* = M\bar{C}^* = N;$$

$$\sum_{j=1}^{M} (Y_j^* - \bar{Y}^*)^2 = N\sigma_Y^2 - S(2, Y) + (N\bar{Y})^2(1/N - 1/M),$$

où

$$N\sigma_Z^2 = \sum_{i=1}^{N} (Z_i - \bar{Z})^2$$

et

$$S(a,Z) = \sum_{k=2}^{r} (1 - 1/k)\left\{\sum_{i:U_i \in A_k} Z_i^a\right\}; \tag{2.2}$$

$$\sum_{j=1}^{M} (C_j^* - \bar{C}^*)^2 = N(1 - N/M) - S(0, Y);$$

$$\sum_{j=1}^{M} (Y_j^* - \bar{Y}^*)(C_j^* - \bar{C}^*) = N\bar{Y}(1 - N/M) - S(1, Y).$$

Pour l'unité $U_i$, posons

$$D_i = Y_i - \bar{Y}; \; W_i = Y_i - RX_i, \quad \text{où} \quad R = \bar{Y}/\bar{X}. \tag{2.3}$$

Comme on ne suppose pas l'existence d'information supplémentaire sur les unités, les comparaisons se feront en fonction d'un ÉASSR. Soit $m$ la taille de l'échantillon tiré de la base imparfaite et $n$, l'effectif correspondant s'il s'agissait d'une base parfaite. Définissons l'efficience d'une base

parfaite par rapport à une base imparfaite, pour n'importe quel estimateur, par l'expression

$$\rho = \frac{\text{EQM d'après un échantillon de taille } m \text{ provenant de la base imparfaite}}{\text{EQM d'après un échantillon de taille } n \text{ qui proviendrait d'une base parfaite}}. \tag{2.4}$$

Posons aussi $f$ comme la fraction de sondage commune lorsque les fractions de sondage sont les mêmes, c.-à-d.,

$$n = fN, \quad m = fM = n(1 + \alpha). \tag{2.5}$$

## 3. RÉSULTATS

Avant de répondre à la question fondamentale que nous nous sommes posée dans la section 1, à savoir s'il faut échantillonner à partir d'une base parfaite ou d'une base imparfaite, nous allons considérer brièvement les deux options dans une perspective de coûts. Si on prévoit que la mise à jour de la base imparfaite coûtera plus cher que la collecte de données qui devra se faire auprès des $(m - n)$ unités additionnelles, il est plus avantageux d'utiliser la base imparfaite avec un échantillon plus grand que de mettre à jour cette base; il en est ainsi lorsque

$$\frac{b_1}{b_0}\left(\frac{m - n}{N}\right) \leq 1, \tag{3.1}$$

où $b_1$ est le coût unitaire de la collecte de données et $b_0$, le coût unitaire de la mise à jour. Notons qu'il faut visiter effectivement $N$ unités pour mettre à jour la base incomplète puisque les $(M - N)$ autres unités sont des répétitions et qu'elles peuvent être identifiées grâce à l'hypothèse b). Notons aussi que même dans le cas d'un ÉASSR issu de la base imparfaite, le nombre supplémentaire d'unités à visiter est tout au plus de $(m - n)$ puisque l'échantillon peut compter plusieurs répétitions de la même unité sous des identités différentes. Ces observations conduisent à l'inégalité (3.1), ce qui confirme la préférence pour une base imparfaite.

Comme nous l'avons mentionné dans la section 1, la base imparfaite ne permettra pas de connaître le nombre total d'unités de population, $N$. L'estimation d'une moyenne et l'estimation d'un total sont donc des problèmes différents; le problème de l'estimation de la moyenne est essentiellement le problème de l'estimation du ratio, mais quant à l'estimation du total, il est possible d'estimer directement et sans biais un total à partir d'un ÉASSR de taille $m$ tiré de la base imparfaite. Par conséquent, il est convenable d'estimer un ratio de population (procédé analogue à l'estimation pour domaine), et de considérer l'estimation de moyenne comme un cas particulier de l'estimation de ratio, puis d'envisager séparément l'estimation d'un total.

### 3.1 Estimation d'un ratio

Pour l'estimation du ratio $R = (\bar{Y}/\bar{X})$, l'estimateur habituel est

$$\hat{R} = \bar{y}^*/\bar{x}^*,$$

où les lettres minuscules représentent les quantités correspondantes pour un échantillon; $\bar{y}^*$ est la moyenne des valeurs $Y^*$ établie d'après un échantillon de taille $m$ issu de la base imparfaite, *etc.*; $\bar{y}^*$ et $\bar{x}^*$ sont des estimateurs non biaisés de $(N\bar{Y}/M)$ et de $(N\bar{X}/M)$ respectivement. En utilisant la méthode delta, on peut définir une formule approximative pour l'EQM de $\hat{R}$, $E(\hat{R} - R)^2$, à savoir

$$\frac{M - m}{m(\bar{X}^*)^2(M - 1)M}\sum_{i=1}^{M} W_i^{*2}; \tag{3.2}$$

utilisant les relations de la section 2, on peut réécrire (3.2) sous la forme

$$\text{EQM}(\hat{R}) = \frac{M(M - m)}{m(N\bar{X})^2(M - 1)}\{N\sigma_W^2 - S(2,W)\},$$

où les valeurs $W$ sont définies en (2.3) et les valeurs $W^*$ calculées de la même manière. Une conséquence de (2.2) est que $S(2,W) \geq 0$, ce qui nous permet d'écrire, d'après (3.2),

$$0 \leq 1 - \frac{S(2,W)}{N\sigma_W^2} \leq 1. \tag{3.3}$$

Maintenant d'après l'équation (2.4), nous pouvons exprimer l'efficience $\rho$ par la formule

$$\rho = \frac{nM(M - m)(N - 1)}{mN(N - n)(M - 1)}\left\{1 - \frac{S(2, W)}{N\sigma_W^2}\right\}. \tag{3.4}$$

Lorsque les fractions de sondage sont égales, $\rho$ peut être exprimée comme suit:

$$\rho = \frac{(1 + \alpha)(N - 1)}{(1 + \alpha)(N - 1) + \alpha}\left\{1 - \frac{S(2,W)}{N\sigma_W^2}\right\}. \tag{3.5}$$

On peut donc affirmer, si l'on s'appuie sur (3.3), que $\rho$, telle qu'elle est définie en (3.5), satisfait l'inégalité

$$0 \leq \rho \leq 1 \tag{3.6}$$

et que, par conséquent, il est avantageux d'utiliser une base imparfaite pour estimer un ratio.

Il convient de souligner que $S(2,W)$ est non décroissant en $\alpha$ et que pour une valeur $\alpha$ fixe, $S(2, W)$ a une valeur plus élevée lorsque les unités qui ont des valeurs $W$

relativement plus élevées se répètent dans la base imparfaite. Comme $\sigma_W^2$ est fixe pour un ensemble donné de $N$ valeurs $W$, il peut y avoir des cas où $\rho$, tel qu'il est défini en (3.4), est plus petit que 1 (de fait, $S(2,W)$ est égal à $N\sigma_w^2$ lorsque les valeurs $W$ sont toutes égales à 0) et par conséquent, il y aura des cas où l'échantillonnage à partir d'une base imparfaite sera préférable à l'échantillonnage à partir d'une base complète, même si la fraction de sondage est moindre dans la première situation.

### 3.2 Estimation d'une moyenne

Comme nous l'avons vu dans la section 3.1, $\bar{y}^*$ est un estimateur non biaisé de $(N\bar{Y})/M$, où $M$ est connu mais $N$, inconnu. Il est donc nécessaire d'estimer $N$ pour obtenir un estimateur de $\bar{Y}$. Notons que $\bar{c}^*$ est un estimateur non biaisé de $(N/M)$ et que, par conséquent,

$$\hat{\bar{Y}} = \bar{y}^*/\bar{c}^*$$

est un estimateur naturel de $\bar{Y}$ du type "estimateur par quotient". Si on remplace $\bar{x}^*$, défini dans la section 3.1, par $\bar{c}^*$, l'EQM de $\hat{\bar{Y}}$ est rendue par l'expression

$$\mathrm{EQM}(\hat{\bar{Y}}) = \frac{M(M-m)}{mN^2(M-1)}\,[N\sigma_D^2 - S(2,D)],$$

où les valeurs $D$ sont définies en (2.3). Si on remplace $W$, dans la section 3.1, par $D$, on peut conclure que (3.6) se vérifie et qu'il est plus avantageux d'utiliser une base imparfaite lorsque (2.5) est vraie.

### 3.3 Estimation d'un total

Si on veut estimer un total, par exemple $N\bar{Y}$, en fonction d'un ÉASSR de taille $m$ tiré d'une base imparfaite, on utilise habituellement l'estimateur

$$\widehat{(N\bar{Y})} = M\bar{y}^*,$$

qui est non biaisé pour $N\bar{Y}$ et dont la variance est définie

$$\begin{aligned}\mathrm{EQM}(M\bar{y}^*) &= \mathrm{Var}(M\bar{y}^*)\\ &= \frac{M(M-m)}{m(M-1)}\\ &\quad\left\{N\sigma_Y^2 - S(2,Y) + (N\bar{Y})^2\left(\frac{1}{N}-\frac{1}{M}\right)\right\}.\end{aligned}$$

On peut exprimer $\rho$ par la formule

$$\rho = \frac{nM(M-m)(N-1)}{mN(N-n)(M-1)}\left\{1 - \frac{S(2,Y) - (N\bar{Y})^2(1/N - 1/M)}{N\sigma_Y^2}\right\}.$$

Il est clair d'après l'expression de $\mathrm{Var}(M\bar{y}^*)$ que

$$\left\{S(2,Y) - (N\bar{Y})^2\left(\frac{1}{N}-\frac{1}{M}\right)\right\}\Big/N\sigma_Y^2, \qquad (3.7)$$

est égal ou inférieur à un. Cependant, il peut arriver que les valeurs $\alpha$ et $Y$ choisies soient telles que l'expression (3.7) sera négative. Dans un tel cas, l'utilisation d'une base imparfaite avec une fraction de sondage relativement plus élevée sera inefficiente, même si (2.5) est vraie. Néanmoins, si le diagramme de dispersion des valeurs $Y^*$ est plus homogène que celui des valeurs $Y$, c.-à-d. si

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \geq \sum_{j=1}^{M}(Y_j^* - \bar{Y}^*)^2, \qquad (3.8)$$

l'expression (3.7) sera toujours non négative. On peut alors tirer des conclusions semblables à celles énoncées dans la section 3.1, par exemple, (3.6) se vérifie lorsque (2.5) est vraie.

## 4. EXEMPLE

Comme nous l'avons souligné plus haut, les unités d'échantillonnage qui devaient servir à l'enquête sur les pêcheurs étaient tirées de la liste de pêcheurs-bénéficiaires qui existait à ce moment-là. Puisque cette enquête était polyvalente, elle a permis d'observer de nombreuses caractéristiques des unités d'échantillonnage qui avaient rapport soit au ménage, soit à l'entreprise de pêche à laquelle appartenait l'unité d'échantillonnage. Comme on ne connaissait que le nombre de bénéficiaires ($M$), le nombre d'entreprises ou de ménages correspondants ($N$) étant inconnu, il n'a pas été possible d'évaluer l'effet de l'utilisation d'une base imparfaite pour cette enquête. Cependant, pour les besoins de cet article, nous allons considérer comme une population en soi les échantillons qui ont été tirés dans une région géographique (portion d'un district administratif du Bengale-Occidental) et nous allons étudier l'effet d'un rééchantillonnage dans cette population. On dénombre dans cette région 27 bénéficiaires ($M$) et 23 entreprises distinctes ($N$); dix-neuf de ces entreprises ont un propriétaire unique ($N_1$) et les quatre autres sont des entreprises à propriété conjointe ($N_2$). Les caractéristiques qui nous intéressent sont le coût de la régénération des plans d'eau ($Y$) et la superficie des plans d'eau exploités ($X$).

Les statistiques sommaires de $Y$ et de $X$ sont les suivantes:

$$\sum Y_i = 58{,}815, \quad \sum X_i = 23.36,$$

$$R = \left(\sum Y_i\right)\Big/\left(\sum X_i\right) = 2{,}517.77,$$

$$S(2,Y) = 212{,}201{,}800, \quad S(2,D) = 145{,}101{,}018,$$

$$S(2,W) = 104{,}505{,}327,$$

$$23\sigma_Y^2 = 442{,}702{,}791, \quad 23\sigma_X^2 = 13.6503 \quad \text{et}$$

$$23\sigma_W^2 = 394{,}790{,}716,$$

où $W$ défini en (2.3).

Pour évaluer l'effet d'un échantillonnage dans la liste des bénéficiaires, qui sont au nombre de 27, nous calculons des estimations des variables suivantes:

$R$ = frais de régénération à l'acre;

$\bar{X}$ = superficie moyenne des plans d'eau exploités, par entreprise (en acres);

$N\bar{X}$ = superficie totale des plans d'eau exploités par l'ensemble des 23 entreprises.

Le tableau ci-dessous donne l'efficience ($\rho$) pour diverses valeurs de $m$ et $n$.

**Tableau 1**

Efficience de l'échantillonnage dans une base parfaite par rapport à l'échantillonnage dans une base imparfaite ($\rho$)

| Taille d'échantillon | | Efficience pour les estimateurs de | | |
|---|---|---|---|---|
| $n$ | $m$ | $R$ | $\bar{X}$ | $N\bar{X}$ |
| 2 | 2 | 0.8695 | 0.6453 | 0.9508 |
| 4 | 4 | 0.8841 | 0.6561 | 0.9668 |
| 6 | 6 | 0.9022 | 0.6696 | 0.9866 |
| 8 | 8 | 0.9225 | 0.6866 | 1.0117 |
| 8 | 9 | 0.7791 | 0.5781 | 0.8519 |
| 10 | 10 | 0.9551 | 0.7088 | 1.0444 |
| 10 | 11 | 0.8172 | 0.6065 | 0.8937 |

On constate que dans la plupart des cas, l'échantillonnage à partir d'une base imparfaite est plus efficient.

## REMERCIEMENTS

Les auteurs tiennent à remercier un des rédacteurs associés et les arbitres pour leurs conseils utiles qui ont contribué à améliorer cet article.

## BIBLIOGRAPHIE

HANSEN, M.H., HURWITZ, W.N., et JABINE, T.N. (1963). The Use of imperfect lists for probability sampling at the U.S. Bureau of Census. *Bulletin de l'Institut Internationale de Statistique*, 40, 497-517, (avec discussion).

INDIAN STATISTICAL INSTITUTE (1988). *A study of Fishermen in West Bengal*: 1985-1986.

RAO, J.N.K. (1968). Some non-response sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

SINGH, R. (1977). A note on the use of incomplete multi-auxiliary information in sample surveys. *Australian Journal of Statistics*, 19, 105-107.

SINGH, R. (1983). On the use of incomplete frames in sample surveys. *Biometrical Journal*, 25, 545-549.

SZAMEITAT, K., et SCHAFFER, K.A. (1963). Imperfect frames in statistics and the consequences for their use in sampling. *Bulletin de l'Institut Internationale de Statistique*, 40, 517-538, (avec discussion).

WRIGHT, T., et TSAO, H.J. (1983). *A frame on frames: An annonated bibliography. Statistical Methods and Improvement of Data Quality*, (Éd. T. Wright). New York: Academic Press, 25-72.

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 211–216
Statistique Canada



# Un nouveau concept pour l'échantillonnage des forêts

FRANCIS A. ROESCH, JR., EDWIN J. GREEN et CHARLES T. SCOTT[1]

RÉSUMÉ

Un concept généralisé est présenté pour toutes les méthodes couramment utilisées d'échantillonnage des forêts. Selon ce concept, la forêt est perçue comme une image bidimensionnelle découpée en pièces comme un casse-tête, les pièces étant définies par les probabilités de sélection individuelles des arbres de la forêt. Ce concept produit un nombre fini d'unités d'échantillonnage sélectionnées de façon indépendante, contrairement à tous les autres concepts généralisés d'échantillonnage des forêts présentés jusqu'à maintenant.

MOTS CLÉS: Échantillonnage des forêts; échantillonnage ppt.

## 1. INTRODUCTION

L'échantillonnage des forêts est souvent accompli sous forme d'un processus en deux parties: un point de la forêt est d'abord choisi au hasard, puis une grappe d'arbres entourant ce point est incluse dans l'échantillon en vertu d'une règle donnée. Les deux règles les plus courantes sont l'échantillonnage par placettes (circulaires, à superficie fixe) et l'échantillonnage par points (horizontal). Dans le premier cas, tous les arbres pour lesquels le centre de la section transversale du tronc à 4.5 pieds au-dessus du sol se situe à l'intérieur d'une distance horizontale constante ($d$) du point choisi au hasard sont inclus dans l'échantillon. Dans le second cas, l'arbre $i$ est ajouté à l'échantillon si ce centre se trouve à l'intérieur d'une distance horizontale $\alpha r_i$ du point choisi au hasard, où $r_i$ est le rayon de la section transversale et $\alpha$ est une constante, choisie en fonction de l'intensité d'échantillonnage désirée. L'arbre $i$ est sélectionné avec une probabilité proportionnelle à $\pi d^2$ dans l'échantillonnage par placettes (la probabilité est la même pour tous les arbres) et avec une probabilité proportionnelle à $\pi r_i^2$ (surface terrière de l'arbre $i$) dans l'échantillonnage par points (les plus gros arbres ont une plus grande probabilité d'être choisis).

L'élément constituant l'unité d'échantillonnage, dans les diverses méthodes d'échantillonnage des forêts, est une question dont ont abondamment débattu les auteurs du domaine de la foresterie. Certains considèrent l'arbre comme l'unité d'échantillonnage (p. ex. Oderwald 1981), tandis que pour d'autres, la grappe d'arbres associée au point (p. ex. Palley et Horwitz 1961; Schreuder 1970), la placette circulaire (p. ex. Cunia 1965), ou le point (p. ex. Husch 1955) constitue l'unité d'échantillonnage. Ces différents points de vue s'appuient sur divers outils statistiques. Par exemple, l'utilisation de l'arbre comme unité d'échantillonnage oblige à utiliser la théorie de l'échantillonnage de populations finies, tandis que l'utilisation du point comme unité d'échantillonnage exige le recours à la théorie un peu plus avancée de l'échantillonnage de populations infinies. En outre, l'échantillonnage par placettes a été présenté, traditionnellement, comme une méthode ayant la placette comme unité d'échantillonnage, tandis que pour l'échantillonnage par points, l'unité d'échantillonnage utilisée a généralement été l'arbre ou le point. Par conséquent, la disparité apparente entre ces mécanismes d'échantillonnage très communs et très semblables est d'origine artificielle.

Nous allons présenter un concept d'unité primaire d'échantillonnage qui s'applique à tous les types de plans d'échantillonnage forestier en vertu desquels les arbres sont sélectionnés d'après l'emplacement d'un point choisi au hasard. Nous allons aussi montrer que ce concept est simple et qu'il donne un nombre fini d'unités d'échantillonnages mutuellement exclusives et prélevées de façon indépendante. Notons, par contraste, qu'à titre d'unités d'échantillonnage, ni les arbres ni les grappes ne possèdent à la fois ces deux caractéristiques; en effet, les arbres ne sont pas sélectionnés de façon indépendante et les grappes d'arbres ne sont pas mutuellement exclusives. Le concept que nous allons présenter diffère aussi de ceux où le point placé au hasard, ou encore la placette, est l'unité d'échantillonnage, car il y a dans ces derniers cas un nombre infini d'unités. Nous allons aussi faire valoir que ce concept, dans bien des cas, est celui qui convient le mieux aux besoins.

## 2. LE CONCEPT DU CASSE-TÊTE

Supposons que la forêt compte $N$ arbres, et que ceux-ci sont désignés 1, 2, ..., $N$. À ces $N$ arbres sont associées les valeurs d'intérêt $\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_N\}$, des cercles $K$ dénotés $K = \{K_1, K_2, \ldots, K_N\}$ et des zones de sélection ayant les tailles $\tilde{A} = \{\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_N\}$. Grosenbaugh et Stover (1957) ont d'abord défini les cercles $K$ dans le contexte de l'échantillonnage par points. Pour nos besoins, le cercle $K$ de l'arbre $i$, $K_i$, est un cercle imaginaire, dont

[1] Francis A. Roesch, Jr., Mathematical Statistician, Institute for Quantitative Studies, Southern Forest Experiment Station, USDA Forest Service, 701 Loyola Avenue, New Orleans, LA 70113; Edwin J. Green, Professor of Forestry, Cook College-Rutgers University, P.O. Box 231, New Brunswick, NJ 08903; et Charles T. Scott, Project Leader, Forest Ecosystem Modeling Unit, Northeastern Forest Experiment Station, USDA Forest Service, 359 Main Road, Delaware, Ohio 43015.

le point milieu est le centre de l'arbre, et dont le rayon est $d$ dans le cas de l'échantillonnage par placettes et $\alpha r_i$ dans le cas de l'échantillonnage par points. La zone de sélection de l'arbre $i$, de taille $A_i$ (en acres), est la partie du cercle $K$ de l'arbre $i$ qui est à l'intérieur de la forêt, et c'est la zone pour laquelle l'arbre sera inclus dans l'échantillon si un point choisi au hasard se trouve dans cette zone.

Dans leur analyse de l'échantillonnage par points, Palley et Horwitz (1961) affirment que "... l'unité primaire d'échantillonnage est une grappe d'arbres associée à un lieu d'origine. Le lieu d'origine est un point dans le cas de l'échantillonnage par points ...". En fait, le lieu d'origine n'est pas un point, parce que la grappe d'arbres n'est pas sélectionnée uniquement à partir de ce point, mais plutôt à partir d'un ensemble infini de points situés à l'intérieur d'une zone précise.

Nous proposons l'approche différente qui consiste à considérer que les unités d'échantillonnage sont les sections mutuellement exclusives de terrain résultant de la superposition des zones de sélection des arbres individuels de la forêt.

La façon dont le terrain est découpé en unités primaires d'échantillonnage est clairement illustrée à la figure 1. La correspondance entre la population, la base d'échantillonnage et l'unité d'échantillonnage, comme elle est présentée par exemple dans Cochran (1977, p. 6), apparaît clairement: la population (*ou l'image du casse-tête*) est divisée en unités d'échantillonnage exhaustives et mutuellement exclusives (*les pièces du casse-tête*) qui, ensemble, forment la base d'échantillonnage. Chaque segment de terrain a une probabilité de sélection définie, et le total de ces probabilités pour l'ensemble des segments est 1. Nous allons appeler cette approche le concept du casse-tête.

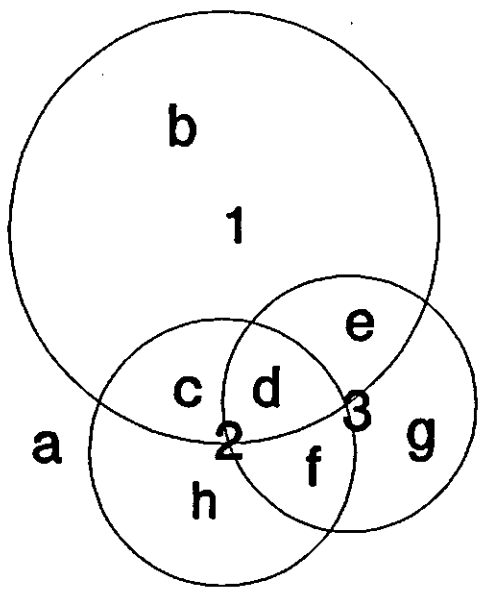


**Figure 1.** Les pièces du casse-tête. Les arbres 1, 2 et 3 ont leur centre là où se trouve le chiffre correspondant. Les cercles qui les entourent représentent les zones de sélection des arbres. Chacun des segments désigné par une lettre constitue une unité d'échantillonnage.

À chaque segment de terrain sont associés des attributs d'intérêt, dont la mesure donnera des valeurs identiques à partir de n'importe quel point de ce segment de terrain. L'élément essentiel est que chaque point est équivalent aux autres à l'intérieur d'un segment particulier. Les segments de terrain sont évidemment sélectionnés avec probabilité proportionnelle à la taille. Dans le cas de l'échantillonnage par points, la taille des segments est déterminée par les surfaces terrières et la distribution spatiale des arbres, ainsi que par la constante $\alpha$ choisie. Dès que $\alpha$ est choisie, la base d'échantillonnage à un point du temps donné est fixée. Dans le cas de l'échantillonnage par placettes, la taille du segment est déterminée par $d$ et par la distribution spatiale des arbres. Par conséquent, peu importe la méthode utilisée pour déterminer les arbres à inclure dans l'échantillon (p. ex. échantillonnage par placettes ou échantillonnage par points), on peut imaginer que dans chaque cas, on découpe l'image du casse-tête d'une certaine façon, on sélectionne des pièces avec probabilité proportionnelle à la taille et on retourne chaque pièce pour en lire à l'endos les attributs.

À l'appui de notre affirmation énoncée plus haut selon laquelle cette méthode est souvent celle qui convient le mieux, notons que l'objectif de la plupart des échantillonnages forestiers est de décrire la *forêt*, et non les arbres individuels. Nos agrégations sont souvent faites par acre ou par hectare, c.-à-d. des unités de la terre forestière, et non des unités de l'arbre. À partir du même endroit, nous pouvons faire des mesures portant sur de nombreux éléments autres que les arbres, par exemple les caractéristiques topographiques et d'autres données sur le site. Il est donc préférable, en général, de considérer des parties de la forêt, plutôt que des arbres individuels, comme les unités d'échantillonnage.

Nous allons nous limiter, dans la présente communication, à un traitement général de l'échantillonnage des forêts, mais notre analyse s'applique à n'importe quel type particulier d'échantillonnage des forêts qui se fonde sur la sélection d'arbres selon une fonction liée à des points choisis au hasard. La seule différence est la définition des segments de terrain, c'est-à-dire la façon dont l'image du casse-tête est découpée. Par exemple, dans l'échantillonnage par placettes, le terrain est divisé en pièces définies par des cercles superposés de taille égale, tandis que dans l'échantillonnage par points, la définition fait intervenir des cercles superposés de tailles proportionnelles à la surface terrière de chaque arbre.

Imaginons, pour les besoins de notre analyse, qu'un point est choisi au hasard à la surface d'une forêt et qu'une fonction quelconque est utilisée pour sélectionner les arbres qui feront partie de l'échantillon. Supposons aussi que notre forêt compte trois arbres (1, 2 et 3) dont les zones de sélection se chevauchent. À la figure 1, le centre des arbres 1, 2 et 3 se trouve là où le chiffre apparaît, et des cercles délimitent la zone de sélection de ces arbres. Chaque segment désigné par une lettre représente une unité d'échantillonnage différente. Si le point tombe dans le segment $a$, la grappe vide est choisie, s'il tombe dans le segment $b$, la grappe contenant seulement l'arbre 1 est

choisie, et s'il tombe dans le segment *d*, les trois arbres sont choisis, *etc.* L'arbre 1 serait donc sélectionné à partir des segments *b, c, d* et *e*. Cela donne une situation qui ressemble à celle décrite dans Kish (1965, section 11.2), c'est-à-dire que si les arbres formaient nos unités primaires d'échantillonnage, notre base d'échantillonnage serait une liste comprenant des inscriptions répétées de la même unité. Dans ce cas, la liste serait formée de grappes d'arbres, et la plupart des arbres seraient associés à plus d'une grappe. Les grappes sont sélectionnées avec probabilité proportionnelle à la taille du segment de terrain. La technique courante de pondération des éléments répétés d'une liste, examinée par Kish, se fonde plutôt sur la sélection d'unités primaires avec probabilité égale.

La méthode du casse-tête, en un sens, réduit la complexité du mécanisme d'échantillonnage, en appliquant d'abord la population d'arbres sur la population des segments de terrain, et en réduisant ainsi la base d'échantillonnage, formée d'une liste de grappes d'arbres dans laquelle les arbres appartiennent à plus d'une grappe, à une liste de segments de terrain uniques. Notre affirmation ci-dessous selon laquelle les simulations d'échantillonnage de forêts peuvent être simplifiées par la méthode du casse-tête est entièrement justifiée par la comparaison entre le coût initial de cette simplification de la liste d'échantillonnage et le besoin de prélever de nombreux échantillons à partir de cette liste.

Pour l'application de la population des arbres sur la population de segments, il faudrait préférablement qu'une observation, pour un segment, soit la somme de valeurs pondérées associées aux arbres, le poids pour chaque arbre étant proportionnel à sa probabilité d'être observé à partir de ce segment particulier. La probabilité qu'un arbre de l'échantillon *i* ait été sélectionné à partir du segment de terrain particulier *j* est:

$$p_{ij} = \left(\frac{A_j}{\tilde{A}_i}\right) Z_{ij},$$

où:

$A_j$ = superficie du segment *j* en acres, et

$$Z_{ij} = \begin{cases} 1 \text{ si le segment } j \text{ fait partie du cercle } k \text{ de l'arbre } i \\ 0 \text{ sinon.} \end{cases}$$

La somme sur *j* des $p_{ij}$ est 1. Nous pouvons maintenant écrire l'observation pour chaque segment sous forme d'une somme de valeurs pondérées associées aux arbres:

$$y_j = \sum_{i=1}^{N} p_{ij}\, \tilde{y}_i. \tag{1}$$

Supposons maintenant que nous choisissions au hasard *m* points à la surface d'une forêt, avec les mêmes hypothèses que ci-dessus (nous effectuons un échantillonnage avec remplacement). Un estimateur non biaisé du total de la valeur d'intérêt pour un échantillon prélevé avec probabilité proportionnelle à la taille est donné par:

$$\begin{aligned} \hat{Y} &= \frac{A_T}{m} \sum_{j=1}^{m} \frac{y_j}{A_j} \\ &= \frac{A_T}{m} \sum_{j=1}^{M} \frac{y_j}{A_j} W_j, \end{aligned} \tag{2}$$

où:

$A_T = \sum_{j=1}^{M} A_j$ ; la superficie totale de la forêt en acres,

$m$ = le nombre de points de l'échantillon,

$M$ = le nombre de segments de terrain, et

$W_j$ = le nombre de fois que la *j*ième unité apparaît dans l'échantillon.

Notons que $W_j$ est un entier entre 0 et *m* inclusivement. $A_j$ et $y_j$ sont fixes et $W_j$ est aléatoire. En outre, nous définissons:

$Y = \sum_{i=1}^{N} \tilde{y}_i$ ; le total de la valeur d'intérêt pour l'ensemble des arbres, et

$Y^* = \sum_{j=1}^{M} y_j$ ; le total de la valeur d'intérêt pour l'ensemble des segments.

Pour montrer que $\hat{Y}$ est non biaisé pour $Y$, nous allons d'abord montrer que $\hat{Y}$ est non biaisé pour $Y^*$, puis que $Y^*$ est égal à $Y$. De la même façon que Cochran (1977, p. 252-255), nous pouvons montrer que $\hat{Y}$ est non biaisé pour $Y^*$:

$$\begin{aligned} E[\hat{Y}] &= E\left[\frac{A_T}{m} \sum_{j=1}^{M} \frac{y_j}{A_j} W_j\right] \\ &= \frac{A_T}{m} \sum_{j=1}^{M} \frac{y_j}{A_j} E[W_j]. \end{aligned} \tag{3}$$

$W_j$ est une variable aléatoire multinomiale dont l'espérance est égale à $m(A_j/A_T)$. Par conséquent

$$E[\hat{Y}] = \sum_{j=1}^{M} y_j = Y^*. \tag{4}$$

Nous pouvons maintenant montrer que $\hat{Y}$ est non biaisé pour $Y$ en montrant que $Y^* = Y$. En substituant le côté droit de l'équation (1) à $y_j$ dans la définition de $Y^*$, nous obtenons:

$$Y^* = \sum_{j=1}^{M} \sum_{i=1}^{N} p_{ij} \tilde{y}_i. \tag{5}$$

En insérant la définition de $p_{ij}$ et en réorganisant l'ordre de sommation, nous obtenons:

$$Y^* = \sum_{i=1}^{N} \tilde{y}_i \left[ \frac{1}{\tilde{A}_i} \sum_{j=1}^{M} A_j Z_{ij} \right]. \tag{6}$$

Puisque

$$\tilde{A}_i = \sum_{j=1}^{M} A_j Z_{ij},$$

le terme entre crochets du côté droit de (6) est égal à 1, et

$$Y^* = \sum_{i=1}^{N} \tilde{y}_i = Y. \quad \text{C.Q.F.D.} \tag{7}$$

Par définition, la variance de $\hat{Y}$ est

$$V(\hat{Y}) = \left(\frac{1}{mA_T}\right) \sum_{j=1}^{M} A_j \left(\frac{A_T y_j}{A_j} - Y\right)^2. \tag{8}$$

L'estimation de la variance de l'échantillon est alors (Cochran 1977):

$$v(\hat{Y}) = \frac{1}{m(m-1)} \sum_{j=1}^{m} \left(\frac{A_T y_j}{A_j} - \hat{Y}\right)^2. \tag{9}$$

Le raisonnement général décrit par les équations (1) à (9) peut être appliqué à n'importe quel type particulier d'échantillonnage forestier conforme au processus en deux parties qui consiste à sélectionner des arbres à partir de points choisis de façon aléatoire.

À titre d'exemple additionnel de l'utilisation du concept du casse-tête, nous allons examiner la base d'échantillonnage obtenue quand l'échantillonnage par points est utilisé pour mesurer la croissance d'une forêt. Pour obtenir le maximum d'efficacité, on prend des mesures à deux moments différents, en utilisant les deux fois les mêmes points choisis au hasard. Ce genre d'échantillonnage relatif à la croissance des forêts est un échantillonnage par points avec deuxième mesure; ce type d'échantillonnage a fait l'objet d'un vaste traitement dans la littérature, le plus récemment par Van Deusen et coll. (1986) et Roesch et coll. (1989, 1991, 1993). Si un échantillonnage par points de ce type est effectué, et que la figure 1 représente le temps 1, le casse-tête pour l'échantillon global pourrait être découpé en pièces comme celles de la figure 2. Les arbres 1, 2 et 3 sont les mêmes que ceux de la figure 1, tandis que l'arbre 4 est apparu dans le peuplement entre les temps 1 et 2. Les cercles intérieurs représentent les zones de sélection de l'échantillon par points au temps 1 (disons $\alpha r_{i1}$, en incluant un indice pour le temps) et les cercles extérieurs représentent les zones de sélection de l'échantillon par points au temps 2 ($\alpha r_{i2}$ est plus grand en raison d'un accroissement de la surface terrière). L'arbre 4 n'a qu'un cercle extérieur car il n'existait pas au temps 1, et l'arbre 2 n'a qu'un cercle intérieur car il est mort avant le temps 2. Le cercle en pointillé représente la zone de sélection qu'aurait eue l'arbre 2 au temps 2 si le temps 2 était survenu juste avant la mort de l'arbre. Par conséquent, le cercle en pointillé ne contribue pas à la définition des segments.

Si le point choisi au hasard tombait dans le segment $a$, les arbres 1 et 3 seraient mesurés les deux fois et l'arbre 2 serait mesuré seulement au temps 1; s'il tombait dans le segment $b$, l'arbre 1 serait mesuré les deux fois et l'arbre 3 ne serait mesuré qu'au temps 2. Cela illustre le fait que même si une autre dimension a été ajoutée à l'échantillon (la dimension temps), le concept d'échantillon forestier demeure le même, puisque la dimension temps peut être appliquée à l'image du casse-tête. Donc, outre les conditions mentionnées ci-dessus, la définition des segments dépend des moments exacts où sont effectuées les deux

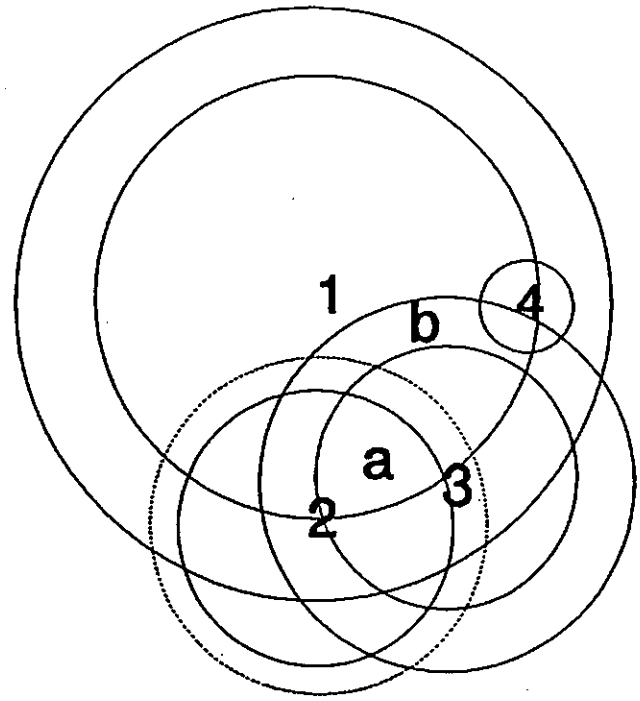


**Figure 2.** Les Pièces du casse-tête définies par emplacement, taille et temps. Un exemple d'unités d'échantillonnage dans le cas d'un échantillonnage par points avec deuxième mesure. Les arbres 1 et 3 ont poursuivi leur croissance et survécu, l'arbre 2 a continué de croître pendant quelque temps avant de mourir et l'arbre 4 est une recrue.

mesures. Cette notion d'unité d'échantillonnage est utile pour comprendre les estimateurs des composantes du changement entre le temps 1 et le temps 2 donnés dans Van Deusen et coll. (1986) et Roesch et coll. (1989 et 1991).

## 3. ANALYSE

Compte tenu de la simplicité du concept du casse-tête, on peut se demander pourquoi cette méthode d'échantillonnage des forêts n'a pas été proposée avant. La raison la plus évidente est probablement que les estimateurs ci-dessus ne peuvent être calculés si les $A_j$ sont inconnues. Puisque la zone de sélection d'un arbre particulier peut être répartie entre de nombreuses pièces du casse-tête et que la taille d'une pièce particulière du casse-tête peut être limitée par des arbres non sélectionnés à partir de cette pièce, les zones de sélection aussi bien des arbres de l'échantillon que des autres arbres doivent être connues pour qu'on puisse calculer les $A_j$ des segments sélectionnés. Par exemple, à la figure 1, si notre point tombait dans la section $c$, nous sélectionnerions les arbres 1 et 2, de sorte que la superficie $c + d$ pourrait être directement calculée. Toutefois, pour calculer $\hat{Y}$ et $v(\hat{Y})$, nous devons connaître la superficie du segment $c$ seul, et l'échantillonnage ne nous donne pas suffisamment d'information à cette fin. Nous allons voir que cette lacune apparente n'est pas importante, en montrant que $\hat{Y}$ peut être reformulé en des termes qui sont calculables. Ce sera, en fait, toujours le cas, peu importe à quelle méthode d'échantillonnage le concept du casse-tête sera appliqué.

Le concept du casse-tête, dans le cas de l'échantillonnage par points, revient à appliquer la population des arbres sur la population connexe des segments de terrain. Nous pouvons reformuler $\hat{Y}$ afin de montrer que ce dernier équivaut à l'estimateur d'échantillonnage par points habituel fondé sur la population des arbres. En développant l'équation (2) de manière à inclure la définition de $y_j$, et après quelques réagencements, on obtient:

$$\hat{Y} = \frac{A_T}{m} \sum_{j=1}^{M} \frac{y_j}{A_j} W_j$$

$$= \frac{A_T}{m} \sum_{j=1}^{M} \frac{\sum_{i=1}^{N} p_{ij} \tilde{y}_i}{A_j} W_j$$

$$= \frac{A_T}{m} \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{\tilde{y}_i Z_{ij} W_j}{\tilde{A}_i} \qquad (10)$$

$$= \frac{A_T}{m} \sum_{i=1}^{N} \frac{\tilde{y}_i}{\tilde{A}_i} \sum_{j=1}^{M} Z_{ij} W_j$$

$$= \frac{A_T}{m} \sum_{i=1}^{N} \frac{\tilde{y}_i}{\tilde{A}_i} w_i,$$

où $w_i$ est égal au nombre de fois que l'arbre $i$ est prélevé dans l'échantillon. L'expression finale dans (10) est l'estimateur de l'échantillonnage par points habituel.

La présente communication n'a pas pour but, par conséquent, de présenter un nouvel ensemble d'estimateurs pour des systèmes d'échantillonnage qui comptent déjà des estimateurs raisonnablement bons, mais plutôt de montrer comment des plans d'échantillonnage qui font l'objet de justifications très disparates dans la littérature se rejoignent à un niveau général. Cette perception différente peut être utile de multiples façons. Nous croyons, tout d'abord, que certains systèmes abstraits d'échantillonnage des forêts pourraient être plus faciles à comprendre s'ils étaient adaptés au cadre décrit ci-dessus. Nous avons constaté, par exemple, que les étudiants saisissent facilement la notion d'échantillonnage par points quand celui-ci est décrit comme une simple méthode de découpage de la forêt en pièces de casse-tête sans chevauchement, qui sont ensuite soumises à un échantillonnage avec probabilité proportionnelle à la taille. Les chercheurs intéressés à élaborer de nouveaux plans d'échantillonnage pour les forêts, ou de nouveaux estimateurs pour les plans existants, peuvent tirer parti de ce concept, car celui-ci leur offre une autre façon de percevoir leurs nouveaux plans et de programmer les simulations d'échantillonnage forestier qui serviront à tester les nouvelles méthodes. La simulation analysée dans Roesch (1993), par exemple, a été simplifiée par l'utilisation du concept du casse-tête plutôt que les autres concepts de base d'échantillonnage forestier qui avaient été proposés jusque-là. La simplification est venue du fait que l'essentiel de la simulation pouvait servir à de nombreux plans d'échantillonnage différents, seules quelques modifications minimes étant apportées au sous-programme de découpage du casse-tête.

Puisque les simulations d'échantillonnage forestier se fondent souvent sur une forêt cartographiée, les $A_j$ peuvent être obtenues directement. Une fois le casse-tête découpé, $y_j$ peut être calculé pour chaque pièce. Puis le programme de simulation choisit simplement les pièces à partir d'une liste, avec probabilité proportionnelle à la taille. Par contraste, voyons ce qu'il en est d'une simulation fondée sur l'utilisation du point comme unité d'échantillonnage. Dans ce cas, un point est sélectionné au hasard, et le programme parcourt la liste des arbres pour trouver tous ceux qui sont suffisamment près de ce point pour être inclus dans l'échantillon. Les attributs d'intérêt sont ensuite calculés. Puisque la probabilité de choisir le même point deux fois pour une population infinie est nulle, cette recherche dans la liste et ce calcul devraient être repris pour chaque point choisi au hasard, et entraîneraient peut-être un calcul répété des attributs pour la même grappe d'arbres. Pour les fins des simulations, la méthode optimale de programmation dépendra de la longueur de la liste d'arbres à parcourir, de la répartition en grappes dans la population des arbres et du nombre de points choisis au hasard.

## 4. CONCLUSION

Nous avons présenté un concept d'échantillonnage forestier généralisé qui utilise un nombre fini de segments de terrain à titre d'unités d'échantillonnage d'une base formée d'une zone terrestre. Nous avons aussi présenté des estimateurs fondés sur ce concept. Le concept du casse-tête devrait aider à comprendre les similitudes et les différences entre différentes méthodes d'échantillonnage des forêts, en intégrant toutes les méthodes au même cadre de référence. S'il est vrai que nous n'utiliserions pas normalement les estimateurs dans la forme présentée pour réaliser un échantillonnage forestier réel, nous pouvons toujours trouver une forme calculable équivalente. L'avantage additionnel de cette approche différente pour les simulations d'échantillonnage n'est pas seulement de nature théorique, mais aussi de nature économique. Compte tenu du temps et de l'argent qu'il faut consacrer à la collecte des données dans les études de foresterie, la capacité de tester facilement les propriétés de différentes méthodes d'échantillonnage avant de les mettre en oeuvre est d'une importance capitale. Nous ne voulons d'aucune façon sous-estimer la nécessité d'effectuer au départ un développement théorique rigoureux des plans d'échantillonnage forestier proposés, mais la simulation de ces plans avant leur mise en oeuvre peut aider à révéler des problèmes jusque-là non détectés. Cette nouvelle approche facilitera de façon générale les comparaisons à l'intérieur de n'importe quel groupe de plans d'échantillonnage des forêts.

## REMERCIEMENTS

Les auteurs tiennent à remercier le rédacteur associé, deux arbitres anonymes, ainsi que Hans Schreuder, pour leurs commentaires utiles.

## BIBLIOGRAPHIE

COCHRAN, W.G. (1977). *Sampling Techniques*, (3e Éd.). New York: John Wiley.

CUNIA, T. (1965). Continuous forest inventory, partial replacement of samples and multiple regression. *Forest Science*, 11, 480-502.

GROSENBAUGH, L.R., et STOVER, W.S. (1957). Point-sampling compared with plot-sampling in southeast Texas. *Forest Science*, 3, 2-14.

HUSCH, B. (1955). Results of an investigation of the variable plot method of cruising. *Journal of Forestry*, 53, 570-574.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

ODERWALD, R.G. (1981). Point and plot sampling - the relationship. *Journal of Forestry*, 79, 377-378.

PALLEY, M.N., et HORWITZ, L.G. (1961). Properties of some random and systematic point sampling estimators. *Forest Science*, 7, 52-65.

ROESCH, F.A. Jr. (1993). Adaptive cluster sampling for forest inventories. *Forest Science*, 39. À paraître.

ROESCH, F.A. Jr., GREEN,E.J., et SCOTT, C.T. (1989). New compatible estimators for survivor growth and ingrowth from remeasured horizontal point samples. *Forest Science*, 35, 281-293.

ROESCH, F.A. Jr., GREEN, E.J., et SCOTT C.T. (1991). Compatible basal area and number of trees estimators from remeasured horizontal point samples. *Forest Science*, 37, 136-145.

ROESCH, F.A. Jr., GREEN, E.J., et SCOTT, C.T. (1993). A test of alternative estimators for volume at time 1 from remeasured point samples. *Canadian Journal of Forest Research*, 23, 598-604.

SCHREUDER, H.T. (1970). Point sampling theory in the framework of equal-probability cluster sampling. *Forest Science*, 16, 240-246.

VAN DEUSEN, P.C., DELL, T.R., et THOMAS, C.E. (1986). Volume growth estimation from permanent horizontal points. *Forest Science*, 32, 415-422.

Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 217–227
Statistique Canada



# Enquêtes par panel: ajout d'une quatrième dimension

GRAHAM KALTON et CONSTANCE F. CITRO[1]

RÉSUMÉ

Les enquêtes qui consistent à recueillir des données dans le temps peuvent viser de nombreux objectifs. Dans la première moitié du présent article, nous examinons diverses options de plans d'enquête - enquêtes à passages répétés, enquêtes par panel, enquêtes par panel avec renouvellement et enquêtes à panel fractionné - pouvant permettre d'atteindre ces objectifs. La deuxième moitié est axée sur les enquêtes par panel. Nous y traitons des décisions qui doivent être prises au moment de la conception d'une enquête par panel, des problèmes posés par la non-réponse aux différentes vagues, du biais de conditionnement et de l'effet de lisière, ainsi que de certaines méthodes permettant l'analyse longitudinale des données d'enquête par panel.

MOTS CLÉS: Enquêtes par panel; enquêtes par panel avec renouvellement; enquêtes à passages répétés; érosion du panel; biais de conditionnement; effet de lisière; analyse longitudinale.

## 1. INTRODUCTION

Les populations visées par les enquêtes changent constamment dans le temps, aussi bien parce que leur composition évolue qu'en raison de modifications des caractéristiques de leurs membres. Des changements de composition surviennent quand des membres entrent dans la population au moment de la naissance (ou de l'atteinte de l'âge adulte), ou encore lorsqu'ils immigrent ou qu'ils sortent d'une institution (si la population étudiée ne comprend pas les personnes en institution); de même, les décès, l'émigration et l'entrée en institution sont des causes de changement. Les caractéristiques changent, par exemple, lorsque des personnes mariées divorcent, ou lorsque le revenu mensuel d'une personne passe de $2 000 à $2 500. Ces changements touchant la population sont à l'origine de toute une série d'objectifs des analyses de l'évolution des données d'enquête en fonction du temps. La présente communication passe en revue les plans d'enquête qui produisent les données nécessaires à l'atteinte de ces divers objectifs.

Cette communication se divise en deux parties. La première présente les aspects généraux de la réalisation d'enquêtes longitudinales, notamment les objectifs de ces enquêtes et les types de plans d'enquête possibles. Cet examen fait l'objet de la section 2. La deuxième partie, qui est le corps principal de la présente communication, s'intéresse au cas particulier des enquêtes par panel, qui consistent à suivre le même échantillon d'unités au fil du temps. Les différents aspects de la conception, de l'exécution et de l'analyse d'une enquête par panel sont exposés à la section 3. La section 4 présente quelques remarques en guise de conclusion.

## 2. ENQUÊTES LONGITUDINALES

La présente section décrit de façon générale les objectifs analytiques propres aux données qui évoluent dans le temps, les plans d'enquête qui conviennent à ces données et la mesure dans laquelle les différents plans peuvent satisfaire aux divers objectifs. L'exposé s'inspire fortement de celui de Duncan et Kalton (1987), qui contient un traitement plus détaillé de ces questions.

L'évolution des caractéristiques et de la composition de la population dans le temps est à la source de toute une gamme d'objectifs ayant trait aux enquêtes longitudinales. Ces objectifs sont, notamment:

(a) l'estimation des paramètres de la population (p. ex. la proportion de la population vivant sous le seuil de la pauvreté) à divers points dans le temps;

(b) l'estimation des valeurs moyennes de paramètres de la population dans le temps (p. ex. moyenne annuelle de la consommation quotidienne de fer);

(c) l'estimation de variations nettes, c'est-à-dire de variations à des niveaux d'agrégation élevés (p. ex. variation de la proportion de chômeurs d'un mois à l'autre);

(d) l'estimation de variations brutes et d'autres facettes de changements touchant les individus (p. ex. proportion de personnes qui étaient sous le seuil de la pauvreté une année et qui ne l'étaient plus l'année suivante);

(e) l'agrégation sur une certaine période de données relatives aux individus (p. ex. sommation des revenus de douze mois pour obtenir le revenu annuel);

(f) l'obtention de données sur des événements qui se produisent au cours d'une période donnée (p. ex. tomber en chômage) et sur leurs caractéristiques (p. ex. durée des périodes de chômage);

(g) l'accumulation progressive d'échantillons, notamment des échantillons de populations rares (p. ex. les femmes qui deviennent veuves);

(h) la tenue à jour d'un échantillon de membres d'une population rare observé à un moment particulier (p. ex. scientifiques et ingénieurs relevés dans une enquête à grande échelle à un moment donné).

[1] Graham Kalton, Westat, 1650 Research Blvd., Rockville (Maryland), É.-U. 20850; Constance F. Citro, National Research Council, 2101 Constitution Ave. N.W., Washington, D.C., É.-U., 20418.

Un certain nombre de plans d'enquête ont été mis au point pour la collecte des données nécessaires à l'atteinte de ces objectifs. En voici une liste:

- *Enquêtes à passages répétés.* Une enquête à passages répétés est une série d'enquêtes transversales distinctes réalisées à différents moments. On ne cherche aucunement à s'assurer que les éléments de l'échantillon soient les mêmes d'un passage à l'autre. Les éléments de l'échantillon sont tirés d'une population définie de la même manière à chaque enquête (p. ex. mêmes frontières géographiques et limites d'âge) et les questions, pour une bonne part, sont identiques d'un passage à l'autre.
- *Enquêtes par panel.* Une enquête par panel consiste à recueillir des données, à différents moments, auprès du même échantillon de répondants.
- *Enquête par panel à passages répétés.* Une enquête par panel à passages répétés est formée d'une série d'enquêtes par panel ayant chacune une durée fixe. Il peut n'y avoir aucun chevauchement des périodes visées par les différents panels, par exemple si un panel ne commence qu'au moment où le précédent se termine (ou plus tard), ou encore il peut y avoir un chevauchement, si deux panels ou plus couvrent partiellement la même période.
- *Enquête par panel avec renouvellement.* À strictement parler, une enquête par panel avec renouvellement est équivalente à une enquête par panel à passages répétés dans laquelle il y a un chevauchement. Dans les deux cas, la durée d'un panel est limitée, et deux panels ou plus sont en cours d'examen au même moment. Toutefois, il est utile de faire une distinction entre ces deux plans, car ils ne visent pas les mêmes objectifs. Les enquêtes par panel avec renouvellement sont largement utilisées pour produire une série d'estimations transversales et d'estimations de variations nettes (p. ex. taux de chômage et évolution de ces taux), tandis que les enquêtes par panel à passages répétés avec chevauchement accordent en outre beaucoup d'importance à des mesures longitudinales (p. ex. durée des périodes de chômage). Par conséquent les enquêtes par panel à passages répétés ont tendance à être de plus longue durée et à comporter moins de panels en cours de traitement à un moment quelconque que les enquêtes par panel avec renouvellement.
- *Enquête avec chevauchement.* Comme une enquête à passages répétés, une enquête avec chevauchement est formée d'une série d'enquêtes transversales menées à différents moments. Toutefois, tandis que l'enquête à passages répétés ne cherche pas à obtenir un chevauchement des échantillons de l'enquête entre deux moments successifs, une enquête avec chevauchement vise précisément à en obtenir un. Le but pourrait être de maximiser le degré de chevauchement des échantillons tout en tenant compte à la fois des variations désirées des probabilités de sélection des éléments de l'échantillon qui demeurent dans la population de l'enquête et des variations de la composition de la population au fil du temps.
- *Enquête à panel fractionné.* Une enquête à panel fractionné est la combinaison d'une enquête par panel et d'une enquête à passages répétés ou d'une enquête par panel avec renouvellement.

Le choix du plan dans une situation particulière dépend des objectifs que l'on veut atteindre. Certains plans sont supérieurs vis-à-vis de certains objectifs, mais moins performants à d'autres égards. Il y a des plans qui ne satisfont aucunement à certains objectifs. Pour un examen détaillé, voir Duncan et Kalton (1987).

Le principal atout d'une enquête à passages répétés est qu'elle comporte le prélèvement d'un nouvel échantillon à chaque passage, de sorte que chaque enquête transversale se fonde sur un échantillon probabiliste de la population qui existe à ce moment précis. Une enquête par panel se fonde sur un échantillon tiré de la population qui existait au début du panel. Bien que, parfois, des efforts soient faits pour ajouter à un panel des échantillons de nouveaux membres à des stades ultérieurs, une telle mise à jour se révèle généralement difficile et est empreinte d'imperfections. De plus, les pertes dues à la non-réponse qui sont enregistrées à mesure qu'un panel vieillit accentuent le problème du biais de non-réponse dont sont entachées, à des points ultérieurs, les estimations de paramètres transversaux tirées du panel. Pour ces raisons, les enquêtes à passages répétés sont supérieures aux enquêtes par panel pour produire des estimations transversales et des moyennes de ces dernières (objectifs (a) et (b)). Dans le cas des moyennes d'estimations transversales, un autre facteur à prendre en considération est la corrélation qui existe entre les valeurs des variables de l'enquête pour la même personne à différents moments. Si cette corrélation est positive, comme c'est le cas généralement, il en résulte un accroissement des erreurs-types des moyennes des estimations transversales provenant d'une enquête par panel. Ce facteur joue donc aussi en faveur des enquêtes à passages répétés par rapport aux enquêtes par panel, lorsqu'on s'intéresse à des moyennes d'estimations transversales.

La représentation plus fidèle des échantillons d'une enquête à passages répétés à des stades ultérieurs semblerait aussi favoriser ce type d'enquêtes par rapport aux enquêtes par panel pour l'estimation de la variation nette (en supposant qu'on veut mesurer cette dernière en tenant compte des changements touchant aussi bien la composition que les caractéristiques de la population). Toutefois, dans ce cas, les corrélations positives des valeurs des variables de l'enquête pour la même personne au fil du temps ont pour effet de diminuer les erreurs-types des estimations de la variation nette établie d'après une enquête par panel. Par conséquent, la présence de ces corrélations joue en faveur de l'enquête par panel pour la mesure de la variation nette.

Les avantages principaux des enquêtes par panel résident dans leur capacité de mesurer la variation brute, ainsi que de permettre l'agrégation, au fil du temps, de données sur les membres de l'échantillon (objectifs (d) et (e)). Les enquêtes à passages répétés ne peuvent répondre à ces objectifs. L'énorme potentiel d'analyse offert par la mesure de variations touchant les mêmes personnes est la principale raison incitant à recourir à une enquête par panel.

Les enquêtes à passages répétés permettent de recueillir des données sur des événements qui surviennent au cours d'une période donnée, ou encore sur la durée d'événements (p. ex. des périodes de maladie) au moyen de questions rétrospectives. Toutefois, les questions portant sur le passé, parce que les répondants éprouvent de la difficulté à se souvenir des dates, engendrent souvent un important problème d'erreur de réponse et font naître le risque d'un biais de télescopage. Une enquête par panel utilisant, pour les événements à étudier, une période de référence correspondant à l'intervalle entre les vagues de collecte des données peut éliminer le problème du télescopage, en utilisant l'interview précédente pour délimiter les événements (ainsi, on ne tiendra pas compte d'une maladie déclarée au cours de la présente interview si la même maladie avait été déclarée à l'interview précédente). De même, une enquête par panel peut permettre de déterminer la durée d'un événement d'après les vagues successives de collecte des données, en limitant la période de référence à l'intervalle entre les vagues.

Des collectes de données répétées au fil du temps peuvent être l'occasion de constituer progressivement un échantillon de membres d'une population rare, par exemple des personnes ayant une maladie chronique rare ou des personnes ayant récemment vécu un deuil. Les enquêtes à passages répétés peuvent être utilisées à cette fin et permettre de créer des échantillons de toutes sortes de populations rares. Les enquêtes par panel, toutefois, ne peuvent permettre d'accumuler que des événements rares nouveaux (p. ex. des deuils) et non des caractéristiques rares stables (p. ex. des personnes ayant une maladie chronique). Si un échantillon de membres ayant une caractéristique rare stable (p. ex. personnes ayant un doctorat) a déjà été établi, une enquête par panel peut être utile à la tenue à jour de cet échantillon, grâce à un apport approprié de nouveaux membres au moment de vagues ultérieures (voir, par exemple, Citro et Kalton 1989).

Les enquêtes par panel avec renouvellement visent principalement l'estimation des niveaux courants et de la variation nette (objectifs (a) et (c)). Dans de telles enquêtes, les éléments font généralement partie du panel pendant une courte période seulement. Par exemple, les membres de l'échantillon de l'enquête mensuelle sur la population active du Canada demeurent dans l'échantillon pendant une période de six mois seulement. Par conséquent, la mesure dans laquelle on peut évaluer les changements individuels et obtenir une agrégation des données dans le temps est limitée par la courte durée des panels. Une caractéristique spéciale des enquêtes par panel avec renouvellement réside dans la possibilité de recourir à l'estimation composite pour améliorer la précision tant des estimations transversales que des estimations de la variation nette (voir Binder et Hidiroglou 1988; Cantwell et Ernst 1993). Voir aussi Fuller et coll. (1993) pour une autre méthode d'utilisation de données antérieures dans le but de tirer des estimations d'un plan d'enquête par panel avec renouvellement.

Comme les enquêtes par panel avec renouvellement, les enquêtes avec chevauchement visent principalement l'estimation des niveaux courants et de la variation nette. Elles peuvent aussi fournir de l'information limitée sur la variation brute et sur des agrégations dans le temps. Les plans d'enquête avec chevauchement s'appliquent dans des situations où un certain chevauchement des échantillons est nécessaire et où les probabilités désirées de sélection des éléments varient dans le temps. C'est le cas notamment des enquêtes menées auprès des établissements, dans lesquelles la probabilité désirée de sélection d'un établissement peut varier d'une enquête transversale à l'autre, en fonction de la variation de la taille de l'établissement et de son type d'activité. En pareil cas, une procédure du type de celle de Keyfitz peut être appliquée afin de maximiser la conservation des éléments de l'enquête précédente, tout en tenant compte des changements des probabilités de sélection et de la composition de la population (voir, par exemple, Keyfitz 1951; Kish et Scott 1971; Sunter 1986). L'échantillon d'entreprises de la division des statistiques du revenu du Internal Revenue Service des États-Unis constitue un exemple de mise en oeuvre d'un plan d'enquête avec chevauchement (Hinkins et coll. 1988).

Une enquête à panel fractionné, parce qu'elle est une combinaison d'une enquête par panel et d'une enquête à passages répétés ou d'une enquête par panel avec renouvellement, peut offrir les avantages de chacune. Toutefois, si un plafond est imposé aux ressources globales disponibles, la taille de l'échantillon de chaque composante sera nécessairement plus faible que si une seule composante était utilisée. En particulier, les estimations des variations brutes et d'autres mesures des variations individuelles provenant d'une enquête à panel fractionné se fonderont sur un échantillon plus petit que ce n'aurait été le cas si toutes les ressources avaient été consacrées uniquement à la composante "enquête par panel".

Lorsqu'on compare différents plans d'enquêtes longitudinales, les coûts doivent être pris en considération. Par exemple, les enquêtes par panel permettent d'éviter les coûts de la sélection répétée d'échantillons qu'exigent les enquêtes à passages répétés, mais comportent des coûts pour le dépistage et le suivi des déplacements des membres de l'échantillon, et parfois aussi pour l'offre d'incitatifs destinés à s'assurer la fidélité des membres du panel (voir la section 3). Si deux plans peuvent l'un et l'autre répondre aux objectifs de l'enquête, les coûts relatifs pour des niveaux de précision donnés des estimations de l'enquête doivent être examinés.

## 3. ENQUÊTES PAR PANEL

Les mesures répétées portant sur le même échantillon que permettent les enquêtes par panel confèrent à ces enquêtes un avantage clé, sur le plan de l'analyse, par rapport aux enquêtes à passages répétés. Les mesures de la variation brute et d'autres composantes de la variation individuelle qu'on peut extraire des données d'une enquête par panel sont la source d'une bien meilleure compréhension des processus sociaux que celle que peut procurer une série de clichés transversaux indépendants. Le potentiel des données longitudinales tirées des enquêtes par panel est depuis longtemps reconnu (voir, par exemple, Lazarsfeld

et Fiske 1938; Lazarsfeld 1948) et, depuis de nombreuses années, des enquêtes par panel sont réalisées dans de nombreux domaines. Les sujets suivants, par exemple, ont fait l'objet d'enquêtes par panel: croissance et développement humains, délinquance juvénile, consommation de drogues, victimisation, comportement des électeurs, études de marketing sur les dépenses de consommation, choix d'études et de carrières, retraite, santé, frais médicaux. (Voir Wall et Williams (1970) pour un examen d'études par panel réalisées il y a longtemps sur la croissance et le développement humains, Boruch et Pearson (1988) pour des descriptions de certaines enquêtes par panel menées aux État-Unis, et le Subcommittee on Federal Longitudinal Surveys (1986) pour des descriptions d'enquêtes par panel fédérales américaines.) On a pu constater ces dernières années un important gain d'intérêt pour les enquêtes par panel dans de nombreux secteurs, notamment dans le domaine de l'information économique sur les ménages. L'enquête permanente américaine "Panel Study of Income Dynamics" a commencé en 1968 (voir Hill 1992, pour une description de la PSID) et des enquêtes par panel de longue durée du même genre ont été amorcées dans de nombreux pays européens au cours de la dernière décennie. Le U.S. Bureau of the Census a lancé l'enquête "Survey of Income and Program Participation" (SIPP) en 1983 (Nelson et coll. 1985; Kasprzyk 1988; Jabine et coll. 1990), et Statistique Canada a introduit l'Enquête sur la dynamique du travail et du revenu (EDTR) en 1993. L'intérêt accru manifesté à l'égard des enquêtes par panel s'est aussi accompagné de la parution d'un nombre croissant d'ouvrages sur la méthodologie de telles enquêtes; citons, à titre d'exemples récents, Kasprzyk et coll. (1989), Magnusson et Bergman (1990) et Van de Pol (1989).

La présente section décrit les principaux aspects qui interviennent dans la conception et l'analyse des enquêtes par panel. Nous nous intéressons principalement aux enquêtes par panel à passages répétés de durée fixe, comme l'enquête SIPP et l'enquête EDRT, mais l'essentiel de notre examen s'applique, de façon générale, à toutes les formes d'enquêtes par panel.

### 3.1 Plan d'une enquête par panel: les choix à faire

La dimension "temps" est une dimension de complexité additionnelle que possède une enquête par panel par rapport à une enquête transversale. En sus de toutes les décisions qui doivent être prises dans le cadre de la planification d'une enquête transversale, une vaste gamme de choix additionnels doivent être faits pour une enquête par panel. Voici les principales décisions à prendre:

- *Durée du panel.* Plus le panel est de longue durée, plus grande est la richesse des données pour les fins de l'analyse longitudinale. Par exemple, plus le panel durera longtemps, plus il y aura de périodes de chômage commençant pendant la durée du panel qui se termineront avant la fin du panel, et donc plus l'estimation de la fonction de survie pour de telles périodes sera précise. En revanche, plus le panel est de longue durée, plus il est difficile de garder un échantillon transversal représentatif aux vagues finales du panel, en raison aussi bien de l'érosion de l'échantillon que des difficultés de mise à jour de l'échantillon en fonction des nouveaux venus dans la population.

  Il peut parfois être bénéfique de faire varier la durée du panel selon divers types de membres. Ainsi, lorsque les objectifs de l'analyse l'exigent, les membres du panel ayant certaines caractéristiques (p. ex. les membres d'une minorité) ou qui vivent certaines événements au cours de la durée du panel normal (p. ex. un divorce) peuvent être gardés dans le panel pour des périodes d'observation plus longues.

- *Durée de la période de référence.* La fréquence de collecte des données dépend de la capacité des répondants de se souvenir de l'information demandée. Ainsi, la PSID, qui comporte des vagues annuelles de collecte des données, exige des répondants qu'ils se souviennent des événements survenus au cours de l'année civile précédente, tandis que l'enquête SIPP, avec des vagues de collecte des données aux quatre mois, exige que l'on se souvienne de ce qui s'est passé au cours des quatre mois précédents. Plus la période de référence est longue, plus le risque d'erreur de mémoire est élevé.

- *Nombre de vagues.* Dans la plupart des cas, le nombre de vagues de collecte des données est déterminé par la durée du panel et la durée de la période de référence. Plus il y a de vagues, plus on risque de subir une érosion du panel et des effets de conditionnement, et plus lourd est le fardeau imposé aux répondants.

- *Panels chevauchants ou non chevauchants.* Dans le cas d'une enquête par panel à passages répétés de durée fixe, une décision doit être prise quant au chevauchement des panels. Prenons le cas, par exemple, d'une proposition d'un groupe d'étude du National Research Council selon lequel l'enquête SIPP devrait comporter un panel de quatre ans (Citro et Kalton 1993). Une possibilité est d'avoir un panel de quatre ans, et de commencer un nouveau panel dès que le précédent prend fin. On pourrait aussi établir des panels de quatre ans dont un nouveau serait amorcé tous les deux ans. Autre possibilité, on pourrait avoir des panels de quatre ans, mais en commencer un nouveau chaque année.

  Un plan comportant des panels non chevauchants a l'avantage d'être simple, car un seul panel à la fois est en cours de traitement. Il produit aussi un vaste échantillon pour l'analyse longitudinale; par exemple, les panels non chevauchants peuvent être approximativement deux fois plus gros que ceux d'un plan comportant en tout temps deux panels chevauchants. Toutefois, cet avantage d'une taille d'échantillon plus grande qu'offrent les panels non chevauchants ne vaut pas pour les estimations transversales, car s'il y a plusieurs panels simultanés, les données qui concernent un moment précis peuvent être combinées pour les besoins de l'estimation transversale. En outre, les estimations visant des données transversales observées vers la fin d'un panel, dans le cas d'un plan sans chevauchement, sont davantage susceptible de souffrir d'un biais dû à l'érosion ou d'un biais

de conditionnement, ou encore de ne pas tenir compte pleinement des nouveaux éléments de la population, que ce n'est le cas pour un plan avec chevauchement, dans lequel l'un des panels est d'origine plus récente. En outre, le plan avec chevauchement permet l'examen de tels biais grâce à une comparaison des résultats des deux panels au cours d'une période donnée, ce que ne permet pas le plan sans chevauchement. Une autre limite du plan sans chevauchement est qu'il est susceptible de ne pas mesurer adéquatement l'effet d'événements comme des modifications législatives. Par exemple, si de nouvelles mesures législatives entrent en vigueur au cours de la dernière année d'un panel non chevauchant, il y aura peu de possibilité d'évaluer leur effet en comparant les situations des mêmes personnes avant, et pendant un certain temps après, l'instauration des nouvelles mesures. Avec des panels chevauchants, l'un des panels offrira une fenêtre d'observation plus large.

- *Taille du panel.* Pour un niveau annuel donné de ressources, la taille d'échantillon de chaque panel est déterminée par les facteurs qui précèdent. On peut obtenir un panel de grande taille pour l'analyse longitudinale en allongeant la période de référence et en utilisant un plan sans chevauchement. Pour des estimations transversales, l'allongement de la période de référence peut permettre d'augmenter la taille de l'échantillon, mais le recours à un plan sans chevauchement n'a aucun effet semblable.

La liste ci-dessus détermine les principaux paramètres d'un plan d'enquête par panel, mais il existe un certain nombre d'autres facteurs qui doivent aussi être examinés:

- *Mode de collecte des données.* Comme pour n'importe quelle enquête, il faut décider si les données seront recueillies par des interviews sur place, par téléphone ou par un questionnaire rempli par le répondant, et si l'on utilisera une méthode assistée par ordinateur (IPAO – interview sur place assistée par ordinateur ou ITAO – interview téléphonique assistée par ordinateur). Dans le cas d'une enquête par panel, une telle décision doit être prise à l'égard de chaque vague de collecte des données, car on peut vouloir faire varier les modes de collecte (par exemple, interview sur place à la première vague pour établir un contact et créer un lien, et interviews par téléphone ou par questionnaire postal à certaines vagues subséquentes). Si le mode de collecte est susceptible de changer d'une vague à l'autre, il faut examiner la question de la comparabilité des données entre les vagues. Parfois, un changement de mode peut entraîner un changement d'interviewer, par exemple si l'on passe d'une interview sur place à une interview téléphonique assistée par ordinateur, effectuée à partir d'un endroit central. Dans un tel cas, l'effet du changement d'interviewer sur la volonté du répondant de continuer de participer au panel et sur la comparabilité des réponses entre les vagues doit aussi faire l'objet d'un examen attentif.
- *Interview avec rétro-information.* Dans les enquêtes par panel, il est possible de rappeler aux participants les réponses données à des vagues antérieures. Cette technique d'interview avec rétro-information permet d'obtenir des réponses plus cohérentes entre les vagues, mais risque de produire un niveau de cohérence exagérément élevé. La facilité d'application de l'interview avec rétro-information dépend de la durée de l'intervalle entre les vagues et du mode de collecte des données. Le traitement des réponses d'une vague de manière à pouvoir les rappeler à la vague suivante est plus facile à accomplir si l'intervalle entre les vagues est long et si des techniques d'interview assistée par ordinateur sont employées. Edwards et coll. (1993) décrivent l'utilisation de rétro-information dans des interviews sur place assistées par ordinateur (IPAO), dans le cadre de la Medical Care Beneficiary Survey, enquête comportant trois interviews par année de chaque répondant.
- *Incitatifs.* De l'argent, ou encore d'autres types d'incitatifs (p. ex. tasses, calculatrices, sacs-repas) peuvent être offerts aux membres de l'échantillon pour qu'ils demeurent fidèles à l'enquête. Dans le cas d'une enquête par panel, les incitatifs peuvent servir non seulement à obtenir la participation initiale du répondant, mais aussi à s'assurer sa coopération pendant toute la durée du panel. Le moment qui convient le mieux à l'offre d'incitatifs dans une enquête par panel (p. ex. à la première vague, à une vague intermédiaire ou à la dernière vague du panel) est une question qui reste à résoudre. Les chercheurs utilisant les enquêtes par panel envoient souvent aux répondants un bulletin périodique relatif à l'enquête, donnant les plus récents faits saillants des résultats, à la fois pour susciter leur bonne volonté à l'égard de l'enquête et pour maintenir un contact avec eux (voir ci-dessous). L'envoi de cartes de souhaits au moment de l'anniversaire de naissance des répondants est un autre moyen fréquemment employé.
- *Règles relatives aux répondants.* Les données des enquêtes sont souvent recueillies auprès de répondants substituts lorsque les répondants devant être interviewés ne sont pas disponibles. Dans le cas d'une enquête par panel, il est alors possible que les répondants ne soient pas les mêmes d'une vague à l'autre, ce qui nuit à la comparabilité des données entre les vagues. Les règles relatives aux répondants, dans une enquête par panel, doivent tenir compte de ce facteur.
- *Plan d'échantillonnage.* La nature longitudinale d'une enquête par panel doit être prise en considération au moment de construire le plan d'échantillonnage visant la première vague. Un échantillonnage par grappes est souvent employé dans les enquêtes transversales avec interviews sur place, de façon à réduire les frais de déplacement et à limiter le travail de construction de la base de sondage à l'établissement de listes d'unités de logement pour certains segments seulement. Ces avantages sont obtenus au prix d'un d'accroissement de la variance des estimations de l'enquête, attribuable à l'utilisation de grappes. Le composition optimale des grappes dépend des divers coûts en cause et de l'homogénéité des variables de l'enquête à l'intérieur des grappes (voir, par exemple, Kish 1965). Dans le cas d'une enquête par panel, l'utilisation et l'ampleur des grappes devrait être

déterminée en tenant compte du panel dans son ensemble et de toutes les vagues de collecte de données. En particulier, l'avantage d'une réduction des frais du travail sur place disparaît pour les vagues au cours desquelles l'interview téléphonique ou le questionnaire postal est utilisé. Par ailleurs, la migration de membres du panel vers des endroits se trouvant à l'extérieur des grappes originales réduit, aux vagues ultérieures, l'avantage que procurait la formation initiale des grappes en termes de réduction des coûts du travail sur place. (Toutefois, certains avantages des grappes initiales demeurent valables dans le cas de la proportion élevée des personnes mobiles qui déménagent à l'intérieur de leur propre quartier.)

Le suréchantillonnage de certains sous-groupes de la population est largement utilisé dans les enquêtes transversales, afin de fournir un nombre suffisant de membres de ces sous-groupes pour permettre des analyses séparées. Voici des exemples de tels sous-groupes: personnes à faible revenu, minorités, groupe d'âge particulier, résidents d'un certain secteur géographique. Un tel suréchantillonnage peut aussi être utile aux enquêtes par panel, mais il faut se montrer prudent dans son application. Dans le cas des panels de longue durée, la prudence se justifie notamment par le fait que les objectifs de l'enquête peuvent changer avec le temps. Le suréchantillonnage visant à répondre à un objectif énoncé au début d'un panel peut se révéler nuisible aux objectifs qui s'imposeront plus tard. Un autre motif de prudence vient du fait que bon nombre des sous-groupes étudiés sont de nature transitoire (p. ex. personnes à faible revenu, personnes vivant dans un secteur géographique donné). Le suréchantillonnage de tels sous-groupes au début d'un panel pourrait avoir une valeur limitée lors des vagues ultérieures: certaines personnes incluses dans un tel suréchantillonnage sortiront du sous-groupe, tandis que d'autre exclues du suréchantillonnage y entreront. Troisièmement, la définition du sous-groupe désiré dans le contexte d'une analyse longitudinale doit être prise en considération. Par exemple, les données de l'enquête SIPP servent à estimer la durée des périodes de participation à divers programmes sociaux. Puisque ces estimations se fondent habituellement sur de nouvelles périodes qui commencent pendant la durée du panel, il pourrait être inutile de suréchantillonner les personnes déjà bénéficiaires de programmes sociaux. Voir Citro et Kalton (1993) pour une analyse du suréchantillonnage dans le contexte de l'enquête SIPP.

- Si l'on désire procéder au suréchantillonnage d'un certain sous-groupe de la population (p. ex. une population minoritaire) dans le cadre d'une enquête par panel, le suréchantillonnage pourrait exiger une vaste opération de sélection. Il faudrait évaluer le coût d'une telle sélection en tenant compte du panel entier, avec toutes ses vagues de collecte des données. Une opération de sélection coûteuse à la première vague pourrait bien être justifiable dans ce contexte.
- *Mise à jour de l'échantillon.* Lorsque le seul objectif d'une enquête par panel est d'effectuer une analyse longitudinale, il peut être suffisant de procéder comme pour une cohorte et de suivre simplement l'échantillon sélectionné au moment de la vague initiale. Cependant, si l'on veut aussi établir des estimations transversales, il peut être nécessaire de mettre à jour l'échantillon à chaque vague afin de tenir compte des nouveaux venus dans la population. Une mise à jour reflétant tous les types de nouveaux venus est souvent difficile, mais il est parfois possible d'établir une méthode assez simple pour incorporer certains types de nouveaux venus. Par exemple, dans un panel de personnes de tous âges, les bébés nés de femmes membres du panel une fois ce dernier amorcé peuvent être inclus comme membres du panel. La population d'inférence de l'enquête SIPP est formée des personnes âgées de 15 ans ou plus. En déterminant, dans l'échantillon initial, les membres des ménages qui n'ont pas encore 15 ans mais qui l'auront avant la fin du panel, en suivant ces derniers pendant la durée du panel et en les interviewant dès qu'ils atteignent 15 ans, on peut effectuer une mise jour d'un panel de l'enquête SIPP tenant compte de cette catégorie de nouveaux venus (Kalton et Lepkowski 1985).

Il importe également de prendre en considération les membres du panel qui sortent de la population à l'étude. Dans certains cas, le départ est irréversible (p. ex. décès), mais dans d'autres cas, il pourrait n'être que temporaire (p. ex. déménagement à l'étranger ou entrée en institution). Si l'on prend des mesures pour garder la trace des participants qui sortent du panel temporairement, ceux-ci peuvent être réadmis dès qu'ils réintègrent la population d'inférence.

Dans des enquêtes par panel comme l'enquête SIPP et la PSID, des données sont recueillies non seulement sur les membres des ménages de l'échantillon initial, mais également sur d'autres personnes – non membres de l'échantillon – avec lesquelles ils vivent lors de vagues ultérieures. L'objectif principal de la collecte de données d'enquête sur des personnes ne faisant pas partie de l'échantillon est de permettre de décrire la situation économique et sociale des membres de l'échantillon. Il convient, toutefois, de se demander si l'on devrait garder dans le panel une partie ou la totalité de ces personnes non membres de l'échantillon lorsqu'elles ne vivent plus avec les ménages du panel. Dans certains types d'analyse, il est utile de continuer de les suivre, mais cela oblige à leur consacrer une part importante des ressources de l'enquête.

Lorsque des données sur des personnes non membres de l'échantillon sont recueillies, il se peut que ces données servent simplement à décrire la situation des membres de l'échantillon, auquel cas les analyses se limitent aux membres de l'échantillon et les non-membres se voient attribuer un poids de zéro. Il est aussi possible d'inclure les personnes ne faisant pas partie de l'échantillon dans des analyses transversales. Dans ce cas, il faut élaborer une pondération appropriée des personnes appartenant et n'appartenant pas à l'échantillon, de façon à tenir compte des multiples façons dont les personnes peuvent figurer dans l'ensemble de données. Huang (1984), Ernst (1989),

ainsi que Lavallée et Hunter (1993), décrivent la méthode de pondération fondée sur des parts équitables qui peut être utilisée à cette fin.

- *Dépistage et suivi.* Une difficulté qui entrave la plupart des enquêtes par panel vient du fait que certains membres du panel ont déménagé depuis la dernière vague et ne peuvent pas être retrouvés. Il y a deux façons de traiter ce problème. Premièrement, on peut tenter d'éviter qu'il ne survienne en mettant sur place un processus de suivi des membres du panel entre les vagues. Une méthode largement utilisée, lorsqu'il y a un long intervalle entre les vagues, consiste à faire des envois postaux aux répondants entre les vagues, par exemple à leur envoyer des cartes d'anniversaire ou des bulletins sur l'enquête, et à demander au service de la poste de fournir des avis de changement d'adresse, le cas échéant. Un autre moyen consiste à demander aux répondants d'indiquer les noms, les adresses et les numéros de téléphone de personnes qui leur sont proches (p. ex. des parents), qui sont peu susceptibles de déménager et qui pourront fournir leur adresse s'ils élisent domicile ailleurs.

  La deuxième façon de traiter les cas de répondants introuvables est de mettre en oeuvre diverses méthodes de dépistage pour tenter de les retrouver. Avec des efforts et de l'ingéniosité, on peut atteindre des taux de réussite élevés. Certaines méthodes de dépistage peuvent être propres à la population à l'étude (p. ex. sociétés professionnelles pour les personnes faisant partie de groupes professionnels), tandis que d'autres peuvent être plus générales, par exemple: annuaire téléphonique, recherche informatisée de numéros de téléphones, annuaires téléphoniques inversés permettant d'obtenir les numéros de téléphone de voisins, suivi du courrier, registres des actes de mariage, immatriculation des véhicules, employeurs et services d'information financière. Il peut se révéler utile d'examiner les registres des décès, notamment dans le cas des enquêtes par panel de longue durée. Les membres du panel qui sont décédés peuvent ainsi être classés de façon appropriée, plutôt que d'être considérés comme des non-répondants. Les méthodes de dépistage sont examinées par Burgess (1989), Clarridge et coll. (1978), Crider et coll. (1971) et Eckland (1968).

### 3.2 Problèmes des enquêtes par panel

Les enquêtes par panel sont confrontées, comme toutes les autres enquêtes, à une vaste gamme de sources d'erreur non due à l'échantillonnage. La présente section n'examine pas toutes ces sources, mais en présente trois qui sont propres aux enquêtes par panel: la non-réponse de vague, le biais de conditionnement et l'effet de lisière.

#### 3.2.1 Non-réponse de vague

La non-réponse enregistrée à la première vague des enquêtes par panel correspond à celle qui s'observe dans les enquêtes transversales. Ce qui distingue les enquêtes par panel, c'est qu'elles font face à une non-réponse additionnelle au moment de vagues subséquentes. Certains membres du panel qui deviennent des non-répondants à une vague particulière ne répondent plus ensuite à aucune autre vague, tandis que d'autres répondent à nouveau à certaines ou à la totalité des vagues subséquentes. Les premiers sont souvent appelés des cas d'érosion, tandis que les seconds sont appelés cas de non-érosion. Les taux globaux de non-réponse de vague des enquêtes par panel s'accroissent au fur et à mesure des vagues, mais dans le cas des enquêtes bien gérées, le taux d'augmentation s'atténue nettement avec le temps. Par exemple, dans le cas du panel de l'enquête SIPP de 1987, la perte était de 6.7% à la vague 1, de 12.6% à la vague 2, puis elle a augmenté lentement jusqu'à 19.0% à la vague 7 (Jabine et coll. 1990). La tendance de la non-réponse à s'atténuer aux vagues ultérieures est rassurante, mais le cumul de la non-réponse sur de nombreuses vagues produit néanmoins des taux de non-réponse élevés aux dernières vagues d'un panel de longue durée. Par exemple, en 1988, après 21 rondes annuelles de collecte des données, le taux de non-réponse de l'enquête PSID pour les personnes qui vivaient dans les ménages faisant partie de l'échantillon en 1968 avait atteint 43.9% (Hill 1992).

Le choix entre les deux méthodes d'usage général qui permettent de traiter les données d'enquête manquantes – pondération et imputation – n'est pas immédiat dans le cas de la non-réponse de vague touchant les enquêtes par panel. Pour l'analyse longitudinale, la méthode de pondération consiste à laisser de côté tous les enregistrements ayant une ou plusieurs vagues manquantes et à essayer de compenser pour leur absence par des rajustements de pondération appliqués aux enregistrements restants. Cette méthode peut entraîner une importante perte d'information lorsque le fichier de données porte sur plusieurs vagues. Dans la méthode d'imputation, en revanche, toutes les données déclarées sont conservées, mais on doit faire des imputations à vaste échelle à titre de compensation pour les vagues manquantes. Un compromis est aussi possible: on peut utiliser l'imputation pour certains profils de non-réponse de vague (p. ex. ceux qui ne comportent qu'une seule vague manquante, et pour lesquels on dispose des données des deux vagues adjacentes), et la pondération pour d'autres (voir, par exemple, Singh et coll. 1990). Pour l'analyse transversale, des fichiers de données distincts peuvent être créés à chaque vague. Ces fichiers peuvent comprendre tous les répondants à cette vague et utiliser soit des rajustements de pondération, soit des imputations, pour tenir compte des non-répondants à cette vague. Kalton (1986) et Lepkowski (1989) examinent les méthodes générales de traitement de la non-réponse de vague, Lepkowski et coll. (1993) traitent des imputations visant à compenser la non-réponse de vague dans la SIPP, et Michaud et Hunter (1993) décrivent les plans relatifs au traitement de la non-réponse de vague dans l'EDTR.

Dans le cas de la non-réponse de vague, il est possible de recueillir une partie ou la totalité des données de la vague manquante en effectuant une interview subséquente. Toutefois, la qualité des données rétrospectives recueillies de cette façon doit être évaluée avec soin. Une expérience visant à mesurer l'utilité de cette approche a été

effectuée auprès du panel de 1984 de la SIPP. À cette fin, une formule de vague manquante a servi à recueillir des réponses relatives à un sous-ensemble minimum de questions de base pour la vague manquante (Huggins 1987; Singh 1993). Les analyses ont montré que les données recueillies à l'aide de la formule de vague manquante affichaient un nombre passablement moindre de transititons que les données repères pour ce qui est de la réception de revenus, des biens et du recours à l'aide gouvernementale. Par conséquent, l'usage de la formule de vague manquante a été abandonné. Les dossiers administratifs peuvent parfois constituer une autre source de données de base pour les vagues manquantes.

### 3.2.2 Effet de conditionnement

Le biais dû au conditionnement, ou effet lié au temps passé dans l'échantillon, est l'effet qui s'exerce sur les réponses d'un membre du panel à une vague donnée de collecte des données en raison de sa participation aux vagues précédentes. L'effet peut refléter simplement un changement de comportement à l'égard de la déclaration. Par exemple, un répondant peut avoir remarqué dans une interview précédente que le fait de répondre "oui" à une question entraîne une série de sous-questions qui, en revanche, ne sont pas posées si la réponse est "non". Dans ce cas, le répondant peut répondre "non" pour éviter d'avoir à répondre à ces questions additionnelles. Du côté positif, un répondant peut avoir constaté à des interviews précédentes qu'il fallait fournir des renseignements détaillés sur le revenu et se préparer aux interviews subséquentes en amassant au préalable l'information nécessaire. L'effet de conditionnement peut aussi se traduire par un changement du comportement réel. Par exemple, un répondant peut s'inscrire au programme des coupons alimentaires après avoir appris l'existence de ce programme grâce aux questions posées à ce sujet à des vagues précédentes de collecte des données.

Une récente étude expérimentale sur l'effet de conditionnement dans une enquête par panel de quatre ans sur les nouveaux mariés a permis de déceler un lien entre la participation à l'enquête et l'harmonie du mariage (Veroff et coll. 1992). Toutefois, cette étude s'est fondée sur des techniques d'interview en profondeur qui vont plus loin que celles utilisées dans la plupart des enquêtes. Un certain nombre d'études sur le conditionnement des panels ayant été effectuées dans des contextes d'enquête plus courants ont révélé que des effets de conditionnement se produisent parfois, mais qu'ils ne sont pas généralisés (Traugott et Katosh 1979; Ferber 1964; Mooney 1962; Waterton et Lievesley 1989).

Les enquêtes par panel avec renouvellement et chevauchement ont comme avantage de permettre la comparaison entre des estimations relatives à la même période obtenues de panels différents. De telles comparaison ont permis de détecter clairement la présence de ce qu'on appelle un "biais lié au groupe de renouvellement" dans les enquêtes sur la population active des États-Unis et du Canada (p. ex. Bailar 1975, 1989 et U.S. Bureau of the Census 1978 pour l'enquête "Current Population Survey" américaine; Ghangurde 1982 pour l'enquête sur la population active canadienne). Le biais lié au groupe de renouvellement peut être le fruit d'un biais de non-réponse et d'effets de conditionnement. Dans des analyses comparant les panels chevauchants de 1985, 1986 et 1987 de l'enquête SIPP, Pennell et Lepkowski (1992) ont observé peu d'écarts entre les résultats des différents panels.

### 3.2.3 Effet de lisière

Dans de nombreuses enquêtes par panel, des données sont recueillies relativement à des sous-périodes de la période de référence, à partir de la dernière vague de collecte de données. Dans l'enquête SIPP, par exemple, des données sont recueillies sur une base mensuelle à l'intérieur de la période de référence de quatre mois qui sépare les vagues. L'effet de lisière correspond à ce qu'on observe souvent avec cette forme de collecte de données, c'est-à-dire que les niveaux de changements déclarés entre des sous-périodes adjacentes (p. ex. adhérer à un programme social, puis cesser d'y participer d'un mois à l'autre) sont beaucoup plus élevés lorsque les données des deux sous-périodes proviennent de vagues différentes que lorsqu'elles viennent de la même vague. On a constaté dans l'enquête SIPP un très vaste effet de lisière, ayant trait aussi bien à l'état de prestataire qu'aux montants des prestations (voir, par exemple, Jabine et coll. 1990; Kalton et Miller 1991). L'effet de lisière s'est également manifesté dans l'enquête PSID (Hill 1987). Murray et coll. (1991) décrivent des méthodes permettant de réduire l'effet de lisière dans l'enquête sur la population active du Canada.

### 3.3 Analyse longitudinale

Il existe une documentation abondante, qui ne cesse de s'accroître, dans le domaine de l'analyse des données longitudinales, notamment plusieurs manuels qui en exposent la matière (p. ex. Goldstein 1979; Hsiao 1986; Kessler et Greenberg 1981; Markus 1979). La présente description ne peut être complète, et doit se limiter à quelques thèmes généraux.

- *Mesure de la variation brute.* Comme nous l'avons déjà signalé, un avantage essentiel de l'enquête par panel par rapport à l'enquête à passages répétés vient de la possibilité de mesurer la variation brute, c'est-à-dire la variation enregistrée au niveau individuel. La méthode de base pour mesurer la variation brute est de dresser un tableau comparatif des réponses fournies à une vague par rapport à celles fournies à la même question, lors d'une autre vague. Cette sorte d'analyse souffre toutefois d'une contrainte importante; en effet, les variations de l'erreur de mesure d'une vague à l'autre peuvent engendrer un biais prononcé dans l'estimation de la variation brute (pour un examen plus détaillé, voir Kalton et coll. 1989; Rodgers 1989; Abowd et Zellner 1985; Chua et Fuller 1987; Fuller 1990; et Skinner 1993).
- *Relation entre les variables dans le temps.* Les enquêtes par panel permettent de recueillir les données nécessaires à l'étude des liens qui existent entre des variables mesurées

à différents moments. Par exemple, à partir de données recueillies sur la cohorte britannique des naissances de 1946 dans le cadre de la National Survey of Health and Development, Douglas (1975) a constaté que les enfants qui étaient hospitalisés pendant plus d'une semaine ou qui subissaient des hospitalisations répétées entre les âges de 6 mois et de 3½ ans affichaient un comportement plus problématique à l'école et avaient des notes de lecture plus faibles à 15 ans. En principe, les enquêtes transversales peuvent utiliser des questions rétrospectives pour recueillir les données nécessaires à ce genre d'analyse. Cependant, les réponses à de telles questions sont souvent entachées par un niveau élevé d'erreur de mémoire et sont parfois empreintes de distorsions systématiques influant sur la relation étudiée.

- *Régression avec termes de différence.* Une régression dont les coefficients sont exprimés sous forme de différences peut permettre d'éviter un certain type d'erreur dans la spécification d'un modèle. Supposons que le modèle de régression approprié pour la personne $i$ au temps $t$ soit

$$Y_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \epsilon_{it},$$

où $x_{it}$ est une variable explicative dont la valeur change avec le temps et $z_{it}$ est une variable explicative qui est constante dans le temps (p. ex. sexe, race). Supposons en outre que $z_{it}$ ne soit pas observée, par exemple qu'elle soit inconnue. Alors $\beta$ peut encore être estimé, mais d'après une régression avec termes de différence:

$$Y_{i(t+1)} - Y_{(t)} = \beta(x_{i(t+1)} - x_{it}) + \epsilon_{i(t+1)} - \epsilon_{it},$$

(Rodgers 1989; Duncan et Kalton 1987).

- *Estimation des durées des périodes.* Les données recueillies dans les enquêtes par panel peuvent servir à estimer la distribution des durées d'événements comme le fait de participer à un programme de protection sociale. Dans des enquêtes par panel comme l'enquête SIPP, certaines personnes se trouvent déjà dans une période de participation au début du panel (périodes tronquées au départ), certaines personnes amorcent une période de participation durant le panel, et certaines périodes se poursuivent après la fin du panel (périodes tronquées à droite). Par conséquent, les périodes ne sont pas toutes observées au complet. La distribution des durées des périodes peut être estimée en appliquant des méthodes d'analyse de survie, par exemple la méthode d'estimation de Kaplan-Meier fondée sur une limite de produit, à toutes les nouvelles périodes (y compris celles qui sont tronquées à droite) qui commencent pendant la durée du panel (p. ex. Ruggles et Williams 1989).
- *Modèles à équations structurelles avec erreurs de mesure.* La séquence de collecte des données dans une enquête par panel se traduit par une mise en ordre claire des variables de l'enquête, dont l'analyse se prête bien à l'utilisation de modèles à équations structurelles. Cette forme d'analyse peut prendre en considération les erreurs de mesure et, grâce à plusieurs mesures répétées, peut traiter des structures d'erreur corrélées (p. ex. Jöreskog et Sörbom 1979).

## 4. CONCLUSION

Les ensembles de données construits à partir d'enquêtes par panels offrent habituellement une grande richesse de contenu pour l'analyse. Ils contiennent des mesures répétées pour certaines variables observées plusieurs fois, et des mesures d'autres variables qui sont examinées à une seule vague. Les interviews répétées du même échantillon donnent l'occasion de recueillir des données sur de nouvelles variables à chaque vague, ce qui produit des données sur toute une gamme de variables observées sur plusieurs vagues. Les données d'un panel peuvent être soumises aussi bien à une analyse longitudinale qu'à une analyse transversale. Les mesures répétées peuvent permettre d'examiner les profils de réponse individuels dans le temps et peuvent aussi être reliées à d'autres variables. Les variables mesurées à une vague unique peuvent être analysées à la lumière d'autres variables observées à cette vague, ainsi que de variables mesurées à d'autres vagues.

La richesse des données d'un panel n'est mise à contribution que si les données sont effectivement analysées, et si elles le sont dans les délais les plus brefs possibles. Mettre en oeuvre une enquête par panel, c'est comme faire tourner une immense roue: les étapes de la conception du questionnaire, de la collecte des données, du traitement et de l'analyse doivent être reprises à chaque vague successive. Il y a un réel danger que l'équipe de l'enquête soit submergée par ce processus, et que les données ne soient pas pleinement analysées. Pour éviter un tel écueil, il faut mettre en place le personnel adéquat et construire une organisation bien intégrée.

Il est souhaitable, par ailleurs, de préserver la simplicité du plan de l'enquête par panel. Le plan devrait être élaboré à la lumière d'objectifs clairement définis. Un bon sens critique doit être exercé au moment d'ajouter au plan des éléments de complexité visant à accroître la richesse des données du panel pour en élargir les applications. Bien que des arguments persuasifs puissent souvent être invoqués en faveur de tels ajouts, ceux-ci doivent être rejetés s'ils menacent l'exécution ordonnée de n'importe quelle étape du processus d'enquête.

Comme il a été indiqué plus haut, les erreurs de mesure ont des effets particulièrement néfastes sur l'analyse des variations individuelles révélées par les données d'une enquête par panel. L'affectation d'une partie des ressources de l'enquête à la mesure de l'ampleur de ces erreurs est donc une décision judicieuse (Fuller 1989). Les erreurs de mesure peuvent être étudiées soit par des études de validité (c.-à-d. par une comparaison des réponses de l'enquête avec des valeurs "vraies" tirées d'une source externe), soit par des études de fiabilité (p. ex. études fondées sur des réinterviews). Les résultats de telles études peuvent ensuite être utilisés dans le processus d'estimation et permettre d'effectuer des rajustements tenant compte des erreurs de mesure.

**BIBLIOGRAPHIE**


ABOWD, H.M., et ZELLNER, A. (1985). Estimating gross flows. *Journal of Business and Economic Statistics*, 3, 254-283.

BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

BAILAR, B.A. (1989). Information needs, surveys, and measurement errors. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 1-24.

BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (Vol. 6), (Éds. P.R. Krishnaiah et C.R. Rao). New York: North Holland, 187-211.

BORUCH, R.F., et PEARSON, R.W. (1988). Assessing the quality of longitudinal surveys. *Evaluation Review*, 12, 3-58.

BURGESS, R.D. (1989). Major issues and implications of tracing survey respondents. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 52-74.

CANTWELL, P.J., et ERNST, L.R. (1993). Nouveaux développements dans l'estimation composite pour l'enquête Current Population Survey. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 139-149.

CHUA, T.C., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

CITRO, C.F., et KALTON, G. (1989). *Surveying the Nation's Scientists and Engineers*. Washington DC: National Academy Press.

CITRO, C.F., et KALTON, G. (1993). *The Future of the Survey of Income and Program Participation*. Washington DC: National Academy Press.

CLARRIDGE, B.R., SHEEHY, L.L., et HAUSER, T.S. (1978). Tracing members of a panel: a 17-year follow-up. *Sociological Methodology*, (Éd. K.F. Schuessler). San Francisco: Jossey-Bass, 389-437.

CRIDER, D.M., WILLITS, F.K., et BEALER, R.C. (1971). Tracking respondents in longitudinal surveys. *Public Opinion Quarterly*, 35, 613-620.

DOUGLAS, J.W.B. (1975). Early hospital admissions and later disturbances of behaviour and learning. *Developmental Medicine and Child Neurology*, 17, 456-480.

DUNCAN, G.J., et KALTON, G. (1987). Issues of design and analysis of surveys across time. *Revue Internationale de Statistique*, 55, 97-117.

ECKLAND, B.K. (1968). Retrieving mobile cases in longitudinal surveys. *Public Opinion Quarterly*, 32, 51-64.

EDWARDS, W.S., SPERRY, S., et EDWARDS, B. (1993). Utilisation de l'IPAO au cours d'une enquête longitudinale: un rapport sur l'enquête Medicare Current Beneficiary Survey. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 25-34.

ERNST, L.R. (1989). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh), New York: John Wiley, 139-159.

FERBER, R. (1964). Does a panel operation increase the reliability of survey data: the case of consumer savings. *Proceedings of the Social Statistics Section, American Statistical Association*, 210-216.

FULLER, W.A. (1989). Estimation of cross-sectional and change parameters: Discussion. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 480-485.

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

FULLER, W.A., ADAM, A., et YANSANEH, I.S. (1993). Estimateurs pour des enquêtes longitudinales avec application à l'enquête Current Population Survey des É.-U. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 349-366.

GHANGURDE, P.D. (1982). Le biais de renouvellement de l'échantillon dans les estimations de l'EPA. *Techniques d'enquête*, 8, 94-111.

GOLDSTEIN, H. (1979). *The Design and Analysis of Longitudinal Studies*. New York: Academic Press.

HILL, D. (1987). Response errors around the seam: analysis of change in a panel with overlapping reference periods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 210-215.

HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.

HINKINS, S., JONES, H., et SCHEUREN, F. (1988). Design modifications for the SOI corporate sample: Balancing multiple objectives. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 216-221.

HSIAO, C. (1986). *Analysis of Panel Data*. New York: Cambridge University Press.

HUANG, H. (1984). Obtaining cross-sectional estimates from a longitudinal survey: Experiences of the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 670-675.

HUGGINS, V. (1987). Evaluation of missing wage data from the Survey of Income and Program Participation (SIPP). *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 205-209.

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

JÖRESKOG, K.G., et SÖRBOM, D. (1979). *Advances in Factor Analysis and Structural Equation Models*. Lanham MD: University Press of America.

KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.

KALTON, G., KASPRZYK, D., et McMILLEN, D.B. (1989). Nonsampling errors in panel surveys. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 249-270.

KALTON G., et LEPKOWSKI, J.M. (1985). Following rules in SIPP. *Journal of Economic and Social Measurement*, 13, 319-329.

KALTON, G., et MILLER, M.E. (1991). The seam effect with Social Security income in the Survey of Income and Program Participation. *Journal of Official Statistics*, 7, 235-245.

KASPRZYK, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues*. Document de travail du SIPP n° 8830. Washington DC: U.S. Bureau of the Census.

KASPRZYK, D., DUNCAN G., KALTON, G., et SINGH, M.P. (Éds.) (1989). *Panel Surveys*. New York: John Wiley.

KESSLER, R.C., et GREENBERG, D.F. (1981). *Linear Panel Analysis*. New York: Academic Press.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size: adjustment for changes in the probabilities. *Journal of the American Statistical Association*, 46, 183-201.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

KISH, L., et SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

LAVALLÉE, P., et HUNTER, L. (1993). Méthodes de pondération pour l'enquête sur la dynamique du travail et du revenu. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 77-88.

LAZARSFELD, P.F. (1948). The use of panels in social research. *Proceedings of the American Philosophical Society*, 42, 405-410.

LAZARSFELD, P.F., et FISKE, M. (1938). The panel as a new tool for measuring opinion. *Public Opinion Quarterly*, 2, 596-612.

LEPKOWSKI, J.M. (1989). Treatment of wave nonresponse in panel surveys. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 348-374.

LEPKOWSKI, J.M., MILLER, D.P., KALTON, G., et SINGH, R. (1993). Imputation pour la non-réponse de vague dans l'enquête Survey of Income and Program Participation (SIPP). *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 115-126.

MAGNUSSON, D., et BERGMAN, L.R. (Éds.) (1990). *Data Quality in Longitudinal Research*. New York: Cambridge University Press.

MARKUS, G.B. (1979). *Analyzing Panel Data*. Beverly Hills, CA: Sage Publications.

MICHAUD, S., et HUNTER, L. (1993). Stratégie pour minimiser l'impact de la non-réponse dans l'enquête sur la dynamique du travail et du revenu. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 103-114.

MOONEY, H.W. (1962). *Methodology in Two California Health Surveys*. Public Health Monograph No. 70, Washington DC: U.S. Department of Health, Education, and Welfare.

MURRAY, T.S., MICHAUD, S., EGAN, M., et LEMAÎTRE, G. (1991). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census. Washington DC: U.S. Department of Commerce, 715-730.

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.

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*, 566-571.

RODGERS, W.L. (1989). Comparisons of alternative approaches to the estimation of simple causal models from panel data. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 432-456.

SINGH, R.P. (1993). Expériences méthodologiques relatives à l'enquête Survey of Income and Program Participation. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 177-188.

SINGH, R., HUGGINS, V., et KASPRZYK, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys*. Document de travail du SIPP n° 9009, Bureau of the Census, Washington DC: U.S. Department of Commerce.

SKINNER, C.J. (1993). Modélisation logistique de données d'enquête longitudinales pouvant comporter une erreur de mesure. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales, Statistique Canada*, 301-309.

SUBCOMMITTEE ON FEDERAL LONGITUDINAL SURVEYS (1986). *Federal Longitudinal Surveys*. Statistical Policy Working Paper 13. Washington DC: Office of Management and Budget.

SUNTER, A.B. (1986). Implicit longitudinal files: A useful technique. *Journal of Official Statistics*, 2, 161-168.

TRAUGOTT, M., et KATOSH, K. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly*, 43, 359-377.

U.S. BUREAU OF THE CENSUS (1978). *The Current Population Survey Design and Methodology*. Bureau of the Census Technical Paper No. 40, Washington DC: U.S. Government Printing Office.

VAN DE POL, F.J.R. (1989). *Issues of Design and Analysis of Panels*. Amsterdam: Sociometric Research Foundation.

VEROFF, J., HATCHETT, S., et DOUVAN, E. (1992). Consequences of participating in a longitudinal study of marriage. *Public Opinion Quarterly*, 56, 315-327.

WALL, W.D., et WILLIAMS, H.L. (1970). *Longitudinal Studies and the Social Sciences*. London: Heinemann.

WATERTON, J., et LIEVESLEY, D. (1989). Evidence of conditioning effects in the British Social Attitudes Panel. *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley, 319-339.

## REMERCIEMENTS

*Techniques d'enquête* désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article, souvent plus d'une fois, durant l'année 1993:

C.H. Alexander, *U.S. Bureau of the Census*
M.G. Arellano, *Advanced Linkage Technologies of America*
J. Armstrong, *Statistique Canada*
T.S. Arthanari, *Indian Statistical Institute*
M. Bankier, *Statistique Canada*
Y. Beaucage, *Statistique Canada*
T.R. Belin, *University of California - Los Angeles*
W. Bell, *U.S. Bureau of the Census*
D. Bellhouse, *University of Western Ontario*
E. Berumen, *IASI*
J. Bethel, *Westat*
D.A. Binder, *Statistique Canada*
G.J. Brackstone, *Statistique Canada*
J.M. Brick, *Westat*
P.A. Chollette, *Statistique Canada*
G.H. Choudhry, *Statistique Canada*
A. Chu, *Westat*
M.J. Colledge, *Statistique Canada*
L. Cox, *U.S. Environmental Protection Agency*
J.-C. Deville, *INSEE*
P. Dick, *Statistique Canada*
J.D. Drew, *Statistique Canada*
R.E. Fay, *U.S. Bureau of the Census*
G. Forsman, *University of Linkoping*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistique Canada*
M.E. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
M.A. Hidiroglou, *Statistique Canada*
H. Hogan, *U.S. Bureau of the Census*
G.J.C. Hole, *Statistique Canada*
D. Holt, *University of Southampton*
D.G. Horvitz, *Retraité*
D.A. Hubble, *U.S. Bureau of the Census*
A.Z. Israëls, *Netherlands Central Bureau of Statistics*
A.E. Johnson, *Agency for Health Care Policy and Research*
G. Kalton, *Westat*
P.S. Kott, *U.S. Department of Agriculture*
P. Lahiri, *University of Nebraska - Lincoln*
D. Lalande, *Statistique Canada*
P. Lavallée, *Statistique Canada*
L. Lebart, *CNRS - Télécom Paris*
H. Lee, *Statistique Canada*
R.J.A. Little, *University of California - Los Angeles*
J. Liu, *Research Triangle Institute*
D. Malec, *National Centers for Health Statistics*
J.T. Massey, *U.S. Department of Health and Human Services*
S.M. Miller, *U.S. Bureau of Labor Statistics*
W.J. Mitofsky, *Voter Research and Surveys*
G. Nathan, *Hebrew University*
H.B. Newcombe, *Expert Conseil*
C.A. Patrick, *Statistique Canada*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
B. Quenneville, *Statistique Canada*
E. Rancourt, *Statistique Canada*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
L. Roy, *Statistique Canada*
D. Royce, *Statistique Canada*
D.B. Rubin, *Harvard University*
K. Rust, *Westat*
I. Sande, *Bell Communications Research*
C.-E. Särndal, *Université de Montréal*
O. Sautory, *INSEE*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *U.S. Internal Revenue Service*
H.T. Schreuder, *U.S. Department of Agriculture*
J. Sedransk, *State University of New York - Albany*
C. Skinner, *University of Southampton*
A. Singh, *Statistique Canada*
T.M.F. Smith, *University of Southampton*
N.L. Spruill, *U.S. Office of the Secretary of Defence*
K.P. Srinath, *Statistique Canada*
C.M. Suchindran, *University of North Carolina - Chapel Hill*
S. Sudman, *University of Illinois - Urbana-Champaign*
J.-L. Tambay, *Statistique Canada*
M.E. Thompson, *University of Waterloo*
A.R. Tupek, *U.S. Bureau of Labor Statistics*
J. Waksberg, *Westat*
W.E. Winkler, *U.S. Bureau of the Census*
K.M. Wolter, *A.C. Nielsen*
J. Wretman, *Stockholm University*
A. Zaslavsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1991: S. Beauchamp et S. Lineger (Photocomposition), et M. Haight (Services de traduction). Finalement on désire exprimer notre reconnaissance à S. DiLoreto, M.M. Kent, C. Larabie et D. Lemire de la Division des méthodes d'enquêtes sociales, pour leur apport à la coordination, la dactylographie et la rédaction.

# Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

## CONTENTS

Volume 42, No. 3, 1993

| | | Page |
|---|---|---|
| Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling | *P. Dellaportas and A. F. M. Smith* | 443 |
| An investigation of changepoints in the annual number of cases of haemolytic uraemic syndrome | *R. Henderson and J. N. S. Matthews* | 461 |
| Estimating annual total heron population counts | *G. E. Thomas* | 473 |
| A bivariate ordered probit model with truncation: helmet use and motorcycle injuries | *A. A. Weiss* | 487 |
| Modelling spatial patterns of trees attacked by bark-beetles | *H. K. Preisler* | 501 |
| Calibration with many variables | *M. C. Denham and P. J. Brown* | 515 |
| Attribute selection in correspondence analysis of incidence matrices | *W. J. Krzanowski* | 529 |
| *General Interest Section* | | |
| Analyses of public use decennial census data with multiply imputed industry and occupation codes | *N. Schenker, D. J. Treiman and L. Weidman* | 545 |
| *Letter to the Editors* | | 557 |
| *Statistical Software Reviews* | | |
| MICLOG; PEST2.1 | | 559 |
| *Statistical Algorithms* | | |
| AS 284 Null distribution of a statistic for testing sphericity and additivity: a Jacobi polynomial expansion | *R. J. Boik* | 567 |
| AS 285 Multivariate normal probabilities of star-shaped regions | *S. L. Lohr* | 576 |
| *Remark* | | |
| AS R91 A remark on Algorithm AS 139: Maximum likelihood estimation in a linear model from confined and censored normal data | *W. D. J. Ryder* | 583 |

*Printed in Great Britain at the Alden Press, Oxford*

# SPECIAL ISSUE
# CONFIDENTIALITY AND DATA ACCESS

## JOS 1993:2

Recent years have seen the unfortunate marriage of two issues of great concern to statistical agencies: confidentiality protection and steadily increasing nonresponse rates. To respond to these growing concerns, the Panel on Confidentiality and Data Access of the Committee on National Statistics and the Journal of Official Statistics have produced this special issue on Confidentiality and Data Access.

### Contents


Preface

Report of the Panel on Confidentiality and Data Access
*George T. Duncan, Virginia A. de Wolf, Thomas B. Jabine, and Miron L. Straf*

Privacy and Advances in Social and Policy Sciences: Balancing Present Costs and Future Gains
*Paul D. Reynolds*

Measures of Disclosure Risk and Harm
*Diane Lambert*

Discussion: *Eleanor Singer, Thomas Plewis, L.W. Cook*

Informed Consent in U.S. Government Surveys
*Robert H. Mugge*

Informed Consent and Survey Response: A Summary of the Empirical Literature
*Eleanor Singer*

Discussion: *Tore Dalenius*

Masking Procedures for Microdata Disclosure Limitation
*Wayne A. Fuller*

Statistical Analysis of Masked Data
*Roderick J.A. Little*

Statistical Disclosure Limitation Practices of United States Statistical Agencies
*Thomas B. Jabine*

Discussion: *Brian V. Greenberg, Donald B. Rubin, Leon Willenborg*

Database Systems: Inferential Security
*Sallie Keller-McNulty and Elizabeth A. Unger*

Discussion: *Martin H. David, Gerald Gates, Teresa F. Lunt, Bo Sundgren*

Confidentiality Legislation and the United States Federal Statistical System
*Joe S. Cecil*

Procedures for Restricted Data Access
*Thomas B. Jabine*

Discussion: *Nancy J. Kirkendall, H.W. Watts, Photis Nanopoulos*

# 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. 10, n° 2) 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, $\omega$; 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.