# TECHNIQUES D'ENQUÊTE

N° 12-001-XPB au catalogue

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

DÉCEMBRE 1998

•

VOLUME 24

•

NUMÉRO 2

Statistique Canada Statistics Canada

Canada

# TECHNIQUES D'ENQUÊTE

## UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

DÉCEMBRE 1998 • VOLUME 24 • NUMÉRO 2

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, 1999

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

Janvier 1999

N° 12-001-XPB au catalogue

Périodicité: semestrielle

ISSN 0714-0045

Ottawa

Statistique Canada Statistics Canada

Canada

# TECHNIQUES D'ENQUÊTE

## Une revue éditée par Statistique Canada

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

**COMITÉ DE DIRECTION**

**Président** G.J. Brackstone

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

**COMITÉ DE RÉDACTION**

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

**Rédacteurs associés**

D.R. Bellhouse, *University of Western Ontario*
D. Binder, *Statistique Canada*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistique Canada*
J. Eltinge, *Texas A&M University*
W.A. Fuller, *Iowa State University*
R.M. Groves, *University of Maryland*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *Central Statistical Office, U.K.*
G. Kalton, *Westat, Inc.*
R. Lachapelle, *Statistique Canada*
P. Lahiri, *University of Nebraska-Lincoln*
S. Linacre, *Australian Bureau of Statistics*
G. Nathan, *Central Bureau of Statistics, Israel*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Bell Communications Research, U.S.A.*
F.J. Scheuren, *Ernst and Young, LLP*
J. Sedransk, *Case Western Reserve University*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
R. Valliant, *Westat, Inc.*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

**Rédacteurs adjoints** P. Dick, H. Mantel, B. Quenneville et D. Stukel, *Statistique Canada*

---

### POLITIQUE DE RÉDACTION

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

#### Présentation de textes pour la revue

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

#### Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 $ par année au Canada et de 47 $ US par année à l'extérieur du Canada. Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au (613) 951-7277 ou au 1 800 700-1033, par télécopieur au (613) 951-1584 ou au 1 800 889-9734 ou par Internet : order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research et la Société Statistique du Canada.

# TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 24, numéro 2, décembre 1998

## TABLE DES MATIÈRES

Dans ce numéro ..... 105

**Enquêtes longitudinales et analyse**
Rédacteurs-coordinateurs Gad Nathan et Christopher Skinner

D.A. BINDER
Les enquêtes longitudinales: Pourquoi ces enquêtes diffèrent-elles de toutes les autres enquêtes? ..... 107

F. BASSI, N. TORELLI et U. TRIVELLATO
Stratégies de collecte de données et de modélisation dans l'estimation de flux bruts relatifs à la population active entachés d'erreurs de classification ..... 117

P.S. CLARKE et R.L. CHAMBERS
Estimation des flux bruts de la population active provenant d'enquêtes donnant lieu à une non-réponse dont il faut tenir compte au niveau du ménage ..... 133

P.-A. SALAMIN
Analyse longitudinale de données de l'enquête suisse sur la population active par régression logistique multidimensionnelle ..... 141

A.H. DORFMAN
Les enquêtes sur les indices de prix en tant qu'études quasi-longitudinales ..... 151

J.-L. TAMBAY, I. ŞCHIOPU-KRATINA, J. MAYDA, D. STUKEL et S. NADON
Traitement de la non-réponse du cycle deux de l'enquête nationale sur la santé de la population ..... 159

M.D. SINCLAIR et J.L. GASTWIRTH
Estimations des erreurs de classification dans l'enquête sur la population active et analyse de leur incidence sur les taux de chômage publiés ..... 171

R.H. RENSSEN
Utilisation de méthodes d'appariement statistique dans l'estimation de calage ..... 185

R. ARNAB
Échantillonnage en deux cycles: Estimation de la population totale ..... 201

E.L. KORN et B.I. GRAUBARD
Intervalles de confiance pour les proportions à petit nombre d'événements positifs prévus estimées au moyen des données d'enquête ..... 209

Remerciements ..... 219

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 105-106
Statistique Canada



# Dans ce numéro

Le présent numéro de *Techniques d'enquête* commence par une section spéciale intitulée «Enquêtes longitudinales et analyse» qui contient six des communications présentées lors de la Réunion satellite de l'AISE-AISO sur les études longitudinales qui s'est tenue à Jérusalem en 1997. Une ou deux communications présentées lors de cette conférence, qui n'ont pu être prêtes à temps pour être publiées dans ce numéro, le seront peut-être ultérieurement. Je suis très reconnaissant à Gad Nathan et à Christopher Skinner d'avoir agi à titre de rédacteurs coordonnateurs pour cette section spéciale. Sans leur ténacité et leur ardeur au travail, ce projet n'aurait jamais vu le jour.

La première communication de la section spéciale, signée par Binder, introduit le sujet en passant en revue l'état actuel de la situation et les défis qui se rattachent aux enquêtes longitudinales par comparaison aux enquêtes transversales. La discussion se divise en quatre parties, examinant tour à tour les problèmes particuliers et les défis qui se posent lors de la conception, de la mise en oeuvre, de l'évaluation et de l'analyse des enquêtes longitudinales.

Bassi, Torelli et Trivellato étudient le problème de l'estimation des flux bruts entre les situations d'activité lorsque les données sont entachées d'erreurs de classification. Ils examinent d'abord diverses stratégies pour recueillir des données longitudinales sur la situation par rapport au marché du travail, et les possibilités qu'elles pourraient offrir en ce qui concerne les erreurs de classification. Ensuite, ils présentent un cadre général de modélisation et un modèle «LISREL modifié» visant à ajuster les estimations de flux bruts afin de corriger les erreurs de classification. Les méthodes sont illustrées au moyen de deux études de cas utilisant des données tirées de la U.S. Survey of Income and Program Participation et de l'Enquête française sur la population active.

Clarke et Chambers étudient l'incidence de la non-réponse dont il faut tenir compte au niveau du ménage. Ils proposent une catégorie de modèles pour la non-réponse dont il faut tenir compte au niveau du ménage. Ils utilisent ensuite des simulations afin de démontrer que les estimations des flux bruts de la population active peuvent être biaisées en présence de cette non-réponse dont il faut tenir compte au niveau du ménage, et que les estimations en fonction de modèles de non-réponse au niveau du ménage peuvent permettre de réduire ce biais. Lorsque le mécanisme de non-réponse au niveau du ménage est correctement spécifié, la source de biais est complètement éliminée; toutefois, même les modèles de non-réponse au niveau du ménage incorrectement spécifiés peuvent réduire ce biais.

Salamin examine le problème de l'estimation d'une variation proportionnelle pour une petite région. Il illustre comment un modèle de régression logistique multidimensionnelle général peut être utilisé pour décrire les données longitudinales obtenues à partir d'un plan d'échantillonnage à renouvellement de panel. Il étudie aussi comment les paramètres de ce modèle peuvent être limités afin de décrire divers types de dépendance entre les observations répétées, ce qui entraîne d'autres estimations des variations fondées sur le modèle. La méthode est illustrée par l'estimation de variations dans la probabilité de trouver du travail dans un canton suisse au moyen des données tirées de l'Enquête suisse sur la population active (ESPA). Par comparaison à de simples différences dans les proportions estimées de personnes occupant un emploi, les estimations fondées sur le modèle comportent de plus petites erreurs types.

Dorfman, dans sa communication, tente de traiter les indices de prix à la consommation d'un point de vue statistique. Il commence par passer en revue la théorie des indices de prix en général, décrivant la méthode stochastique et les objections qui sont formulées à son endroit. Ensuite, il propose une modification de cette méthode stochastique en fonction d'une modélisation à espace d'états qui contourne la majorité des critiques formulées contre celle-ci. La méthode est illustrée au moyen de données sur les prix et les quantités du thon en boîte.

Dans la dernière communication de la section spéciale, Tambay, Şchiopu-Kratina, Mayda, Stukel et Nadon décrivent le traitement de la non-réponse dans l'Enquête nationale sur la santé de la population. Les données recueillies au cours du premier cycle de l'enquête sont considérées comme des prédicteurs éventuels de la non-réponse au cours du cycle deux. Un algorithme DAICH (détection automatique d'interactions du chi carré) est utilisé afin de déterminer les classes de pondération pour la correction de la non-réponse au cours du cycle deux. La communication décrit aussi brièvement le plan d'échantillonnage suivi et les autres étapes du calcul des poids d'estimation.

Sinclair et Gastwirth étudient le problème des erreurs de classification en ce qui concerne les situations d'activité dans la Current Population Survey de la U.S. Bureau of the Census. À cette fin, ils adaptent la méthode mise au point par Hui et Walter, qui convient bien aux données dichotomiques utilisant des données de réinterview, à des données trichotomiques. Contrairement aux autres méthodes, celle-ci ne présuppose pas que les données de réinterview sont dépourvues d'erreurs, mais part plutôt de l'hypothèse qu'il existe une erreur à la fois dans les données de l'interview initiale et dans celles de la réinterview. Les auteurs font une estimation empirique en comparant les taux d'erreurs estimés générés par leur méthode à ceux des autres méthodes existantes telles que celle de Poterba et Summers, et arrivent à la conclusion que le degré de sous-estimation de l'erreur tend à être plus élevé lorsque le taux véritable de chômage est effectivement élevé. Finalement, plutôt que de supposer un taux d'erreur constant d'un bout à l'autre, ils tentent d'effectuer une analyse qui présume que les taux d'erreur ne sont constants qu'au sein de regroupements temporels associés à divers niveaux de chômage.

Renssen examine la difficulté de combiner des informations sur des variables tirées de deux grandes enquêtes distinctes, au moyen des informations auxiliaires provenant d'une troisième enquête de moindre importance qui recueille toutes les variables. En se fondant sur des concepts comme l'appariement statistique et l'étalonnage, il propose des méthodes pour la production de tableaux à double entrée, la production de fichiers de microdonnées et l'estimation des corrélations. En ce qui concerne la production des tableaux à double entrée, sa démonstration conduit à l'étude de deux ensembles différents de contraintes liées à l'étalonnage, l'un appelé stratification incomplète à double entrée et le deuxième, stratification synthétique à double entrée. Lors d'une étude de simulation utilisant les données tirées d'une étude pilote effectuée pour l'Enquête-ménage sur les conditions de vie aux Pays Bas, le calage fondé sur la stratification synthétique à double entrée se révèle bien supérieure.

Arnab évalue différentes stratégies d'échantillonnage en deux occasions. L'échantillon utilisé lors de la deuxième occasion est une combinaison d'un sous-échantillon du premier et d'un nouvel échantillon non apparié. Diverses stratégies de sous-échantillonnage du premier échantillon et d'estimation d'un total lors de la deuxième occasion sont mises en comparaison. Il examine les stratégies existantes décrites dans la documentation et en propose deux nouvelles. L'efficience des diverses stratégies est comparée à la fois analytiquement et empiriquement.

Finalement, Korn et Graubard examinent le problème de la production d'intervalles de confiance pour les proportions comportant un dénombrement positif prévu faible. Signalant au passage que les intervalles binomiaux de Clopper-Pearson utilisés traditionnellement dans les contextes hors enquête sont inappropriés lorsqu'il est question de les utiliser avec des données d'enquête complexes, ils proposent une modification de ces intervalles. Par des simulations, ils comparent ensuite les intervalles proposés aux autres qui sont communément utilisés, dont les intervalles de transformation logit, les intervalles de Breeze (1990) fondés sur une approximation de Poisson et des intervalles linéaires fondés sur la normalité. Ils illustrent aussi les trois autres méthodes proposées avec des applications utilisant des données tirées de la National Health and Nutrition Examination Survey et la Hispanic Health and Nutrition Examination Survey.

Le rédacteur en chef

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 107-115
Statistique Canada



# Les enquêtes longitudinales: Pourquoi ces enquêtes diffèrent-elles de toutes les autres enquêtes?

DAVID A. BINDER[1]

RÉSUMÉ

Nous examinons dans cet article divers aspects de la conception et de l'analyse des études dans le cadre desquelles les mêmes unités sont examinées à différents moments, au fil des ans. Ces études incluent les enquêtes longitudinales, ainsi que les analyses longitudinales d'études rétrospectives et de données administratives ou données du recensement. Nous nous intéressons ici tout particulièrement aux problèmes particuliers qui résultent du caractère longitudinal de l'étude, en examinant quatre des principales composantes de l'étude longitudinale, à savoir la conception, la mise en oeuvre, l'évaluation et l'analyse. Une attention particulière doit être portée à chacune de ces étapes durant la planification d'une étude longitudinale. Au nombre des questions qui sont liées à la nature longitudinale des études, mentionnons les suivantes: concepts et définitions, bases de sondage, méthodes d'échantillonnage, collecte des données, traitement de la non-réponse, imputation, estimation, validation des données ainsi que analyse et diffusion des données. En présumant que les exigences fondamentales liées à la conduite d'une enquête transversale sont connues, nous énonçons ci-après les questions et les problèmes qui se dessinent pour bon nombre d'études longitudinales.

MOTS CLÉS: Bases de sondage; données administratives; collecte des données; non-réponse; imputation; estimation; analyse des données.

## 1. RAISONS JUSTIFIANT LA CONDUITE D'UNE ÉTUDE LONGITUDINALE

Chaque année, des milliers d'enquêtes sont menées par des organismes statistiques à travers le monde. Ces enquêtes visent habituellement à recueillir de l'information pour étayer la prise de décisions ou l'élaboration de politiques. Ces enquêtes ne sont pas menées uniquement à des fins historiques, mais aussi pour recueillir de l'information sur les mesures que l'on pourrait prendre pour faciliter la mise en oeuvre de divers changements de politiques. La plupart des enquêtes sont basées sur des données transversales, c'est-à-dire sur des données obtenues dans le cadre d'une enquête réalisée auprès d'une population donnée, à un moment précis. Des sommaires sont également établis sur la population à l'étude, au moment de l'enquête. Cependant, il arrive très souvent que le but d'une enquête soit, non pas de brosser le tableau de la situation réelle au moment de l'enquête, mais plutôt de connaître les répercussions de divers changements envisagés. Ou encore, il se peut que l'on veuille surveiller les effets d'une modification de politique dont l'entrée en vigueur est prévue prochainement. Le facteur temps est ici l'élément déterminant. Lorsqu'on cherche à obtenir de l'information sur certains phénomènes, par exemple sur l'état de santé ou le niveau de scolarité d'une population, il faut alors examiner les divers déterminants de ces phénomènes. Cependant, il arrive parfois que la relation temporelle elle-même ne soit pas claire, c'est-à-dire quant aux causes ayant précédé les effets. Ces causes pourraient être mesurées si, plutôt que de faire une enquête transversale, les enquêtes étaient échelonnées sur une certaine période, sous forme d'une série d'enquêtes transversales ou encore en utilisant le même panel de répondants, d'un cycle d'enquête à un autre. C'est cette notion basée sur le bon sens qui est à l'origine de la popularité croissante des études longitudinales. Ces dernières offrent en outre l'avantage de réduire l'importance des effets des variables non observées, lorsque les différences dans le temps sont définies en comparant les données recueillies auprès des mêmes répondants.

La prolifération des études longitudinales s'explique également du fait que les sources de données administratives peuvent aujourd'hui être utilisées plus efficacement, ce qui rend faisables certaines études longitudinales. De plus en plus de données administratives sont disponibles. Ces données sont souvent recueillies automatiquement auprès des mêmes personnes, durant une période de temps. Par conséquent, même si les données obtenues de sources administratives ne sont pas idéales, elles peuvent néanmoins constituer une bonne source d'information substitutive pour les enquêtes.

L'avantage de concevoir une étude longitudinale vient de ce que la même méthodologie peut être utilisée pour chacun des cycles de l'enquête, ce qui en retour peut accroître la validité des conclusions qui en seront tirées. Souvent, la conduite d'enquêtes répétées auprès des mêmes répondants s'avère le meilleur moyen de comprendre les profils des changements économiques et sociaux. Il existe une autre approche, moins souhaitable mais néanmoins susceptible de donner des résultats satisfaisants, qui consiste à répéter l'enquête sans nécessairement faire appel aux mêmes répondants, ce qui peut réduire les coûts. L'élément

[1] David A. Binder, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Immeuble R.H. Coats, Parc Tunney, Ottawa, (Ontario), Canada, K1A 0T6.

principal à retenir est le suivant: lorsqu'on cherche à comprendre certains phénomènes dans le temps, il devient alors nécessaire de procéder à une collecte de données de façons répétées.

Un certain nombre de facteurs liés aux coûts doivent être considérés avant de décider de la nature d'une nouvelle étude longitudinale. Bien sûr, il faut évaluer les avantages en regard de ces divers coûts. Les enquêtes longitudinales offrent la possibilité d'étudier des questions dans de multiples domaines, dont quelques-uns sont examinés ci-après. Dans le domaine des soins de santé, par exemple, il est notamment intéressant d'étudier l'évolution de l'état de santé et les déterminants à l'origine de ces changements. En d'autres mots, quels sont les facteurs de risque et quel est, en fait, l'incidence de ces risques sur l'état de santé à long terme? En recueillant des données auprès des mêmes personnes sur une période de temps donnée, il devient possible d'évaluer ces facteurs, non seulement dans le cadre d'études à petite échelle, typiques des essais cliniques, mais également d'enquêtes nationales à grande échelle sur la santé de la population. Il faut reconnaître toutefois que le type d'information recueilli dans le cadre d'une enquête longitudinale nationale sera fort différent de l'information obtenue durant un essai clinique.

Un autre domaine où il est intéressant d'obtenir des observations dans le temps est celui du travail et du revenu. Il ne suffit pas, par exemple, de connaître le changement net, dans le temps, de la situation d'activité ou du taux d'activité de la main-d'oeuvre. Il est bon également de savoir quelles personnes sont passées, disons de la situation de sans emploi à celle de personne active ou inactive. Les profils d'emploi ont changé ces dernières années. On retrouve aujourd'hui davantage de femmes sur le marché du travail et aussi davantage d'emplois à temps partiel. Les changements d'emploi sont également plus fréquents. Les enquêtes longitudinales peuvent fournir des réponses à bon nombre de questions qui permettent en retour de mieux comprendre ces phénomènes. Il peut être intéressant, par exemple, d'étudier les caractéristiques des postes de débutants occupés par les personnes auparavant en chômage, ou encore l'efficacité de différentes stratégies de recherche d'emplois ou des programmes gouvernementaux de formation.

La durée des épisodes de pauvreté est une autre question qui retient de plus en plus l'attention afin, par exemple, de déterminer combien de temps une personne à faible revenu demeure dans cette situation. Quels sont les facteurs qui détermineront si cette situation sera à long terme? Quelle est l'importance de la scolarité et d'autres facteurs en ce qui a trait à la pauvreté et à la durée des épisodes de pauvreté?

Dans le domaine de l'éducation, maintenant, il est intéressant d'examiner la transition de l'école au milieu du travail, au moment de l'entrée sur le marché du travail, et l'étude longitudinale peut convenir davantage que les autres études à l'examen de ce phénomène. Un autre exemple relié aux études est celui de l'efficacité des divers programmes d'études, comme les programmes de formation professionnelle ou de formation des adultes.

Dans le domaine de la justice et de la victimisation, il existe de nombreux exemples où l'observation des mêmes personnes dans le temps peut être bénéfique. Que l'on pense par exemple au suivi des personnes qui ont été victimes d'un acte de violence, dans le but d'évaluer les répercussions de cet événement à long terme. De même, il peut être utile d'observer les personnes qui ont eu des démêlés avec la justice, pour évaluer leurs profils de comportement subséquents et les déterminants de ces profils.

Les études sur les comportements des consommateurs présentent un grand intérêt pour les commerçants et autres intervenants, notamment celles portant sur les habitudes d'achat des consommateurs. Bon nombre de chercheurs trouveraient ainsi fort utile d'avoir un historique des achats des consommateurs.

Pour les décideurs, les études sur les effets de la fluctuation des paiements de transfert aux particuliers, au fil des ans, peuvent s'avérer très utiles. Or, une étude longitudinale peut déterminer notamment pendant combien de temps la personne dépendra de ces paiements, si certains de ces paiements créent ou non une dépendance, quelles sont les caractéristiques des bénéficiaires et quels sont les effets à long terme d'une participation à divers programmes d'aide.

En matière d'économie, les caractéristiques longitudinales de diverses entreprises présentent un grand intérêt. Elles permettent entre autres de mesurer l'efficacité d'une entreprise, son degré d'utilisation de la technologie, les effets à long terme de cette utilisation et l'évolution de la productivité au fil des ans. Diverses questions éclairantes sur les caractéristiques démographiques des entreprises pourraient être posées, par exemple quelles sont les caractéristiques qui font qu'une entreprise est vouée à l'échec ou quelles sont les conditions économiques favorables à la création d'entreprises. Il est également utile d'examiner les conditions propices à la fusion des entreprises. Ce sont là tous des phénomènes dont la mesure peut être facilitée par des études longitudinales.

Au cours des dernières années, de nombreuses entreprises ont subi des changements structurels et seules les études longitudinales permettent d'observer certains de ces changements à un micro-niveau. En effet, un grand nombre de mesures ne peuvent être estimées que lorsque les répondants sont interrogés plus d'une fois.

L'agriculture est un autre domaine qui suscite de l'intérêt, en raison de l'évolution constante des pratiques et du profil agricoles. Il est ainsi pertinent d'examiner l'évolution des exploitations agricoles, tant en termes de produits cultivés que de la taille des entreprises, ou encore de savoir dans quelle mesure les caractéristiques des exploitants agricoles ont changé.

Comme nous venons de l'indiquer, les études longitudinales offrent de multiples applications et comportent de nombreuses facettes. Il existe également de nombreuses

dimensions à leur conception et leur analyse. Dans les quatre sections qui suivent, nous résumons ces questions en regard de quatre volets distincts, soit: la conception, la mise en oeuvre, l'évaluation et l'analyse. Bon nombre de ces questions ont été discutées dans Kasprzyk, Duncan, Kalton et Singh (1989) et Armstrong, Darcovich et Lavallée (1993). Certaines questions liées à la conception et aux séries chronologiques sont traitées dans Binder et Hidiroglou (1988); quelques références plus récentes sont également citées.

## QUESTION 1: CONCEPTION

La planification préliminaire, durant l'étape de conception, est essentielle au succès de l'étude longitudinale. Il faut en effet s'assurer que seules des informations pertinentes et précises seront recueillies auprès des répondants, afin de maximiser les avantages potentiels de l'étude longitudinale. Ceci suppose qu'il faut planifier dès le départ les analyses qui seront effectuées à partir de l'enquête longitudinale, afin d'assurer la collecte des données nécessaires à ces analyses. Duncan et Kalton (1987) présentent un excellent résumé de bon nombre de ces questions. Webber (1994), pour sa part, décrit la stratégie d'essai utilisée pour la planification de la Survey on Labour and Income Dynamics aux États-Unis. Toujours aux États-Unis, Huggins et Fischer (1994) discutent des plans en vue du remaniement de la Survey of Income and Program Participation, en se basant sur leurs propres expériences. Les études longitudinales peuvent être plus coûteuses qu'une série d'études transversales. Aussi les avantages découlant de la collecte de données longitudinales doivent-ils être encore plus grands, les coûts étant eux-mêmes plus élevés. Il est en outre essentiel de s'assurer du maintien du financement de l'étude longitudinale, car il faut parfois attendre le deuxième ou le troisième cycle de l'enquête avant d'obtenir des résultats concrets. Bien sûr, on ne peut planifier de la même façon une étude longitudinale et une série d'enquêtes transversales dont les données seront combinées pour créer une base de données longitudinales. De toute évidence, la première est préférable; cependant, il arrive souvent, en raison de la structure de l'organisme de sondage, que l'on possède déjà une série de données transversales et donc que le regroupement de ces données offre une solution de rechange acceptable (voir Hughes et Hinkins 1995).

En général, une attention particulière doit être portée à la conception de la base de données, pour toute enquête longitudinale prévoyant des analyses basées sur des mesures longitudinales, comme l'étude d'épisodes ou de périodes. Certains organismes statistiques sont actuellement à faire la transition, des enquêtes transversales aux enquêtes longitudinales. Une telle transition nécessite cependant une bonne planification. À titre d'exemple, la mise à jour des bases de données servant aux enquêtes longitudinales diffère considérablement des méthodes de mise à jour utilisées pour les enquêtes transversales et il est à prévoir que la conduite d'un nombre de plus en plus grand d'enquêtes longitudinales mettra en lumière de nombreuses questions liées aux infrastructures et à l'organisation, en particulier en ce qui a trait à la mise à jour des bases de données et aux opérations. Ces changements peuvent en outre avoir une grande incidence sur l'organisme statistique.

Durant la planification d'une enquête longitudinale, il est important de déterminer si les utilisateurs auront ou non besoin aussi d'estimations transversales. Veut-on obtenir des renseignements sur les répondants qui participent à l'enquête durant une période donnée et aussi être en mesure de produire des estimations à un moment précis, comme s'il s'agissait d'une enquête transversale? Le cas échéant, il faudra en tenir compte dans la conception et la mise en oeuvre de l'enquête (voir Lavallée 1995). Une telle préoccupation doit également être prise en considération, si les variables à l'étude requièrent la comparaison d'estimations transversales dans le temps, plutôt que de véritables mesures longitudinales comme l'étude des autocorrélations entre des unités communes dans le cadre d'une enquête auprès des entreprises.

Les concepts et les définitions qui sous-tendent les enquêtes longitudinales sont habituellement établis de concert avec les utilisateurs des données. Même la définition de l'unité longitudinale à observer dans le temps peut devoir être clarifiée pour des populations dynamiques, par exemple pour les enquêtes auprès des ménages ou des entreprises. Il est important de comprendre les exigences de l'utilisateur et de discuter de ce qui peut être mesuré dans le temps, tout en maintenant un niveau de qualité adéquat. Durant la planification de l'enquête, ces exigences doivent être examinées avec soin en regard de ce qui est faisable, au plan opérationnel, durant la réalisation de l'enquête proprement dite. Étant donné les coûts potentiels de ces études, il est souvent avantageux de mener au préalable un essai complet, en particulier des questionnaires. Une autre question qui mérite d'être étudiée plus à fond est celle de l'établissement d'un plus grand nombre de mesures longitudinales types, communes entre les pays, ce qui permettrait aux gouvernements et aux chercheurs d'établir de meilleures comparaisons internationales.

Un autre volet important de la conception des études longitudinales concerne la création, l'utilisation et la mise à jour des bases de sondage au fil des ans, de manière à faciliter la mise en oeuvre de l'étude. À titre d'exemple, il peut arriver qu'une enquête par panel sur les établissements soit basée sur un registre des entreprises dont la composition évolue sans cesse sous l'effet des additions, des retraits et des fusions. Dans de telles circonstances, il est important de bien définir les unités à inclure dans ces panels, au fil des ans.

La fréquence accrue des enquêtes longitudinales au cours des dernières années s'explique notamment par le

plus grand nombre de fichiers administratifs pouvant servir de bases de sondage pour les études longitudinales. Les fichiers administratifs peuvent également contenir des données utiles, en plus de servir de bases de sondage. Cependant, ces données doivent habituellement faire l'objet de certaines manipulations avant d'être utilisées à des fins statistiques pour une étude longitudinale. En général, il faut étudier avec soin, dès l'étape de la conception, l'incidence qu'auront sur l'étude les modifications apportées à la base de sondage.

Une pratique courante consiste à prendre un certain nombre de fichiers administratifs différents et à les apparier pour créer une base de sondage. D'autres études longitudinales sont basées uniquement sur l'information contenue dans divers fichiers administratifs. Manifestement, la difficulté tient au fait que ces fichiers administratifs évoluent au fil des ans et donc qu'il devient nécessaire de modifier les échantillons prélevés de ces fichiers et aussi d'adopter des mesures spéciales pour s'assurer que les analyses demeurent pertinentes.

L'étude longitudinale est souvent basée sur une enquête existante ou sur un recensement mené antérieurement, qui sert ensuite de fondement à la base de sondage utilisée pour le suivi des répondants au fil des ans. Un des inconvénients d'une telle approche est qu'il devient difficile d'obtenir des estimations transversales lorsque les additions dans la population sont exclues de la base de sondage. Il faut parfois avoir recours à des techniques de couplage des enregistrements pour la mise à jour de la base de sondage – une technique habituellement sujette à l'erreur.

Dans de rares cas, il est souvent avantageux d'utiliser non seulement une, mais plusieurs bases de sondage. Une telle méthode assure une représentation adéquate des populations susceptibles d'être sous-représentées dans une base de sondage unique; cependant, elle peut également nécessiter l'utilisation d'une méthode de couplage des enregistrements et de techniques de pondération complexes.

La méthode d'échantillonnage à utiliser, une fois la base de sondage constituée, est un autre aspect important à considérer durant la phase de conception. Kalton et Citro (1993) traitent d'un certain nombre d'enquêtes longitudinales différentes, notamment des enquêtes à passages répétés, c'est-à-dire d'une série d'enquêtes transversales; des enquêtes par panel, où certains répondants sont sélectionnés et suivis dans le temps; des enquêtes par panel à passages répétés, pour lesquelles de nouveaux panels sont sélectionnés à différents moments, durant une période donnée; des enquêtes par panel avec renouvellement, où il y a, à chaque cycle, suppression d'un panel et addition d'un nouveau panel; des enquêtes avec échantillons chevauchants, pour lesquelles de mêmes répondants se retrouvent d'un cycle à un autre sans qu'il s'agisse pour autant d'un plan d'échantillonnage par panel fixe ainsi que des enquêtes par panel fractionné, où il peut y avoir combinaison d'enquêtes par panel et d'enquêtes par panel à passages répétés ou avec renouvellement. Le plan d'échantillonnage doit assurer le prélèvement d'un échantillon suffisant auprès de la population à l'étude, ainsi que de tout groupe témoin. Les données administratives se sont avérées très utiles pour la conception d'un échantillon pour bon nombre de ces enquêtes, car elles procurent souvent une base de sondage adéquate.

Comme l'a fait remarquer un des examinateurs, un élément clé de la phase de conception est celui de la stratégie à utiliser pour tenir compte de la réduction de l'échantillon sous l'effet de l'attrition, de la non-réponse, du retrait de la population-cible, *etc.* Au nombre des possibilités qui s'offrent, mentionnons l'élargissement de l'échantillon lors des cycles subséquents; une telle stratégie risque toutefois d'altérer la représentativité de la cohorte. Une autre stratégie serait de commencer avec un échantillon plus large et de ne pas remplacer les unités perdues, cette dernière stratégie étant proposée notamment par Singh, Petroni et Allen (1994).

Avant de convenir d'un plan d'échantillonnage particulier, il faut tenir compte des techniques de pondération et d'estimation connexes. Il faut aussi déterminer la périodicité ou la fréquence de l'enquête. De toute évidence, lorsque les variables à l'étude changent souvent, il serait plus souhaitable de mener l'enquête plus fréquemment. D'une part, des enquêtes plus fréquentes augmentent les coûts et le fardeau de réponse; d'autre part, une fréquence réduite peut accroître le biais dû à l'erreur de mémoire. Ces compromis entre les coûts et la qualité sont habituellement difficiles à quantifier.

Très souvent, lorsqu'on a besoin à la fois d'estimations transversales et longitudinales, il faut s'assurer d'avoir des estimations transversales valides pour la sélection des échantillons supplémentaires, ceci du fait qu'il peut y avoir des membres de la population dans les estimations transversales qui ne faisaient pas partie de la base de sondage lors des cycles précédents et qui ne seraient donc pas représentés dans l'échantillon. Czajka (1994) étudie ceci en regard de l'estimation du revenu.

Il est également utile, durant la phase de planification, de prévoir des échantillons pour fins d'évaluation. En effet, divers types de biais peuvent être introduits dans les enquêtes longitudinales, dont certains peuvent survenir du seul fait que la collecte des données, à chaque cycle, se fait auprès des mêmes répondants. Aussi, serait-il bon d'envisager l'addition d'échantillons supplémentaires qui serviraient uniquement à des évaluations visant à mesurer certains de ces effets. Ces échantillons seraient composés d'individus de la population-cible qui ne faisaient pas partie de l'enquête longitudinale. De tels échantillons sont des plus utiles pour évaluer les mesures transversales.

## QUESTION 2: MISE EN OEUVRE

La deuxième question que nous abordons ici est celle de la mise en oeuvre d'une étude longitudinale. La première

étape consiste à choisir le mode de collecte des données, parmi les divers choix qui s'offrent. L'interview assistée par ordinateur est une technique qui se répand de plus en plus et qui élargit le choix des instruments d'enquête pouvant être utilisés. Cette technique facilite entre autres l'utilisation de l'interview dépendante, où le répondant ou l'intervieweur a accès aux réponses des cycles précédents; ceci peut augmenter, ou réduire, certains biais. Hill (1994) évalue cette méthode en regard de la Survey of Income and Program Participation aux États-Unis.

Bien sûr, comme ce sont les mêmes répondants qui sont interrogés à diverses reprises, la question du fardeau de réponse prend alors encore plus d'importance que lors d'une enquête transversale unique. Il faut éviter de surcharger le répondant, car ceci risque d'accroître le taux de refus aux cycles subséquents. Michaud, Dolson, Adams et Renaud (1995) proposent de réduire le fardeau du répondant en faisant un usage accru des données administratives. La réduction de l'attrition due à la non-réponse est un objectif important des enquêtes longitudinales et il y aurait lieu d'envisager l'utilisation d'incitatifs financiers ou autres pour aider à préserver l'intégrité de l'échantillon au fil des ans (voir Lengacher, Sullivan, Couper et Groves 1995). Un autre moyen de réduire l'attrition est de recueillir de l'information pour faciliter les efforts de recherche des répondants et aussi de garder contact avec les répondants au fil des ans; McGuigan, Ellickson, Hays et Bell (1995) examinent diverses méthodes de recherche, de repondération et de modélisation de la sélection de l'échantillon, pour résoudre les problèmes d'attrition.

Dans certaines enquêtes longitudinales, la collecte des données se fait de façon rétrospective, c'est-à-dire que les questions posées font référence à la fois à des événements antérieurs et présents; une telle situation peut donner lieu à un effet dit de lisière. Les changements observés durant les périodes de référence peuvent donc dépendre des périodes pour lesquelles les données ont été obtenues de façon rétrospective.

Les dossiers administratifs peuvent être utiles pour enrichir la base de données et éviter ainsi que toutes les données soient recueillies directement du répondant (voir Michaud et coll. 1995). Bien sûr, ceci peut dépendre de la qualité des données administratives, de leur disponibilité et de la réciprocité entre l'information contenue dans les dossiers administratifs et les variables à l'étude; voir Stearns, Kovar, Hayes et Koch (1996) pour un exemple sur cette question. Cependant, lorsqu'on utilise des données administratives ou des fichiers combinés, ces divers fichiers peuvent comporter des lacunes statistiques et il faut alors se demander comment pallier ce problème.

En général, la modification de la structure de la base de sondage peut causer des problèmes durant l'exécution des analyses longitudinales. Des changements peuvent également survenir en ce qui a trait à certaines caractéristiques clés du répondant; dans le cas par exemple d'un registre des entreprises, si l'information sur la classification de l'industrie change, du fait que les entreprises ont modifié la nature des produits offerts, il est alors important que la base de données reflète ce changement de la classification pour assurer une utilité maximale des analyses longitudinales. Cependant, ceci risque également de compliquer l'analyse.

De nombreux problèmes surviennent, lorsque la base de données est créée en combinant des échantillons prélevés d'une série d'enquêtes individuelles. L'intégration de cette information peut créer des difficultés, du fait que des méthodes différentes ont pu être utilisées pour les diverses enquêtes, ce qui risque de se solder par une variation dans la qualité de l'information, d'une base de données à une autre.

Pour bon nombre d'enquêtes longitudinales, les questions liées au couplage des enregistrements sont importantes. Le couplage est en effet utilisé à différentes étapes du traitement et, dans certains cas, les études longitudinales sont basées uniquement sur ces fichiers couplés. Le couplage des enregistrements est ainsi fréquemment utilisé pour créer et mettre à jour les bases de sondage, entre autres pour lier des fichiers administratifs dans le temps, lier des fichiers administratifs et des bases de sondage ou encore lier des bases de sondage distinctes. Dans le cas par exemple d'enquêtes sur les établissements, il se peut que l'on veuille créer des dossiers longitudinaux composites qui soient basés sur plusieurs enquêtes indépendantes à passages répétés, étant donné que bon nombre des établissements sont interrogés à chaque cycle de l'enquête. Le couplage est souvent utilisé pour déterminer quelles unités correspondent aux mêmes établissements, ainsi que pour déterminer les ajouts à une base de sondage. Bien sûr, les erreurs dues au couplage peuvent être importantes; voir Scheuren et Winkler (1993) à ce sujet.

Dans d'autres cas, aucun répondant véritable n'est suivi au fil des ans. Le couplage des enregistrements sert ici plutôt à créer des populations artificielles, par le biais de l'appariement statistique. Ces populations sont ensuite analysées, comme si elles étaient réelles.

Une autre question liée à la mise en oeuvre est celle du traitement de la non-réponse. On sait que la non-réponse aux enquêtes longitudinales n'est pas entièrement le fruit du hasard et qu'elle tend à varier, selon la sous-population. Une attention particulière doit donc être portée au choix de la méthode d'imputation ou de repondération à utiliser: voir, par exemple, Tambay, Şchiopu-Kratina, Mayda, Stukel et Nadon (1998). Lorsque des données administratives servent de fondement à une étude longitudinale, il risque d'y avoir des données manquantes, d'où la nécessité de prévoir des méthodes spéciales pour y remédier.

L'imputation et la repondération sont les deux méthodes habituellement utilisées pour pallier les données manquantes. La repondération est utilisée habituellement lorsque la non-réponse est propre à un cycle particulier. L'imputation est utilisée plus fréquemment lorsqu'on observe une non-réponse partielle durant un cycle donné de

l'enquête. Il peut y avoir des avantages à utiliser l'imputation pour des études longitudinales plutôt que transversales. Dans le premier cas, l'imputation est basée sur les données longitudinales contenues dans la base de données pour le même répondant, et non sur l'information obtenue des autres répondants, au même moment. Pour résoudre les problèmes d'attrition et de non-réponse cyclique, on peut modéliser les taux d'attrition et utiliser ces modèles pour compenser pour la non-réponse, par l'ajustement des poids. Diverses méthodes d'ajustement des poids ont été examinées pour la Survey of Income and Program Participation et les résultats sont présentés dans Rizzo, Kalton et Brick (1994), Folsom et Witt (1994) et An, Breidt et Fuller (1994). Singh, Wu et Boyer (1995) ont pour leur part étudié ce problème, en regard du cas difficile que présente l'estimation des flux bruts.

Le calcul des poids peut poser de nombreuses difficultés. Il existe diverses approches et techniques pour calculer les poids transversaux et longitudinaux. Les premiers sont utilisés pour les mesures de la population qui font référence à un moment précis dans le temps, alors que les deuxièmes doivent être utilisés lorsque sont incluses des données sur plus d'un cycle. L'analyste peut choisir d'utiliser des poids propres à la personne, qui diffèrent des poids qui s'appliquent à des ménages (voir Kalton et Brick 1995). À titre d'exemple, pour certaines variables comme le revenu familial, il serait préférable d'utiliser des poids qui s'appliquent aux ménages plutôt qu'à la personne. La pondération devient plus complexe lorsqu'il y a utilisation de bases de sondage multiples. L'utilisation efficace de données administratives peut engendrer des complications encore plus grandes dans le plan de pondération lui-même (voir, par exemple, Stearns et coll. 1996).

De nombreux facteurs peuvent rendre un échantillon sous-représentatif. Ainsi, l'absence de représentativité peut être due à des problèmes de couverture résultant d'une immigration dans la population. Dans d'autres cas, le sous-dénombrement peut être attribuable à l'attrition. À l'opposé, le surdénombrement peut résulter de l'inclusion de certains cohabitants non échantillonnés d'un ménage, ce qui suppose que ces personnes sont incluses dans l'échantillon du seul fait qu'elles vivent avec une personne appartenant à l'échantillon initial (voir Lavallée 1995 et Kalton and Brick 1995). D'autres types de surdénombrement général peuvent se produire et des techniques de pondération spéciales doivent être utilisées pour s'assurer qu'aucun biais n'est introduit. Cependant, dans le cas notamment des enquêtes longitudinales, ceci peut devenir assez complexe. Les données administratives peuvent alors être utilisées pour déterminer si, oui ou non, l'échantillon est représentatif et pour fournir de l'information en vue d'apporter les corrections nécessaires.

Comme la majeure partie des estimations, aux fins d'une étude longitudinale, consistent à mesurer des changements plutôt qu'à mesurer un phénomène à un moment précis, ceci soulève des questions quant à la façon de définir les variances pour ces estimations de la mesure du changement et peut donc nécessiter l'adoption de nouvelles procédures. En général, l'estimation de la variance peut devenir assez complexe lorsque les statistiques sont des fonctions complexes des observations longitudinales. À titre d'exemple, les limites des tranches de revenu peuvent changer au fil des ans et il peut être intéressant d'étudier les mouvements d'une personne, d'une tranche de revenu à une autre.

Le désir d'inclure des données obtenues d'enquêtes transversales permanentes pour produire de nouvelles mesures intégrées, en utilisant toute l'information disponible des diverses sources, peut compliquer encore davantage les estimations.

## QUESTION 3: ÉVALUATION

Le troisième aspect dont nous discutons ici est celui de l'évaluation des données et des méthodes. Même si les évaluations peuvent se faire indépendamment de la mise en oeuvre, les résultats de ces évaluations devraient influer sur l'enquête elle-même, soit en modifiant les méthodes d'estimation, soit en modifiant la façon dont l'enquête est conçue et mise en oeuvre durant les cycles subséquents.

Il existe de nombreuses sources de biais qui peuvent être étudiées. Le biais peut être dû à l'utilisation d'une technique d'interview dépendante où le répondant et l'intervieweur reçoivent de l'information faisant référence à un cycle antérieur de l'enquête. L'effet de lisière peut résulter d'études rétrospectives; voir, par exemple Murray, Michaud, Egan et Lemaître (1991). D'autres sources de biais peuvent être introduites, lorsque la non-réponse est informative, c'est-à-dire lorsque la propension à la non-réponse est liée à la variable à l'étude, par exemple lorsque la non-réponse dans le ménage est liée aux flux bruts à l'intérieur du ménage (flux bruts faisant ici référence aux changements dans la classification de la personne): voir Clarke et Chambers (1989). D'autres biais peuvent être dus à des erreurs de mesure ou de classification: voir, par exemple, Bassi, Torelli et Trivellato (1998). Enfin, il peut y avoir introduction d'un biais de conditionnement du fait que le répondant devient plus sensibilisé à certaines questions qui lui sont posées, par exemple sur la dynamique du travail, et donc que son comportement change du seul fait de sa participation à l'enquête.

Il convient également d'évaluer l'effet des erreurs de réponse et des erreurs de l'intervieweur sur l'analyse. À des méthodes d'interview différentes peuvent correspondre des taux d'erreur différents. Le fait qu'il y ait ou non roulement des intervieweurs peut aussi avoir une incidence sur certaines analyses, tout comme l'utilisation ou non de répondants substituts.

D'autres évaluations peuvent être faites pour mesurer l'effet de l'attrition et évaluer diverses techniques d'imputation et autres méthodes de traitement de la non-réponse (voir Tin 1996, pour une évaluation de l'attrition

par des méthodes économétriques). Pour leur part, Schejbal et Lavrakas (1995) étudient l'effet de l'attrition du panel lors d'une enquête téléphonique locale à double base de sondage. Corder, Manton et Woodbury (1994) étudient des méthodes pour améliorer la couverture et réduire l'attrition dans le cadre d'une enquête nationale sur les soins de longue durée (National Long Term Care Survey). L'attrition du panel peut être due à un refus de participer ou à l'impossibilité de rejoindre les personnes; or l'effet de l'attrition peut être très différent selon qu'il s'agit d'une enquête longitudinale ou transversale, et ces différences doivent être étudiées. Allen et Petroni (1994) discutent quant à eux du problème de la correction des données en fonction des personnes qui ont déménagé.

Enfin, il faut concevoir des études de contrôle de la qualité qui tiennent compte des caractéristiques particulières des enquêtes longitudinales. Un grand nombre d'études sur le contrôle de la qualité, outre les études habituelles prévues pour les enquêtes transversales, peuvent être utilisées pour les enquêtes longitudinales, car la répétition de l'enquête peut permettre une identification plus efficace des cas sujets à erreur. Comme la stabilité des données dans le temps est importante pour les études longitudinales, les changements méthodologiques peuvent avoir un impact sur les mesures longitudinales à l'étude et ceux-ci doivent être étudiés. Enfin, les données administratives peuvent fournir des évaluations utiles, certaines de ces données pouvant aider à valider les résultats.

## QUESTION 4: ANALYSE

Le dernier volet de notre examen porte sur les questions liées à l'analyse, et principalement sur l'analyse potentielle de l'étude longitudinale. Les causes ou déterminants des divers résultats présentent un grand intérêt pour les utilisateurs de données. Cependant, la modélisation de ces causes peut être complexe, en particulier si l'enquête elle-même est complexe. Un grand nombre de ces questions sont examinées dans Singh et Whitridge (1990) et dans Hidiroglou et Michaud (1998).

Les mesures des flux bruts et autres mesures du changement brut sont quelques exemples des types d'analyses courantes. Flux brut fait référence au changement, pour une personne, d'une catégorie à une autre. Il s'agit, en d'autres mots, de la transition de la catégorie $A$ à la catégorie $B$ entre deux moments précis, alors que le flux net correspond au résultat net du changement pour un ensemble de personnes, dans le temps. L'incidence d'une erreur dans la mesure du flux brut soulève des questions difficiles. S'il se produit des erreurs de mesure assez importantes à chaque cycle, il en résultera un impact significatif sur le biais des estimations du flux brut, même si cela n'a aucune incidence sur le flux net en soi. Parfois, le renouvellement de l'échantillon aggrave ce problème, car il peut être difficile de tenir compte adéquatement du renouvellement de l'échantillon lorsqu'on mesure les flux bruts. Il faut prévoir un traitement spécial pour les panels qui s'ajoutent à l'échantillon à un cycle donné et pour ceux qui ont été éliminés de l'échantillon au cycle précédent, afin d'obtenir une bonne estimation de ces flux. Pour la mesure des flux bruts, il importe également de différencier les changements dans la population des flux bruts proprement dits. Autrement dit, les changements qui surviennent d'un cycle à un autre sont une combinaison des changements dans la taille de la population et des changements à l'intérieur de la population. La situation peut devenir encore plus complexe lorsque les flux bruts sont eux-mêmes analysés en regard d'autres données, par exemple la dynamique du revenu.

Comme l'a fait remarquer un examinateur, il est important d'éduquer les utilisateurs sur la façon d'analyser efficacement les données longitudinales. La récente prolifération des enquêtes longitudinales fait naître de nombreuses possibilités en termes d'analyses nouvelles; cependant, il est possible que bon nombre des analystes qui n'ont étudié jusqu'ici que les enquêtes transversales ne connaissent pas les techniques qui conviennent le mieux aux enquêtes longitudinales.

Pour les nombreuses enquêtes qui utilisent des bases de sondage s'appuyant sur des données administratives, il peut s'avérer nécessaire de tenir compte des changements dans la base de sondage durant l'analyse, car les inclusions dans la base de sondage peuvent dépendre des changements apportés aux procédures administratives et des changements dans les conditions propres à la personne. La composition d'un fichier des prestataires de l'assurance-chômage, par exemple, évolue en fonction des critères d'admissibilité et de la situation personnelle de l'individu.

La mesure du changement peut souvent être décomposée en diverses composantes. À titre d'exemple, une distinction peut être établie entre les mouvements des unités de l'échantillon, d'un domaine à un autre, et les changements dans les données, pour les unités d'un même domaine. Holt et Skinner (1989) présentent une discussion intéressante sur diverses composantes du changement.

Dans le cas des analyses plus complexes, comme la modélisation des séries chronologiques, la plupart des modèles classiques de séries chronologiques ne prennent pas en considération le fait que l'information provient d'un sondage. Par conséquent, la modélisation des séries chronologiques ne tient pas compte adéquatement des erreurs d'échantillonnage attribuables au type d'enquête.

Par ailleurs, certaines mesures peuvent dépendre d'autres enquêtes transversales. Il se peut, par exemple, que les tranches de revenu utilisées pour l'analyse de l'enquête longitudinale soient tirées d'une autre enquête, celle-ci de nature transversale; or ceci risque d'accroître la complexité de l'analyse, car ces tranches ne sont pas nécessairement stables.

L'utilisation ou non de poids d'échantillonnage et la façon de les utiliser, le cas échéant, créent des difficultés pour nombre d'analystes, car de nombreux modèles clas-

siques utilisés pour l'analyse des données dans le temps ne comportent pas de poids d'échantillonnage. Il faut donc mettre au point des méthodes pour intégrer efficacement les poids d'échantillonnage dans l'analyse. L'utilisation de poids s'avère souvent la méthode à privilégier dans le cas des enquêtes à grande échelle, car ceci aide à éviter l'erreur de spécification du modèle.

Dans certains cas, les erreurs dues au traitement des données, par exemple au couplage des enregistrements, devront être incluses dans l'analyse ou, du moins, faire l'objet d'une étude pour en comprendre l'incidence (voir, par exemple, Dorinski et Huang 1994).

Les données administratives sont souvent utilisées pour l'analyse, car ces données sont parfois plus facilement utilisables que les données recueillies. Cependant, comme les données administratives peuvent présenter des problèmes au niveau conceptuel ou autre, des méthodes d'analyse spéciales sont parfois requises pour assurer un usage efficace de ces données.

Enfin, il y a lieu de mentionner les difficultés associées à la diffusion des données. Pour de nombreux phénomènes, il faut définir des mesures longitudinales sommaires. Or, souvent, ces mesures ne cadrent pas avec les tableaux habituellement utilisés pour les enquêtes transversales. De plus, un grand nombre d'analyses requièrent un accès à des micro-données ce qui, en retour, risque de créer des problèmes en ce qui a trait à la protection de la confidentialité de l'information. En effet, il se peut que les mesures habituelles prises au moment de la diffusion de fichiers de micro-données provenant d'enquêtes transversales ne suffisent pour la diffusion des résultats d'enquêtes de nature longitudinale, du fait que les bases de données longitudinales sont beaucoup plus riches et donc que la possibilité de pouvoir identifier une personne à partir de ces données devient beaucoup plus grande. Comme la protection de la confidentialité du répondant est d'une importance capitale, une approche conservatrice doit parfois être utilisée, même si celle-ci ne répond pas entièrement aux exigences des utilisateurs.

## SOMMAIRE

Nous avons discuté brièvement d'un grand nombre de questions auxquelles s'intéressent actuellement les chercheurs qui étudient la conception et l'analyse des études longitudinales et nous avons constaté que plusieurs d'entre elles devront faire l'objet d'une analyse plus poussée. Des réponses à bon nombre de ces questions apparaîtront à mesure que notre expérience des enquêtes longitudinales augmentera; cependant, cet approfondissement de nos connaissances fera naître également beaucoup de nouvelles questions. Ceci ouvre donc la voie à de nombreuses études et recherches significatives.

## REMERCIEMENTS

L'auteur tient à remercier les examinateurs et le rédacteur adjoint, pour leurs nombreuses suggestions utiles.

## BIBLIOGRAPHIE

ALLEN, T.M., et PETRONI, R.J. (1994). Mover nonresponse adjustment research for the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 662-667.

AN, A.B., BREIDT, F.J., et FULLER, W.A. (1994). Regression weighting methods for SIPP data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 434-439.

ARMSTRONG, J., DARCOVICH, N., et LAVALLÉE, P. (Éds.) (1993). *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales*, Statistique Canada.

BASSI, F., TORELLI, N., et TRIVELLATO, U. (1998). Stratégies de collecte de données et de modélisation dans l'estimation de flux bruts relatifs à la population active entachés d'erreurs de classification. *Techniques d'enquête*, 24, 117-132.

BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics Volume 6: Sampling* (Éds. P.R. Krishnaiah et C.R. Rao). North Holland: Amsterdam, 187- 211.

CLARKE, P.S. et CHAMBERS, R.L. (1998). Estimation des flux bruts de la population active provenant d'enquêtes donnant lieu à une non-réponse dont il faut tenir compte au niveau du ménage. *Techniques d'enquête*, 24, 133-140.

CORDER, L.S., MANTON, K.G., et WOODBURY, M. (1994). Improving coverage, response rates, and nonresponse follow-up via a longitudinal list sample design: The National Long-Term Care Surveys. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 63-84.

CZAJKA, J.L. (1994). Income stratification in panel surveys: Issues in design and estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 791-796.

DORINSKI, S.M., et HUANG, H. (1994). Use of administrative data in SIPP longitudinal estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 656-661.

DUNCAN, G.J., et KALTON, G. (1987). Issues of design and analysis of surveys across time. *Revue Internationale de Statistique*, 55, 97-117.

FOLSOM, R.E., et WITT, M.B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 428-433.

HIDIROGLOU, M.A., et MICHAUD, S. (Éds.) (1998). *Recueil: Symposium 98, Analyses longitudinales pour enquêtes complexes*, Statistique Canada. À paraître.

HILL, D.H. (1994). The relative empirical validity of dependent and independent data collection in a panel survey. *Journal of Official Statistics*, 10, 359-380.

HOLT, D., et SKINNER, C.J. (1989). Components of change in repeated surveys. *Revue Internationale de Statistique*, 57, 1-18.

HUGGINS, V.J., et FISCHER, D.P. (1994). The redesign of the survey of income and program participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 668-673.

HUGHES, S., et HINKINS, S. (1995). Creation of panel data from cross-sectional surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 408-413.

KALTON, G., et BRICK, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 37-51.

KALTON, G., et CITRO, C. (1993). Enquêtes par panel: ajout d'une quatrième dimension. *Techniques d'enquête*, 19, 217-227.

KASPRZYK, D., DUNCAN, G., KALTON, G., et SINGH, M.P. (Éds.) (1989). *Panel Surveys*. Wiley: New York.

LAVALLÉE, P. (1995). Pondération transversales des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-36.

LENGACHER, J.E., SULLIVAN, C.M., COUPER, M.P., et GROVES, R.M. (1995). Once reluctant, always a reluctant? Effects of differential incentives on later survey participation in a longitudinal study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1029-1034.

McGUIGAN, K.A., ELLICKSON, P.L., HAYS, R.D., et BELL, R.M. (1995). Tracking, weighting, and sample selection modeling to correct for attrition. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 402-407.

MICHAUD, S., DOLSON, D., ADAMS, D., et RENAUD, M. (1995). Combining administrative and survey data to reduce respondent burden in longitudinal surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 11-20.

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, 715-730.

RIZZO, L., KALTON, G., et BRICK, M. (1994). Adjusting for panel nonresponse in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 422-427.

SCHEJBAL, J.A., et LAVRAKAS, P.J. (1995). Panel attrition in a dual-frame local area telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1035-1039.

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

SINGH, A.C., et WHITRIDGE, P. (Éds.) (1990). *Analysis of data in time. Recueil: Symposium 89, L'analyse des données dans le temps*, Statistique Canada.

SINGH, A.C., WU, S., et BOYER, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-401.

SINGH, R.P., PETRONI, R.J., et ALLEN, T.M. (1994). Oversampling in panel surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 674-679.

STEARNS, S.C., KOVAR, M.G., HAYES, K., et KOCH, G.G. (1996). Estimates of national hospital use from administrative data and personal interviews. *Journal of Official Statistics*, 12, 47-61.

TAMBAY, J.L., ŞCHIOPU-KRATINA, I., MAYDA, J., STUKEL, D., et NADON, S. (1998). Traitement de la non-réponse du cycle deux de l'enquête nationale sur la santé de la population. *Techniques d'enquête*, 24, 159-169.

TIN, J. (1996). Program participation and attrition: The empirical evidence. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 669-674.

WEBBER, M. (1994). The survey of labour and income dynamics: lessons learned in testing. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 85-99.

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 117-132
Statistique Canada



# Stratégies de collecte de données et de modélisation dans l'estimation de flux bruts relatifs à la population active entachés d'erreurs de classification

FRANCESCA BASSI, NICOLA TORELLI et UGO TRIVELLATO[1]

RÉSUMÉ

Les flux bruts entre situations d'activité sont d'une grande importance pour comprendre la dynamique du marché du travail. Les flux observés sont généralement entachés d'erreurs de classification qui peuvent introduire des biais importants. Dans le présent article, nous passons en revue certaines des stratégies les plus couramment utilisées pour recueillir des données longitudinales sur la situation par rapport au marché du travail, ainsi que des méthodes de modélisation qui ont été créées pour corriger les flux bruts lorsqu'ils présentent des erreurs de classification. Nous présentons un cadre général pour l'estimation des flux bruts. Nous donnons des exemples des diverses spécifications de modèle appliquées à des données qui ont été recueillies avec des stratégies différentes. Nous considérons en particulier deux cas, c'est-à-dire les flux bruts de la U.S. Survey of Income and Program Participation et les flux bruts de l'Enquête française sur la population active, qui est une enquête annuelle dans laquelle on recueille des données mensuelles rétrospectives.

MOTS CLÉS: Erreurs de classification corrélées; modèles de classe latente; données longitudinales; erreurs de mémoire; effet de «lisière».

## 1. INTRODUCTION

Les flux bruts entre situations d'activité représentent un puissant moyen pour analyser la dynamique du marché du travail. Les mouvements bruts ont trait à des variations qui ont lieu au niveau des individus, leur estimation dépend donc de la disponibilité de données longitudinales.

Les effets d'une classification erronée d'unités par rapport à leur situation au sein de la population active peuvent être à l'origine de fausses transitions. Même si l'on peut présupposer que ces erreurs s'annulent lors de l'estimation de flux nets, on ne peut pas ne pas en tenir compte dans l'estimation des flux bruts.

Diverses stratégies permettent de corriger les flux bruts afin de tenir compte d'erreurs de classification. Essentiellement, ces stratégies sont fondées sur

a) des hypothèses concernant l'origine des erreurs de classification, qui peuvent être dues
   a1) au plan d'échantillonnage (enquêtes par panel, éventuellement à renouvellement de panel; enquêtes rétrospectives; des combinaisons d'enquête par panel et d'enquête rétrospective, *etc.*), ou
   a2) au contenu et à la structure du questionnaire (existence d'un ou de plusieurs indicateurs de la variable observée, types de questions (questions fondées sur un épisode ou sur un événement), *etc.*);
b) des hypothèses concernant le processus qui est à l'origine des transitions entre les situations relatives à la population active.

Dans le présent article, nous passons en revue certaines des stratégies les plus couramment utilisées pour recueillir des données longitudinales sur la situation d'activité, ainsi que des méthodes de modélisation qui ont été créées pour corriger des flux bruts lorsqu'ils présentent des erreurs de classification. Nous montrons que la plupart des spécifications habituelles qui sont proposées dans les textes spécialisés peuvent être considérées comme des cas particuliers d'une formulation générale qui permet de déterminer les avantages et les inconvénients de chaque spécification et d'envisager une stratégie d'estimation commune.

Le présent article porte principalement sur des applications sensées de cette méthode de modélisation générale dans l'estimation de flux bruts à partir de données d'enquête recueillies par des stratégies différentes. Nous considérons deux cas en particulier: la U.S. Survey of Income and Program Participation et l'Enquête française sur la population active, qui est une enquête annuelle à renouvellement de panel comportant des informations mensuelles rétrospectives.

L'organisation de l'article est décrite ci-après. Dans la section 2, nous abordons brièvement diverses stratégies de collecte de données longitudinales sur l'activité, ainsi que le rôle probable que jouent ces stratégies dans l'introduction d'erreurs de classification, d'après leur description dans les études qui traitent des méthodes d'enquête. Dans la section 3, nous présentons une méthode assez générale qui permet de modéliser des flux bruts entachés d'erreurs de classification, c'est-à-dire qui permet d'estimer à la fois des flux bruts réels et des probabilités de réponse conditionnelles. Nous donnons également des exemples de la façon dont on peut spécifier comme cas particuliers de la méthode

[1] Francesca Bassi, Nicola Torelli et Ugo Trivellato, Dipartimento di Scienze Statistiche, Via San Francesco, 33, 35121, Padova, Italy.

précitée certains modèles bien connus qui sont utilisés pour la correction de flux bruts observés (section 3.1). Nous présentons ensuite un cadre pratique pour la formulation des modèles susmentionnés; ce cadre est fourni par des modèles de classe latente, et plus précisément par le modèle «LISREL modifié» qui a été proposé par Hagenaars (1990), et qui est un outil général qui permet de décrire de type causal existant entre des variables nominales observées et non observées (section 3.2).

La dernière et principale partie de l'article (section 4) est consacrée à une présentation détaillée de deux études de cas. Nous appliquons la même méthode de modélisation aux deux cas: des informations a priori sur les caractéristiques de mesure de l'enquête (et si possible, sur le processus réel) sont combinées à des recherches de spécifications, afin d'obtenir des modèles parcimonieux et (nous l'espérons) sensés. Comme nous l'avons mentionné précédemment, ces deux études de cas sont raisonnablement différentes, principalement en ce qui a trait au plan d'échantillonnage; cette différence s'avère utile pour illustrer diverses spécifications de modèle et diverses stratégies utilisées pour obtenir et tester la formulation finale.

Concernant les deux études de cas, on constate ce qui suit:

a) le modèle LISREL modifié s'est avéré suffisamment souple pour modéliser le mécanisme à l'origine des erreurs contenues dans les données longitudinales qui ont été recueillies d'après des plans d'échantillonnage différents, ainsi que pour modéliser le processus à l'origine des transitions réelles au sein de la population active;
b) en particulier dans la partie du modèle relative aux mesures, nous avons été en mesure d'intégrer le patron d'erreurs corrélées ainsi que les effets que produisent ces erreurs, qui sont particulièrement importants dans le cas d'enquêtes qui comportent des éléments rétrospectifs;
c) les transitions observées sont corrigées dans le sens prévu d'après des preuves théoriques et empiriques relatives aux effets des erreurs de mesure (et non automatiquement dans le sens de la mobilité, comme c'est le cas avec des stratégies fondées sur l'hypothèse d'erreurs de classification indépendantes).

## 2. LE RÔLE DES STRATÉGIES DE COLLECTE DE DONNÉES

Les informations servant à l'estimation des flux bruts relatifs à la population active proviennent de données longitudinales, c'est-à-dire d'observations portant sur les mêmes unités et effectuées à diverses époques. Depuis peu, on s'efforce davantage de recueillir des données longitudinales. Cela est vrai également dans le cas d'enquêtes dont le but principal est d'évaluer la situation d'individus au sein de la population active, dans le cas d'une population donnée. D'autre part, cet intérêt à l'égard de la collecte et de l'utilisation de données longitudinales soulève de nouvelles questions concernant l'origine et le patron des erreurs de mesure (ou de classification), ainsi qu'au sujet des effets possibles de ces erreurs sur les estimations des quantités qu'on souhaite connaître. Parmi les études de référence générales portant sur les sources d'erreurs de classification touchant des données longitudinales recueillies dans le cadre d'enquêtes au fil du temps, il y a celles de Duncan et Kalton (1987) et de Kalton et Citro (1993). Dans la présente section, nous abordons brièvement certains des principaux effets des erreurs de classification sur les stratégies de modélisation utilisées pour corriger des flux bruts.

Un argument qui est avancé habituellement au sujet de l'incidence des erreurs de mesure dans l'estimation de flux bruts est que ces erreurs mènent à une surestimation des variations. Cela est vrai lorsqu'on présuppose que les erreurs de mesure ne sont pas corrélées au fil du temps. Or, dans de nombreux cas, cette hypothèse n'est pas réaliste (voir Skinner et Torelli 1993; Singh et Rao 1995; van de Pol et Langeheine 1997) et devrait être reconsidérée en tenant compte attentivement de la stratégie de collecte de données qui a été effectivement adoptée. De manière générale, si les données longitudinales sont recueillies par interrogation rétrospective (du moins en partie), on peut affirmer que des erreurs de mémoire entraînent des erreurs corrélées.

Des hypothèses précises concernant les erreurs de classification ne peuvent être introduites avec succès dans des modèles statistiques appropriés que si l'on dispose d'informations supplémentaires sous forme de connaissances a priori plausibles concernant le mécanisme qui est à l'origine des erreurs, ou des données supplémentaires relatives à la situation au sein de la population active.

Les stratégies de modélisation utilisées pour corriger des flux bruts entachés d'erreurs de classification devraient aussi tenir compte du processus de mesure qui est effectivement utilisé, étant donné que le nombre d'erreurs de classification et la direction d'éventuels biais sont liés à la stratégie qui a été adoptée pour recueillir les données longitudinales.

Comme on le sait bien, des données longitudinales peuvent être obtenues à l'aide de diverses stratégies d'enquête. Il est utile à cet égard de faire tout au moins la distinction entre, d'une part, les enquêtes par panel et, d'autre part, les enquêtes rétrospectives. En outre, la disponibilité d'indicateurs multiples mérite une attention particulière.

Les enquêtes par panel représentent la façon la plus naturelle de recueillir des données longitudinales. Parmi ce type d'enquêtes, les enquêtes à renouvellement de panel jouent un rôle important. En effet, il s'agit du type de méthode qui est utilisé pour la plupart des enquêtes sur la population active (EPA), dont le but principal est l'estimation des statistiques de la population active. Dans

le cas d'EPA comportant un plan d'échantillonnage à renouvellement de panel, les données longitudinales sur la séquence (généralement brève) de situations peuvent être obtenues en appariant des données relatives à des individus qui participent à deux ou à plusieurs enquêtes successives. Dans les EPA, la période de référence, ainsi que les notions et les définitions utilisées pour classer les répondants sont généralement conformes aux recommandations du Bureau international du travail (BIT) (Hussmanns, Mehran et Verma 1990); cela donne des mesures des situations d'activité assez précises et comparables entre elles dans l'espace et dans le temps. Les données relatives à l'activité sont recueillies également par l'entremise d'enquêtes-ménages polyvalentes. Dans ces enquêtes, on accorde moins d'attention à la situation par rapport au marché du travail que dans le cas des enquêtes mentionnées précédemment, et les périodes de référence, ainsi que les notions et les définitions peuvent être moins conformes aux recommandations du BIT.

Les données longitudinales peuvent être recueillies également par l'entremise d'enquêtes rétrospectives. Des enquêtes transversales peuvent comprendre des questions rétrospectives qui servent à obtenir des informations sur la séquence de situations d'activité qui sont vécues par les répondants. Dans ce cas, la stratégie d'interrogation a une importance déterminante en ce qui a trait à la réduction des erreurs de mémoire, des télescopages, *etc.* Les méthodes utilisées pour améliorer l'exactitude des déclarations dans les enquêtes rétrospectives sont fondées sur des éléments de la psychologie cognitive et sur les techniques d'enquête (pour une analyse de ces méthodes, voir O'Muircheartaigh 1996). En outre, de nombreuses études empiriques fournissent des preuves du nombre et de l'orientation de biais dus à des erreurs de mémoire. Il est utile d'ajouter à cet égard que dans le cas d'études rétrospectives, des facteurs liés à la durée de la période pour laquelle le répondant doit faire appel à sa mémoire, l'importance des événements qui sont considérés ou la difficulté que comporte la récupération de données relatives à des événements passés mènent généralement à une simplification des questions non conforme aux conventions du BIT relatives aux enquêtes sur la population active.

La pratique répandue qui consiste à utiliser une combinaison de stratégie par panel et de stratégie rétrospective offre des possibilités intéressantes pour estimer des flux bruts de population active en présence d'erreurs de classification. Les enquêtes par panel font appel à des questions rétrospectives, du moins pour un nombre limité de sujets, afin de couvrir la période comprise entre deux vagues successives (c'est le cas notamment de la Survey of Income and Program Participation, comme nous le verrons dans la section 4.2). Lorsqu'on utilise une telle stratégie mixte, les principales caractéristiques du processus de mesure doivent être considérées attentivement, étant donné que celles-ci peuvent avoir une incidence considérable sur la formulation de modèles sensés permettant de tenir compte des erreurs de classification. Des caractéristiques plus spécifiques du processus de mesure découlent également de la prise en considération des particularités du plan d'échantillonnage.

Dans une optique différente, la disponibilité de mesures multiples de la situation par rapport au marché du travail, c'est-à-dire des données sur la situation d'activité d'une personne à une époque donnée qui sont fournies par deux ou plusieurs sources différentes, offre une excellente possibilité de modéliser des erreurs de classification. Ces informations ont une grande importance, d'une façon générale, et en particulier lorsqu'il faut prendre en considération des types d'erreurs de classification corrélées assez compliqués. Des indicateurs multiples de la situation d'activité d'une personne peuvent être recueillis lors d'une même interview, ou lors d'interviews distinctes (par exemple, lors de vagues différentes d'une enquête par panel).

Le premier cas n'est pas très courant, mais parfois, des questions portant sur la situation d'activité sont posées dans des contextes différents et de diverses façons. Par exemple, on demande d'abord une auto-classification du répondant par rapport au marché du travail; puis, dans une autre section du questionnaire, on pose une série de questions qui permettent de classer le répondant d'après des définitions standard relatives à la population active. (Pour un exemple différent, voir le cas de la Survey of Income and Program Participation, dans la section 4.2.)

Le deuxième cas couvre plusieurs situations. Au moins deux d'entre elles méritent d'être étudiées:

a) la situation où les données proviennent d'études comportant des réinterviews, et qui sont souvent recueillies spécialement pour obtenir des informations sur les probabilités d'erreurs de classification (dans un tel cas, la pratique courante consiste à assimiler les données de réinterview à des données de validation; pour en savoir davantage sur les méthodes classiques utilisées pour corriger des flux bruts d'après des données de réinterview, voir Abowd et Zellner 1985, Poterba et Summers 1986, ainsi que Chua et Fuller 1987);

b) la situation où les données sont recueillies rétrospectivement dans le cadre d'enquêtes par panel mais renvoient à une époque déjà couverte par l'interview précédente, ou sont recueillies lors d'une enquête supplémentaire qui est effectuée occasionnellement et qui couvre la ou les périodes de référence de l'enquête par panel actuelle. Il est évident que dans ce cas, des mesures différentes d'une ou de plusieurs variables peuvent être entachées d'erreurs de classification comportant en général des caractéristiques différentes.

Un grand nombre de questions qui sont soulevées ici sont expliquées dans les études de cas présentées dans la section 4, où la présence simultanée de données de panel et

de données rétrospectives, ainsi que d'indicateurs multiples d'une même variable latente, est utilisée pour obtenir des modèles parcimonieux.

## 3. ESTIMATION DE FLUX BRUTS ENTACHÉS D'ERREURS DE CLASSIFICATION

### 3.1 Un cadre général

La spécification de modèles statistiques qui est effectuée en vue de corriger des flux bruts de population active pour tenir compte d'erreurs de classification devrait permettre de prendre en considération la nature des données dont on dispose (comme nous l'avons indiqué dans la section précédente) et des hypothèses solides au sujet des processus qui sont à l'origine, premièrement, des transitions entre les situations d'activité (par ex., des chaînes de Markov) et, deuxièmement, des erreurs de mesure (par ex., des erreurs non corrélées et des erreurs corrélées).

Dans le cas le plus simple, nous considérons des données de panel où pour chaque période $t = 1, ..., T$, on observe une variable discrète $Y_t$ relative à une unité générique faisant partie d'un échantillon aléatoire de taille $n$. Dans nos études de cas, les unités sont des personnes et les périodes sont des mois ou des trimestres. $Y_t$ représente une valeur ou une situation choisie parmi un nombre $r$ de valeurs ou de situations distinctes. $Y_t$ est une mesure imparfaite d'$y_t$, qui représente la situation réelle d'une unité générique à l'époque $t$. En général, il n'est pas nécessaire de présupposer qu'$y_t$ varie au cours du même ensemble de situations $1, 2, ..., r$, mais afin de simplifier les choses, et sans diminuer la portée générale, nous allons considérer ici le même ensemble de situations que dans le cas d'$Y_t$.

Les stratégies qui servent à estimer des flux bruts sont fondées sur une spécification appropriée de la probabilité combinée relative aux processus réel et observé $P(Y_1, ..., Y_T, y_1, ..., y_T)$. L'analyse statistique s'appuie donc sur la marginalisation par rapport à des grandeurs observées:

$$P(Y_1, ..., Y_T) = \sum_{y_1=1}^{r} \cdots \sum_{y_T=1}^{r} P(Y_1, ..., Y_T, y_1, \ldots y_T). \quad (3.1)$$

Les modèles sont fondés sur des spécifications parcimonieuses de la fonction de probabilité combinée $P(Y_1, ..., Y_T, y_1, \ldots y_T)$. Cette parcimonie peut être obtenue essentiellement en décomposant la fonction en un produit de probabilités conditionnelles découlant d'un ensemble approprié d'hypothèses concernant la structure de dépendance existant entre les composantes $Y_1, ..., Y_T, y_1, ..., y_T$.

Pour nos besoins, un point de départ pratique pour la spécification des modèles est offert par des hypothèses formulées, premièrement, au sujet de la structure du processus qui est à l'origine des transitions réelles entre les situations d'activité, et, deuxièmement, au sujet du processus de mesure (en utilisant, par exemple, des connaissances solides ou des preuves empiriques tirées de la stratégie de collecte de données qui a été adoptée).

Dans le cas d'un modèle dont le but est de faire la distinction entre les transitions réelles et observées relatives au marché du travail, un exemple typique d'application de cette idée est fourni par les modèles de Markov à classe latente (van de Pol et Langeheine 1990). Les probabilités suivantes sont spécifiées dans le cas d'une unité générique:

$$q_t^{l_t j_t} = P(Y_t = l_t \mid y_t = j_t) \qquad t = 1, ..., T \quad (3.2)$$

$$\pi_t^{j_t j_{t-1}} = P(y_t = j_t \mid y_{t-1} = j_{t-1}) \qquad t = 2, ..., T \quad (3.3)$$

$$\pi_1^{j_1} = P(y_1 = j_1) \quad (3.4)$$

Les probabilités conditionnelles (3.2) représentent le rapport existant entre la situation réelle et la situation observée, c'est-à-dire la probabilité de déclarer, à l'époque $t$, la situation $l_t$ alors que la situation réelle est $j_t$. Il est clair que cette spécification implique l'hypothèse d'indépendance locale, c'est-à-dire que $Y_1, ..., Y_T$ sont indépendantes, étant donné $y_1, ..., y_T$. Les probabilités conditionnelles (3.3) représentent la dynamique du marché du travail, c'est-à-dire la probabilité qu'une transition de $j_{t-1}$ à $j_t$ ait lieu lorsqu'on passe de l'époque $t - 1$ à $t$. D'après (3.3), le processus de transition réel évolue en suivant une chaîne de Markov de premier ordre. Enfin, les probabilités (3.4) décrivent la situation initiale du processus de Markov.

La probabilité marginale dans le cas de la séquence observée (3.1) est donnée alors par:

$$P(Y_1 = l_1, \ldots Y_T = l_T) = \sum_{j_1=1}^{r} \cdots \sum_{j_T=1}^{r} \pi_1^{j_1} \prod_{t=2}^{T} q_t^{l_t j_t} \pi_t^{j_t j_{t-1}} \quad (3.5)$$

Dans le cas de quatre points de mesure, le modèle (3.5) est représenté de façon équivalente par le schéma de parcours de la figure 1, où les flèches indiquent des effets directs entre les variables.

$$\begin{array}{cccc} y_1 & \rightarrow & y_2 & \rightarrow & y_3 & \rightarrow & y_4 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ Y_1 & & Y_2 & & Y_3 & & Y_4 \end{array}$$

**Figure 1.** Schéma de parcours d'un modèle de Markov à classe latente dans le cas de quatre points de mesure.

Il est utile de noter que l'hypothèse d'indépendance locale équivaut à l'hypothèse d'erreurs de classification indépendant (ECI). Comme nous l'avons indiqué dans la section précédente, l'hypothèse ECI est sévèrement critiquée et paraît absolument déraisonnable dans le cas d'une collecte de données effectuée au moyen de questions rétrospectives.

Comme autre exemple, dans le cas de $T = 2$, des stratégies classiques utilisées pour corriger des flux bruts

d'après des études à réinterview peuvent être représentées dans le cadre décrit précédemment. Dans ce cas, on a recours à des informations supplémentaires; c'est-à-dire que les paramètres $q_t$ sont estimés à l'extérieur du cadre d'observation et introduits dans (3.5) afin d'obtenir directement $P(y_1, y_2)$.

Le même cadre peut être utilisé pour inclure des hypothèses plus générales au sujet du processus latent comme du processus de mesure, jusqu'à l'inclusion d'erreurs de classification corrélées en série. Comme cas intéressant, nous considérons le modèle proposé par Pfeffermann, Skinner et Humphreys (1998). Sans tenir compte ici de la situation initiale, voici comment ces auteurs expriment les probabilités de réponse conditionnelles:

$$q_t^{i_t j_t} = P(Y_t = i_t \mid y_t = j_t, Y_{t-1} = i_{t-1}) \qquad t = 2, ..., T, \quad (3.6)$$

ils maîtrisent ainsi l'hypothèse d'erreurs de classification indépendantes.

Une formulation similaire, qui vise à introduire, du moins en partie, une dépendance entre la situation observée à l'époque $t$ et la séquence de situations réelles existant aux époques $t$ et $t-1$, a été proposée par van de Pol et Langeheine (1992), qui élargissent le modèle afin de pouvoir inclure également une chaîne de Markov de deuxième ordre pour le processus de transitions réelles.

La stratégie de modélisation servant à estimer les flux bruts peut être élargie davantage dans diverses directions, dont celles décrites ci-après.

a) Le modèle peut être élargi tout naturellement pour exploiter la disponibilité d'indicateurs multiples d'une même situation réelle non observée. Cela implique que les probabilités de réponse, comme celles contenues dans (3.2), sont définies pour une ou plusieurs autres variables observées supplémentaires, qui sont traitées comme des mesures imparfaites de la même situation latente $y_t$. La figure 2 montre, à titre d'exemple, un modèle de Markov à classe latente dans le cas de deux indicateurs par variable latente et quatre points temporels. Dans ce modèle, chaque paire d'indicateurs qui renvoie à un point temporel donné est présumée être indépendante, sous condition de la variable latente correspondante; c'est-à-dire que la corrélation entre les indicateurs est complètement expliquée par leur rapport avec $y_t$.
b) L'hétérogénéité observée au niveau individuel, dans le processus de transition ou dans le processus de mesure, peut être introduite par l'entremise de conditions relatives à un ensemble de covariables $X_t$. Un exemple de cette approche est contenu dans l'étude de Pfeffermann, et coll. (1998). Ces auteurs utilisent des informations sur des covariables au niveau de l'unité et modélisent leur incidence sur la situation d'activité à l'aide d'un logit multinomial.
c) L'hétérogénéité non observée peut également être prise en considération, ce qui mène à des modèles à classe latente mixtes (van de Pol et Langeheine 1990). Un exemple simple de ce genre de modélisation est le modèle des personnes qui déménagent et des personnes qui restent, dans lequel on présuppose un comportement différent, au niveau latent, dans le cas de groupes d'unités, lorsque le nombre d'unités faisant partie d'un groupe ne peut être observé directement.

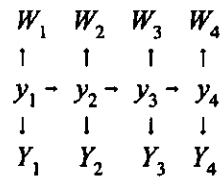



**Figure 2.** Schéma de parcours d'un modèle de Markov à classe latente dans le cas de quatre points de mesure et deux indicateurs pour chaque variable latente

### 3.2 Utilisation de modèles à classe latente et de modèles connexes dans l'estimation de flux bruts entachés d'erreurs de mesure

Les modèles à classe latente sont un exemple particulier de la formulation de modèles généraux que nous avons décrite dans la section précédente. Dans ces modèles, la situation d'activité réelle joue le rôle de variable latente, tandis que la situation observée sert d'indicateur de cette variable. Certaines des spécifications mentionnées dans la section précédente comprennent la dépendance entre les erreurs de classification. Une méthode générale et pratique qui permet de traiter cette dépendance, et qui comprend des modèles à classe latente standard avec erreurs de classification corrélées, est le modèle appelé LISREL modifié qui a été proposé par Hagenaars (1990).

Le modèle LISREL modifié est un prolongement de l'analyse de parcours de Goodman (1973), qui est un outil servant à décrire de type causal existant entre des variables nominales par l'entremise d'un système d'équations logit. Essentiellement, le prolongement intègre des variables latentes. Ainsi, un modèle LISREL modifié réunit un sous-modèle de mesure, qui spécifie la dépendance des indicateurs par rapport à des variables latentes, et un sous-modèle structurel qui spécifie des rapports ordonnés existant entre des variables latentes et entre d'éventuelles variables extérieures. Comme le suggère son nom, ce modèle peut être considéré également comme l'analogue pour variables discrète du modèle LISREL bien connu pour variables continues (Joreskög et Sörbom 1988).

Les modèles LISREL modifiés permettent d'introduire des erreurs de classification corrélées en série en insérant des effets directs entre les indicateurs (Hagenaars 1988). La présence d'effets directs implique que l'association entre les variables observées n'est pas entièrement expliquée par les effets des variables latentes sur leurs indicateurs, mais

qu'il existe une source d'association supplémentaire entre les indicateurs qui se situe au-delà de la partie qui est expliquée par le rapport existant entre ces indicateurs et les variables latentes.

Une fois qu'un modèle raisonnable a été spécifié, il faudrait vérifier l'identification. Le modèle comporte un grand nombre de variables non observables, et l'identification de tous les paramètres n'est pas assurée automatiquement.

Deux stratégies, utilisées éventuellement conjointement, offrent des possibilités raisonnables de réaliser l'identification: i) l'imposition de restrictions d'égalité plausibles à l'ensemble de paramètres; ii) le recours à des indicateurs multiples de la situation réelle non observée. Le modèle de Markov à classe latente représenté dans la figure 1, par exemple, n'est pas identifié sans que l'on n'impose des restrictions supplémentaires à ses paramètres. Si l'on présume que la chaîne latente est homogène au fil du temps, ou si les probabilités de réponse font l'objet de restrictions pour qu'elles soient égales au fil du temps, on peut montrer que le modèle est identifié (Lazarsfeld et Henry 1968). La disponibilité d'indicateurs multiples relatifs à la situation réelle non observée peut également aider à identifier des modèles de mesure complexes. Les critères d'identification utilisés pour certaines spécifications très particulières se sont avérés efficaces (par exemple, on peut montrer que le modèle représenté dans la figure 2 est identifié, mais des règles générales permettant de vérifier l'identification globale n'ont pas encore été proposées. Il est conseillé de vérifier au moins l'identification locale, c'est-à-dire l'identifiabilité des paramètres inconnus dans un voisinage de la solution du maximum de vraisemblance. Selon Goodman (1974), une condition suffisante pour permettre l'identification d'un modèle à classe latente est que la matrice d'information soit à rang complet. La condition énoncée par Goodman peut être difficile à vérifier du point de vue du calcul. En outre, dans le cas de certains ensembles de données, il peut arriver que la matrice d'information ne soit pas à rang complet, simplement parce que certaines estimations sont très proches des bornes de l'espace des paramètres. Une autre façon, empirique, de vérifier l'identifiabilité consiste à estimer le modèle en utilisant divers ensembles de valeurs de départ. Si des ensembles de valeurs de départ différents donnent la même valeur pour la fonction de vraisemblance logarithmique, mais des estimations de paramètres différentes, le modèle n'est pas identifié.

Pour ce qui est de l'estimation, les modèles LISREL peuvent être traités comme des modèles log-linéaires dirigés à variables latentes (Hagenaars 1997). Un modèle log-linéaire dirigé donne une séquence de modèles logit multinomiaux parcimonieux (comportant éventuellement des variables latentes) qui sont estimés par étapes. À chaque étape, on considère une variable dépendante et on estime un modèle logit multinomial sur un tableau de contingence qui a été réduit au-dessus des variables qui n'ont pas d'incidence directe sur la variable dépendante dans l'ordre causal. Finalement, les estimations obtenues à chaque étape sont réunies, afin d'obtenir des paramètres estimés pour le modèle complet. La modélisation log-linéaire dirigée donne exactement les mêmes estimations, les mêmes erreurs-types et les mêmes statistiques de contrôle que la méthode standard de Goodman, mais en utilisant des tableaux marginaux plus simples. Si le modèle causal contient une ou plusieurs variables latentes, il faut utiliser une méthode d'estimation appropriée, par exemple, une application de l'algorithme à mémoire externe (Meng et Rubin 1993).

La validité empirique du modèle causal complet peut être testée en comparant les fréquences prévues estimées avec les fréquences observées, dans le tableau complet, à l'aide du rapport de vraisemblance $L^2$ et de la statistique de Pearson $X^2$. Cependant, en raison de la nature de la structure des données observées relatives aux transitions qui ont lieu sur le marché du travail, de nombreuses cellules indiquent des fréquences observées très faibles. Pour cette raison, les critères habituels $X^2$ et $L^2$ devraient être utilisés uniquement pour obtenir une indication générale d'ajustement, étant donné que leur distribution asymptotique $\chi^2$ n'est plus garantie, en raison de la répartition clairsemée et non équilibrée des fréquences dans le tableau de contingence.

Diverses stratégies permettent d'élargir et d'améliorer l'évaluation des modèles; trois d'entre elles méritent d'être mentionnées dans le présent contexte:

i) Un modèle restreint subordonné à un modèle plus vaste peut être vérifié à l'aide du test conditionnel, c'est-à-dire en considérant la différence existant entre les deux modèles en ce qui a trait à la valeur de $L^2$, qui est distribuée asymptotiquement comme $\chi^2$ dans des conditions plus faibles (Goodman 1981, et Haberman 1978).

ii) En général, le recours à des critères multiples peut être une stratégie sensée. Des indices fondés sur le critère d'information, comme le critère d'information d'Akaike (CIA) ou le critère d'information de Bayes (CIB), peuvent être utiles pour comparer des modèles de remplacement non hiérarchiques. Un autre avantage du CIA et du CIB est que, dans la procédure de sélection, ces critères soupèsent la qualité d'ajustement d'un modèle par rapport à la parcimonie de celui-ci, en considérant les degrés de liberté du modèle et la taille de l'échantillon. (CIA = $L^2$ - 2 x degrés de liberté. CIB = $L^2$ - ln(N+1) x degrés de liberté.) Le modèle que l'on préfère dans ce contexte est celui qui présente la plus faible valeur du CIA ou du CIB.

iii) Des méthodes de rééchantillonnage de Monte Carlo peuvent être mises en oeuvre pour simuler la distribution asymptotique de $X^2$ et $L^2$ (Langeheine, Pannekoek et van de Pol 1995).

## 4. DEUX ÉTUDES DE CAS

### 4.1 Le cadre général

Dans la présente section, nous présentons deux applications du modèle LISREL modifié qui servent à corriger des flux bruts observés relatifs au marché du travail. Les données utilisées proviennent des deux enquêtes mentionnées ci-après, qui sont fondées sur des plans d'échantillonnage en partie différents:

1) la U.S. Survey of Income and Program Participation (SIPP), enquête-ménage par panels multiples qui recueille des informations rétrospectives sur les antécédents d'activité entre vagues;
2) l'enquête française sur la population active (EFPA), enquête rétrospective annuelle comportant des périodes de référence qui se chevauchent sur un mois.

Pour chacune des études de cas, nous spécifions un modèle sur la base d'informations a priori qui ont trait tant au processus de transition qu'au mécanisme à l'origine des erreurs. Dans la spécification des modèles, les informations a priori sont essentielles à l'obtention de modèles parcimonieux et plausibles.

Tous les modèles sont écrits sous forme d'un modèle LISREL modifié et estimés à l'aide d'un algorithme à mémoire externe. Nous avons utilisé le programme *EM* (Vermunt 1993) et vérifié, dans le cas de tous les modèles, les maximums locaux.

Les deux derniers modèles s'avèrent plutôt complexes, étant donné qu'ils intègrent la corrélation existant entre les erreurs de classification et des hypothèses spécifiques sur le comportement des répondants. Ce fait, ainsi que le caractère clairsemé et non équilibré du tableau de contingence observé, typique des transitions relatives à la population active, exige des critères d'évaluation de la qualité d'ajustement autres que $L^2$ et $X^2$. Dans la première étude de cas, des modèles de remplacement ont été évalués au moyen du CIB et sur la base de connaissances de fond relatives au marché du travail américain. Dans le deuxième cas, des modèles de remplacement ont été comparés au moyen du test conditionnel.

Dans les sections qui suivent, les modèles sont présentés sous forme logique et verbale, tandis que la formulation mathématique pour le modèle final est donnée dans l'annexe pertinente.

### 4.2 Les données de la SIPP

La SIPP est une enquête-ménage par panels multiples effectuée par le U.S. Bureau of the Census dans le but de recueillir des informations sur des sujets comme l'emploi, le revenu, la participation à des programmes sociaux, *etc.* La population de référence est composée des personnes non institutionnalisées âgées de plus de quatorze ans.

Cette enquête permanente a démarré en 1984. De façon régulière, un nouvel échantillon de ménages, appelé «panel», a été choisi tous les ans et suivi durant une période de deux ans et demi (pour une description détaillée de la SIPP, voir U.S. Department of Commerce 1991, ainsi que Citro et Kalton 1993).

Chaque panel est divisé de façon aléatoire en quatre «groupes de renouvellement» et interviewé huit fois à intervalles de quatre mois. Pour des raisons pratiques, chaque groupe de renouvellement est interviewé une fois par mois au cours d'une période de quatre mois consécutifs, et au moyen de questions rétrospectives, on recueille des informations sur la période de quatre mois qui s'écoule avant la tenue des interviews subséquentes. Chaque série d'interviews effectuées auprès de l'ensemble de l'échantillon est désignée par le terme «vague».

Nous allons nous référer au panel de 1986, qui a commencé au mois de février 1986 et qui a pris fin au mois d'août 1988. Nous allons considérer la période intermédiaire comprise entre janvier 1986 et janvier 1987, au sujet de laquelle nous disposons d'informations provenant de quatre groupes de renouvellement. La figure 3 montre le plan d'échantillonnage de l'enquête en ce qui a trait à notre échantillon.

Les informations sur l'activité sont recueillies principalement dans la section du questionnaire intitulée «Labour Force and Recipiency» (pour un renseignement supplémentaire, qui est recueilli dans une autre section du questionnaire, voir ci-après), où l'on demande à chaque répondant de déclarer sur une base hebdomadaire ses antécédents d'activité pour la période de quatre mois (dix-huit semaines) précédente, en passant en revue une série de questions filtrées. On demande tout d'abord au répondant s'il ou elle occupait un emploi ou avait une entreprise a un moment quelconque durant la période de référence. Dans le cas d'une réponse négative, on demande au répondant d'indiquer s'il ou elle a passé du temps à chercher du travail ou s'il ou elle était mis(e) en disponibilité, et dans l'affirmative, au cours de quelles semaines exactement. Par contre, si le répondant répond oui à la première question, (c'est-à-dire s'il ou elle a travaillé durant un certain temps) et mentionne un emploi ou une entreprise comportant une certaine durée au cours de la période de référence, il est invité à passer à la section suivante du questionnaire. Aux répondants qui ne mentionnent pas une situation stable par rapport au marché du travail, on pose une longue série de questions, afin de déterminer la situation d'activité dans laquelle ils se trouvaient au cours de chacune des semaines de référence.

Les informations recueillies sur une base hebdomadaire sont généralement consignées pour obtenir une classification mensuelle fondée sur les trois catégories courantes suivantes: personne occupée (O), en chômage (C) et hors de la population active (H). Dans le cas de personnes qui occupent différentes catégories au cours d'un même mois, la situation d'activité mensuelle est celle identifiée par la catégorie «modale» en ce qui a trait aux semaines du mois en question (Martini 1989).

| Mois d'interview | Groupe de renouv. | Vague | Mois de référence | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Février | 2 | 1 | Oct | Nov | Déc | Jan | | | | | | |
| Mars | 3 | 1 | | Nov | Déc | Jan | Fév | | | | | |
| Avril | 4 | 1 | | | Déc | Jan | Fév | Mars | | | | |
| Mai | 1 | 1 | | | | Jan | Fév | Mars | Avr | | | |
| Juin | 2 | 2 | | | | | Fév | Mars | Avr | Mai | | |
| Juillet | 3 | 2 | | | | | | Mars | Avr | Mai | Juin | |
| Août | 4 | 2 | | | | | | | Avr | Mai | Juin | Juil |
| Septembre | 1 | 2 | | | | | | | | Mai | Juin | Juil | Août |

**Figure 3.** Plan de renouvellement pour le panel de 1986 de la SIPP (deux premières vagues)

Les flux bruts observés entre deux mois civils génériques sont alors obtenus de la manière énoncée ci-après.

a) Dans le cas des personnes appartenant aux trois premiers groupes de renouvellement, d'après des données rétrospectives recueillies lors d'une même interview. Ces flux observés sont appelés transitions «intra-vague» (IV).
b) Dans le cas des personnes appartenant au quatrième groupe de renouvellement, en réunissant les informations recueillies lors de deux interviews différentes, effectuées à quatre mois d'intervalle. Ces flux observés sont appelés transitions «entre vagues» (EV).

Dans l'estimation des variations mensuelles, un problème particulier que l'on rencontre dans le cas des données de la SIPP est l'effet dit «de lisière» (Young 1989): on observe davantage de variations lorsque les données relatives à deux mois qui se suivent sont recueillies dans deux vagues différentes (la transition chevauche la lisière entre les deux vagues) que dans le cas où les données proviennent d'une même interview. L'effet de lisière est présent dans l'ensemble de l'enquête; des preuves de cet effet en ce qui a trait à plusieurs variables importantes sont décrites par Martini (1988), Marquis et Moore (1989), ainsi que par Kalton et Miller (1991).

Le tableau 1 illustre ce phénomène dans le cas de notre échantillon du panel de 1986 de la SIPP. La ligne 4→1 contient les transitions IV moyennes; les lignes 1→2, 2→3 et 3→4 contiennent les transitions EV moyennes relatives à la position des deux mois de référence pertinents dans chaque vague (par exemple, la ligne 1→2 contient les transitions qui ont eu lieu dans chaque vague entre les deux premiers mois de référence). Le tableau 1 montre clairement que les transitions IV observées décrivent un marché du travail beaucoup plus stable que les transitions EV. De plus, la stabilité IV augmente à mesure qu'on recule dans la vague (de 3→4 à 1→2).

Une cause plausible de l'origine de l'effet de lisière et du patron systématique de transitions observées dans l'ensemble d'une vague est la différence entre l'incidence des erreurs de mesure contenues dans les données recueillies selon la stratégie IV et l'incidence des erreurs de mesure touchant les données obtenues selon la stratégie EV. Plus précisément, il est probable que les erreurs de classification présentent des degrés de corrélation différents selon le type de flux observés (IV ou EV). La stabilité plus grande que reflètent les transitions IV pourrait être due à une corrélation élevée entre les erreurs de classification En effet, s'il n'y avait pas de corrélation entre les erreurs, en particulier dans le cas des transitions IV, on ne s'attendrait pas à observer des preuves de l'existence d'un effet de lisière.

Un grand nombre de causes plausibles des erreurs corrélées sont évoquées dans les textes spécialisés portant sur la psychologie cognitive et sur les techniques d'enquête, en particulier sur les effets des erreurs de mémoire (voir à cet égard Bernard, Killworth, Kronenfeld et Sailer 1984, ainsi que O'Muircheartaigh 1996); parmi ces causes, on mentionne un effet de «conditionnement»: les répondants ont tendance à donner la même réponse lorsqu'on recule à l'intérieur de la vague, et dans des cas extrêmes, ils répètent machinalement la même réponse pour les quatre mois.

**Tableau 1**
Transitions mensuelles (×100) dans le cas du panel de 1986 de la SIPP, janvier 1986 à janvier 1987

| | Type | OO | OC | OH | CO | CC | CH | HO | HC | HH |
|---|---|---|---|---|---|---|---|---|---|---|
| 1→2 | IV | 98,27 | 1,04 | 0,69 | 15,46 | 79,63 | 4,91 | 1,15 | 1,42 | 97,43 |
| 2→3 | IV | 97,91 | 1,13 | 0,96 | 17,34 | 75,96 | 6,70 | 1,38 | 1,71 | 96,91 |
| 3→4 | IV | 97,85 | 1,20 | 0,95 | 19,23 | 73,25 | 7,52 | 1,28 | 1,69 | 97,03 |
| 4→1 | EV | 94,03 | 2,10 | 3,87 | 26,81 | 42,20 | 30,99 | 5,65 | 3,77 | 90,58 |

Une abondante documentation empirique montre que ce genre d'effet de conditionnement est la principale source d'erreurs de classification dans le cas des données de la SIPP. D'autres sources d'erreurs potentielles, typiques des enquêtes par panel, n'ont pas d'incidence importante sur les données de la SIPP. Les études de vérification de dossiers administratifs fournissent peu de preuves, s'il en existe, d'un effet lié au temps de présence dans l'échantillon (Chakrabartry et Williams 1989; McCormick, Butler et Singh 1992). Comme considération d'ordre général, nous pouvons dire que dans les données de la SIPP, l'effet de

lisière prédomine sur les autres sources d'erreurs qui peuvent biaiser les estimations de flux bruts.

En récapitulant, le recours à une méthode fondée sur la modélisation pour obtenir des flux bruts non biaisés à partir de données de la SIPP est justifié par les deux arguments suivants:

a) la présence manifeste d'erreurs de classification corrélées;
b) le recours à des informations a priori sur le mécanisme à l'origine des données tirées de deux sources;
   b1) des preuves précises tirées des flux bruts observés de la SIPP, comme l'effet de lisière et l'accroissement de la stabilité lorsqu'on recule à l'intérieur de la vague (mentionnées précédemment);
   b2) des explications générales contenues dans la documentation sur les enquêtes portant sur le comportement des répondants.

Afin de supprimer, dans les données de la SIPP, les erreurs de classification contenues dans les flux bruts relatifs à la population active, nous avons élaboré un modèle d'après les hypothèses et les informations suivantes:

a) le processus de transition réelle suit une chaîne de Markov de premier ordre;
b) des données sur les transition IV entachées d'erreurs de classification corrélées, d'après un patron qui est spécifié ci-après;
c) dans le cas des transitions EV, l'hypothèse d'erreurs de classification indépendantes est confirmée;
d) les groupes de renouvellement sont des échantillons équivalents aux fins d'une modélisation également, c'est-à-dire que les répondants se comportent de la même manière dans les quatre groupes de renouvellement;
e) les données de la SIPP donnent deux indications sur la situation mensuelle par rapport au marché du travail de chaque répondant: les informations détaillées recueillies dans la section du questionnaire intitulée Labour Force and Recipiency, dont il a été question précédemment, ainsi que les renseignements supplémentaires recueillis dans la section intitulée Earnings and Employments, dans laquelle on demande aux répondants, sur une base hebdomadaire, s'ils occupaient ou non un emploi au cours de la période de référence.

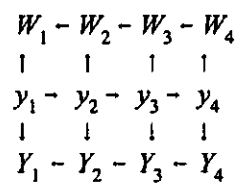



**Figure 4.** Schéma de parcours d'un modèle LISREL modifié pour quatre points de mesure et deux indicateurs pour chaque variable latente (pour connaître la signification des symboles, voir le texte principal)

La figure 4 représente le schéma de parcours d'une version simplifiée du modèle (c'est-à-dire une version qui n'a pas pour but de représenter en détail le patron d'erreurs de classification corrélées, ni de prendre en considération le fait que nous sommes en présence de quatre groupes de renouvellement) pour quatre points temporels, c'est-à-dire quatre mois consécutifs. Dans ce cas, $y_t$ ($t = 1, 2, 3, 4$) représente des variables latentes, $Y_t$ et $W_t$ représentent des indicateurs, et les flèches représentent des effets directs entre paires de variables. L'indicateur $Y_t$ renvoie à la situation d'activité déclarée, qui est décrite au moyen des trois catégories habituelles (O, C et H), tandis que $W_t$ renvoie à la variable binaire «emploi/aucun emploi». Compte tenu du fait que les informations sont recueillies dans deux sections distinctes du questionnaire, et par des méthodes d'interview différentes, on peut présupposer qu' $Y_t$ et $W_t$ sont indépendantes, étant donné $y_t$. D'autre part, les effets directs entre les indicateurs sont à l'origine d'erreurs de classification corrélées au fil du temps; la réponse donnée pour l'époque $t + 1$ a une incidence sur celle qui est donnée pour l'époque $t$. Il est à noter également qu'une variable supplémentaire $G$, qui comporte quatre catégories, devrait être ajoutée au schéma, afin de rendre compte du nombre de personnes faisant partie des groupes de renouvellement. Tous les indicateurs dépendent de $G$, étant donné que les unités faisant partie de divers groupes sont interviewés au cours de mois civils différents.

L'équation de base du modèle décompose la proportion contenue dans la cellule générique du tableau de contingence à neuf caractères pour l'obtention du produit des probabilités conditionnelles énoncées dans l'annexe A, équations (A1) à (A7). Une version préliminaire du modèle a été proposée par Bassi, Croon, Hagenaars et Vermunt (1994).

L'équation (A1) définit la probabilité d'appartenir à l'un des quatre groupes de renouvellement. Les équations (A2) et (A3) définissent respectivement la situation initiale et les probabilités de transition de la chaîne de Markov de premier ordre latente. Les équations (A4) et (A5) définissent les probabilités de réponse concernant l'indicateur $Y_t$; enfin, les équations (A6) et (A7) définissent les probabilités analogues concernant l'indicateur dichotomique $W_t$. Les probabilités de réponse sont définies de manière à ce que la réponse qui est donnée pour un mois donné dépende à la fois de la situation réelle actuelle ($y_t$) et des situations «passée» réelles et «passée» déclarées ($y_{t+1}$ et $Y_{t+1}$). Le mot «passée» renvoie à la façon de penser des répondants lorsqu'ils répondent à des questions rétrospectives: ils commencent à se souvenir à partir du moment le plus proche de l'interview et remontent jusqu'à la fin de la période de référence.

Un ensemble complexe de contraintes a été imposé aux probabilités de réponse de (A4), (A5), (A6) et (A7), afin de tenir compte, premièrement, de l'effet de conditionnement, et, deuxièmement, du fait que les quatre groupes de renouvellement sont des échantillons équivalents en ce qui a trait au mécanisme qui est à l'origine des erreurs.

Ces contraintes sont formulées en détail dans l'annexe A. Essentiellement, elles intègrent une connaissance *a priori* du comportement des répondants et nous permettent de spécifier un modèle parcimonieux. En particulier, les équations (A8) à (A14) correspondent aux énoncés suivants:

a) concernant les erreurs de classification IV, on présuppose, en suivant Hubble et Judkins (1989),
   a1) qu'un répondant qui déclare une situation d'activité erronée va continuer à donner la même réponse pour le mois qui suit immédiatement, en remontant dans la vague (A8);
   a2) cependant, si la situation à l'époque $t+1$ est déclarée correctement, la probabilité de réponse pour le mois qui précède immédiatement dépend uniquement de la situation réelle actuelle (A9);
   a3) le même mécanisme à l'origine d'erreurs agit dans le cas des deux indicateurs. Dans le cas de $W_t$, nous énonçons qu'une réponse correcte est donnée lorsque la situation réelle est O et qu'un emploi est déclaré, et lorsque la situation réelle est C ou H, et qu'aucun emploi n'est déclaré, (A10) et (A11).
b) Les probabilités de réponse établies sont les mêmes dans le cas de tous les groupes de probabilités, (A12) à (A15). Par exemple, les égalités dans (A12) signifient que les probabilités de réponse, dans le cas des personnes faisant partie du groupe de renouvellement 1, pour le mois d'avril, sont égales aux probabilités de réponse relatives aux personnes faisant partie du groupe 4, pour le mois de mars, aux probabilités relatives au personnes du groupe 3, pour le mois de février, et aux probabilités relatives aux personnes du groupe 2, pour le mois de janvier. (Les probabilités sont établies de manière à ce qu'elles soient égales car elles renvoient à la réponse donnée pour le dernier mois de la vague.)

Le modèle a été estimé pour corriger les flux bruts observés relatifs au trimestre compris entre janvier et avril 1986 (tableau 2). La comparaison entre les flux observés et les flux estimés met en évidence le fait que le modèle réduit l'effet de lisière: les transitions IV sont corrigées dans le sens d'un marché du travail plus dynamique, tandis que les transitions EV sont corrigés dans le sens opposé. Il est utile de noter que les effets de la correction par modélisation sont plus évidents dans le cas des flux qui partent de la situation «sans emploi», qui sont caractérisés par une plus grande mobilité.

La qualité de l'ajustement du modèle a été évaluée au moyen de critères multiples, comme le CIB et le test conditionnel pour modèles hiérarchiques, ainsi qu'en faisant appel à l'interprétabilité et à la cohérence par rapport à des connaissances de fond sur la dynamique du marché du travail américain des années 80.

### 4.3 Les données de l'enquête française sur la population active

La deuxième étude de cas porte sur les flux observés sur le marchés du travail dans le cadre de l'enquête française sur la population active (EFPA), qui est effectuée annuellement par l'INSEE.

La population de référence de l'EFPA est composée de tous les membres des ménages français qui sont âgés de quinze ans et plus dans l'année au cours de laquelle doit avoir lieu l'interview. Le plan d'échantillonnage de cette enquête prévoit tous les ans le renouvellement d'un tiers de l'échantillon. Les informations sur l'activité sont recueillies au moyen de questions rétrospectives qui portent sur la période de référence de treize mois qui a précédé l'interview. À chaque répondant, on demande d'indiquer, sur une base mensuelle, quelle était sa situation par rapport au marché du travail durant la période de référence; pour ce faire, le répondant dispose d'une grille où il ou elle peut indiquer sa situation au cours de chaque mois. Cette grille comporte huit catégories: indépendant, occupé pour une durée déterminée, occupé de façon permanente, sans emploi, en formation, étudiant(e), service militaire et autre (retraité(e), ménagère, *etc*).

**Tableau 2**
Transitions mensuelles observées et estimées dans le cadre de la SIPP (×100), janvier à avril 1986

| | | OC | OC | OH | CO | CC | CH | HO | HC | HH |
|---|---|---|---|---|---|---|---|---|---|---|
| J-F | IV | 98,11 | 1,17 | 0,72 | 14,53 | 80,16 | 5,31 | 0,90 | 1,57 | 97,53 |
| | EV | 94,08 | 2,17 | 3,75 | 23,58 | 44,30 | 32,12 | 5,62 | 3,45 | 90,93 |
| | Estimé | 97,25 | 1,47 | 1,28 | 16,08 | 77,16 | 6,76 | 1,59 | 1,32 | 97,09 |
| F-M | IV | 98,66 | 0,92 | 0,42 | 16,06 | 78,67 | 5,27 | 0,64 | 1,65 | 97,71 |
| | EV | 94,88 | 1,91 | 3,21 | 21,90 | 48,54 | 29,56 | 4,99 | 4,11 | 90,90 |
| | Estimé | 97,83 | 1,20 | 0,97 | 19,40 | 74,01 | 6,59 | 1,21 | 1,50 | 97,29 |
| M-A | IV | 98,71 | 0,64 | 0,65 | 20,76 | 71,74 | 7,50 | 1,47 | 1,05 | 97,48 |
| | EV | 95,59 | 1,52 | 2,89 | 30,48 | 34,92 | 34,60 | 6,34 | 3,78 | 89,88 |
| | Estimé | 98,11 | 0,95 | 0,94 | 26,42 | 65,75 | 7,83 | 2,17 | 0,71 | 97,12 |

Aux fins de notre analyse, nous avons regroupé ces huit catégories sous les trois catégories habituelles (O, C, H). Nous considérons comme des personnes occupées les répondants qui se classent dans les trois premières catégories, comme des personnes en chômage les répondants qui se classent dans la quatrième catégorie, et comme personnes hors du marché du travail les répondants qui se classent dans les catégories restantes.

Nous analysons les informations qui ont été recueillies au cours des deux vagues consécutives de mars 1991 et de mars 1992, auprès d'un sous-échantillon de personnes, c'est-à-dire les personnes qui ont répondu à trois interviews consécutives (janvier 1990, mars 1991 et mars 1992) et qui étaient âgées entre 18 et 20 ans en 1992, pour un total de 5 427 personnes. Les périodes de référence des deux vagues que nous considérons se chevauchent en mars 1991. Nous avons donc deux renseignements sur la situation d'activité pour le mois en question: un qui a été recueilli en mars 1991 et l'autre qui a été obtenu au moyen d'une question rétrospective, douze mois plus tard.

Le patron de transitions mensuelles observées que présente notre échantillon de l'EFPA révèle quelques faits intéressants, qui sont attribuables en grande partie au jeune âge des répondants.

Les transitions montrent un degré modéré de variation saisonnière liée au calendrier scolaire. De juin à juillet, par exemple, nous observons une proportion plus grande que la moyenne de gens qui arrivent sur le marché du travail comme personnes occupées; par contre, d'août à septembre, une proportion de gens plus grande que la moyenne quitte la situation d'emploi (vraisemblablement pour étudier).

La distribution marginale des trois situations entre mars 1990 et mars 1992 montre que les personnes faisant partie de notre échantillon arrivent progressivement sur le marché du travail: en mars 1990, on constate que 44 % d'entre elles occupent un emploi ou sont au chômage, tandis qu'en mars 1992, cette proportion a augmenté à 54 %.

Le renseignement double pour le mois de mars 1991 fournit quelques preuves brutes concernant les erreurs de réponse contenues dans les données: 8 % des répondants déclarent une situation différente dans l'une des interviews par rapport à l'interview précédente. Dans le cas de la période de février à avril 1991, on observe deux types de flux: un flux intra-vague (IV), c'est-à-dire que les informations sur la situation par rapport au marché du travail ont été recueillies au cours d'une même interview, et un flux entre vagues (EV), c'est-à-dire que les informations ont été recueillies au cours de deux interviews distinctes (tableau 3).

Comme prévu, les transitions IV dénotent un marché du travail plus stable que les transitions EV. Ce fait peut être considéré comme une indication de la présence d'erreurs de classification corrélées dans les données. Nous avons abordé de manière approfondie les patrons et les causes de la corrélation entre erreurs dans les deux sections précédentes, et les considérations qu'elles contiennent peuvent en grande partie être élargies aux données de l'enquête française.

**Tableau 3**
Transitions observés dans le cadre de l'EFPA (×100), de février à avril 1991

| | | OO | OC | OH | CO | CC | CH | HO | HC | HH |
|---|---|---|---|---|---|---|---|---|---|---|
| F-M | IV | 98,19 | 1,67 | 0,14 | 9,11 | 90,65 | 0,24 | 0,28 | 0,11 | 99,61 |
| | EV | 93,17 | 3,58 | 3,25 | 25,18 | 65,23 | 9,59 | 3,75 | 1,96 | 94,29 |
| M-A | IV | 98,60 | 1,04 | 0,36 | 8,89 | 90,37 | 0,74 | 0,24 | 0,29 | 99,47 |
| | EV | 93,24 | 3,33 | 3,43 | 25,90 | 63,79 | 10,31 | 3,79 | 2,07 | 94,14 |

En général, dans le cas d'une enquête rétrospective comportant une période aussi longue pour laquelle les répondants doivent faire appel à leur mémoire, nous nous attendons à ce que le manque de mémoire soit la cause principale des erreurs de classification. Nous nous attendons également à ce que la probabilité de donner une réponse incorrecte augmente à mesure que s'accroît l'écart entre le mois de référence et le mois où a lieu l'interview. Ce fait peut être considéré comme la principale cause de corrélation entre erreurs de classification, conjointement avec le télescopage et les effets de conditionnement, qui pourraient également toucher les données de l'enquête française (voir Magnac et Visser 1995).

On peut raisonnablement penser que l'effet global des erreurs de classification entraîne une sous-estimation de la mobilité en ce qui a trait au marché du travail français.

Sur la base de ces considérations, nous avons spécifié un modèle dans le but de supprimer les erreurs de mesure dans les flux bruts observés à intervalles trimestriels (tableau 4). La dernière colonne du tableau 4 indique le pourcentage observé de personnes qui ont changé de situation entre les deux mois considérés (MO = mobilité observée). En moyenne, au cours des cinq transitions IV, on observe 6,122 % de mobilité entre deux mois consécutifs.

Comme dans les études de cas précédentes, nous désignons par $y_t (t = 1, 2, 3, 4, 5, 6)$ les situations réelles par rapport au marché du travail, et au moyen de lettres majuscules, $Y_t (t = 2, 3, 4, 5, 6)$, leurs indicateurs, qui représentent les situations d'activité observées en mars 1992 (en se référant à mars, juin, septembre et décembre 1991, ainsi qu'à mars 1992); $W_t (t = 1, 2)$ représente les situations d'activité observées en mars 1991 (en se référant à décembre et mars 1991). Comme auparavant, $y_t$, $Y_t$ et $W_t$ renvoient à l'une des trois catégories habituelles (O, C et H).

Le modèle est spécifié en décomposant la proportion indiquée dans la cellule générique du tableau de contingence à sept caractères, comme dans l'annexe B, équations (B1) à (B6).

Étant donné que nous observons deux indicateurs pour un mois seulement, un modèle dans lequel on présuppose des effets directs entre les indicateurs serait sous-identifié. C'est pourquoi nous ne sommes pas en mesure de modéliser explicitement la dépendance entre les situations observées.

**Tableau 4**
Taux de transition trimestriels observés (×100), EFPA, décembre 1990 à mars 1992
(MO = mobilité observée)

| | | OO | OC | OH | CO | CC | CH | HO | HC | HH | MO |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D90-M91 | IV | 94,77 | 4,25 | 0,98 | 24,53 | 72,40 | 3,07 | 0,98 | 0,66 | 98,36 | 5,08 |
| | EV | 91,50 | 4,86 | 3,64 | 31,60 | 56,84 | 11,56 | 4,40 | 2,10 | 93,50 | 10,16 |
| M91-J91 | IV | 96,03 | 3,02 | 0,95 | 23,21 | 74,32 | 2,47 | 1,28 | 0,68 | 98,04 | 4,54 |
| | EV | 91,48 | 4,63 | 3,89 | 35,01 | 54,20 | 10,79 | 4,84 | 2,14 | 93,02 | 12,04 |
| J91-S91 | IV | 94,29 | 3,94 | 1,77 | 20,93 | 78,29 | 0,78 | 4,71 | 2,95 | 92,34 | 7,85 |
| S91-D91 | IV | 93,73 | 4,48 | 1,79 | 23,63 | 74,89 | 1,48 | 3,22 | 1,65 | 95,13 | 7,23 |
| D91-M92 | IV | 93,90 | 4,80 | 1,30 | 21,67 | 76,74 | 1,59 | 1,70 | 0,59 | 97,71 | 5,91 |

La seule façon de rendre compte d'erreurs de classification présentes dans les données de l'EFPA consiste à laisser dépendre les situations observées de transitions latentes. Par ailleurs, cela semble être une présupposition sensée dans le cas d'enquêtes rétrospectives. En effet, les flux entre deux situations différentes peuvent facilement faire l'objet d'un classement erroné, parce que dans certaines circonstances, il peut vraiment être difficile de situer correctement les événements. Par exemple, des employés qui perdent leur emploi ou qui prennent leur retraite (flux OC et CH) vont généralement prendre les jours de congé auxquels ils ont droit, et peuvent ne pas savoir exactement à quel moment ils ont perdu ou quitté leur emploi. Il peut être également difficile de se rappeler à quel moment on s'est joint à la population active, surtout si l'arrivée sur le marché du travail a eu lieu après avoir quitté l'école (flux HC et HO)(van de Pol et Langeheine 1997).

Le modèle LISREL modifié, formulé mathématiquement dans l'annexe B, est fondé sur les hypothèses de fond énoncées ci-après.

Au niveau latent, les transitions suivent une chaîne de Markov non stationnaire de premier ordre (équations (B1) et (B2)). En effet, l'importance de la saisonnnalité dans le cas des transitions observées incite à éviter d'imposer une stationnarité de quelque ordre que ce soit à la chaîne de Markov latente.

Les probabilités relatives aux réponses dans le cas de données recueillies au cours des deux vagues dépendent de la transition latente qui a lieu entre $t$ et $t+1$ (les équations (B3) et (B4) renvoient à des données recueillies en mars 1992, et les équations (B5) et (B6) à des données recueillies en mars 1991).

Afin de décrire en détail le mécanisme qui est à l'origine des erreurs et de spécifier un modèle plus parcimonieux, nous avons imposé les contraintes suivantes aux probabilités relatives aux réponses:

a) les probabilités de réponse qui ont trait au même mois faisant partie d'années subséquentes (décembre et mars) sont considérées comme étant égales;
b) les probabilités de réponse à l'époque $t$, étant donné que la situation réelle n'a pas changé entre l'époque $t$ et l'époque $t+1$, sont considérées comme étant constantes au fil du temps;
c) les probabilités de réponse sont considérées comme étant égales pour juin et septembre 1991;
d) en général, les répondants qui changent de situation entre le mois $t$ et $t+1$ (transitions OC, OH, CO, CH, et HC) déclarent, à l'époque $t$, soit la situation réelle à l'époque $t$, soit la situation réelle à l'époque $t+1$; autrement dit, les répondants ne déclarent pas une situation qu'ils n'ont pas quitté ou à laquelle ils n'ont pas accédé;
e) cependant, si la transition latente a lieu entre les situations H et O, nous admettons les trois réponses à l'époque $t$, c'est-à-dire que nous considérons que les gens qui trouvent un emploi peuvent confondre leur situation précédente (à l'époque $t$) et hésiter entre C et H.

La contrainte c) est imposée principalement pour obtenir un modèle parcimonieux. Cette contrainte intègre l'idée selon laquelle les probabilités de réponse relatives à des mois qui se situent plus ou moins dans la partie centrale de la période de référence ne varient pas beaucoup.

Les contraintes b) et d) reflètent le fait que les probabilités de réponse dépendent de transitions latentes. Nous nous attendons à ce que ces probabilités ne varient pas beaucoup au fil du temps lorsqu'il n'y a pas de changement latent (contrainte b)), tandis que nous prévoyons que la probabilité de mal situer des changements, surtout dans le cas de situations ambiguës, augmente avec la longueur de la période pour laquelle le répondant doit faire appel à sa mémoire. Les contraintes d) visent à tenir compte de l'effet de télescopage.

La figure 5 montre le schéma de parcours du modèle estimé.

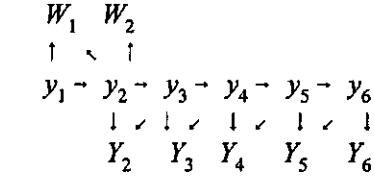



**Figure 5.** Schéma de parcours d'un modèle LISREL modifié, avec six points de mesure et deux indicateurs pour une seule variable latente

Le tableau B1, dans l'annexe B, montre le patron des restrictions imposées aux probabilités relatives aux réponses (a) à e)); le tableau indique quels sont les paramètres qui sont établis sur un pied d'égalité et lesquels sont établis à 0, afin d'introduire dans le modèle les contraintes précitées, définies par les équations (B1) à (B6).

Le modèle final a été choisi après avoir comparé plusieurs modèles, comme le montre le tableau 5.

**Tableau 5**
Choix du modèle (ME = mobilité estimée)

| MODÈLE | $L^2$ | dl | $\Delta L^2$ | valeur de $p$, test conditionnel | ME |
|---|---|---|---|---|---|
| A | 2509,5759 | 2124 | | | 5,424 |
| A1 | 3450,1716 | 2154 | 940,5957 | 0 | 4,918 |
| A2 | 3849,9470 | 2178 | 399,7754 | 0 | 5,798 |
| B | 816,1620 | 2076 | | | 5,888 |
| B1 | 855,2282 | 2094 | 39,0662 | 0,01 | 5,818 |
| B2 | 864,9657 | 2106 | 9,7375 | 0,40 | 5,906 |
| B3 | 879,5996 | 2121 | 14,6339 | 0,10 | 6,252 |

Nous avons commencé l'analyse en estimant un modèle fondé sur l'hypothèse d'erreurs de classification indépendantes (ECI) (modèle A dans le tableau), qui, comme prévu, présente un mauvais ajustement.

Les modèles qui suivent sont fondés sur l'étude de Magnac et Visser (1995). Ces auteurs considèrent des transitions mensuelles au cours d'une période plus longue que la nôtre (de janvier 1989 à mars 1992), mais dans le même échantillon de personnes. Ils présupposent que la situation par rapport au marché du travail qui existe dans le mois de l'interview est correctement déclarée, tandis que la probabilité de commettre des erreurs augmente avec l'écart qui sépare le mois de référence et l'époque de l'interview, d'après une fonction déterministe du temps. Les probabilités de réponse sont présumées constantes tout au long des vagues de l'enquête, et l'on présuppose que les transitions réelles suivent une chaîne de Markov stationnaire de premier ordre. Notre modèle A1 est une version du modèle de Visser et Magnac assujettie à moins de contraintes (aucune hypothèse de stationnarité n'est émise) et qui est appliquée à des transitions trimestrielles relatives à la période comprise entre décembre 1990 et mars 1992. Notre modèle A2 ajoute au modèle A1 l'hypothèse de stationnarité de premier ordre au niveau latent. Ces modèles donnent tous deux un rendement assez mauvais et (d'après la colonne ME) nous constatons que, en moyenne, ils corrigent le marché du travail observé vers la stabilité, ce qui est un résultat qui est contraire aux preuves relatives aux effets des erreurs de classification dans les enquêtes rétrospectives.

Le modèle B introduit la corrélation entre les erreurs de classification en laissant dépendre chaque indicateur de la transition réelle qui a eu lieu entre les époques $t$ et $t + 1$; en outre, il inclut la contrainte (a). La qualité de l'ajustement augmente de manière frappante (voir $L^2$). Tous les modèles subséquents sont subordonnés au modèle B et des restrictions supplémentaires peuvent être évaluées par un test conditionnel. Le modèle B1 introduit des contraintes sous b), et le modèle B2 les contraintes supplémentaires sous c); le modèle B3 est notre modèle final.

Le tableau 6 présente les taux de transition estimés à l'aide de notre modèle le mieux ajusté. Le marché du travail français est corrigé dans le sens d'une plus grande mobilité. La mobilité moyenne estimée est de 6,252 %. En outre, les probabilités de réponses estimées montrent un patron cohérent par rapport à l'idée selon laquelle la probabilité de commettre des erreurs augmente avec la longueur de la période pour laquelle le répondant doit faire appel à sa mémoire.

## REMERCIEMENTS

Les recherches pour le présent article ont été financées par les subventions CNR n° 94.02242.CT10 et MURST n° 02.09.02.110 et n. 02.09.02.124. Nous remercions Michael Visser et Thierry Magnac de nous avoir fourni des données anonymées tirées de l'enquête française sur la population active. Une version préliminaire du présent article a été présentée lors de la Réunion satellite de l'AISE/AISO sur les études longitudinales, qui a eu lieu à Jérusalem, du 27 au 31 août 1997, où nous avons bénéficié de commentaires et de discussions. Nous remercions particulièrement deux examinateurs pour leurs critiques fouillées et leurs suggestions.

**Tableau 6**
Taux de transition trimestriels observés (×100), EFPA, décembre 1990 à mars 1992
(ME = mobilité estimée)

| | OO | OC | OH | CO | CC | CH | HO | HC | HH | ME |
|---|---|---|---|---|---|---|---|---|---|---|
| D90-M91 | 94,85 | 4,48 | 0,67 | 12,70 | 66,28 | 21,02 | 1,09 | 1,55 | 97,36 | 6,27 |
| M91-J91 | 95,65 | 1,37 | 2,98 | 28,43 | 62,35 | 9,22 | 3,61 | 1,48 | 94,91 | 7,49 |
| J91-S91 | 93,71 | 4,25 | 2,04 | 14,88 | 82,50 | 2,62 | 4,11 | 3,49 | 92,40 | 7,70 |
| S91-D91 | 98,32 | 1,67 | 0,01 | 15,42 | 83,75 | 0,83 | 3,80 | 0,47 | 95,73 | 4,24 |
| D91-M92 | 93,23 | 5,02 | 1,75 | 9,99 | 88,65 | 1,36 | 2,07 | 1,28 | 96,65 | 5,56 |

## ANNEXE A

**Spécification finale de modèle pour les données de la SIPP, en ce qui a trait aux probabilités conditionnelles**

(1) Décomposition de modèle de base

$$z_g = P(G = g) \tag{A1}$$

$$\pi_1^{j_1} = P(y_1 = j_1) \tag{A2}$$

$$\pi_t^{j_t j_{t-1}} = P(y_t = j_t \mid y_{t-1} = j_{t-1}) \qquad t = 2, 3, 4 \tag{A3}$$

$$q_{yt}^{l_t j_t l_{t+1} j_{t+1} g}$$

$$= P(Y_t = l_t \mid y_t = j_t, Y_{t+1} = l_{t+1}, y_{t+1} = j_{t+1}, G = g)$$

$$t = 1, 2, 3 \tag{A4}$$

$$q_{y4}^{l_4 j_4 g} = P(Y_4 = l_4 \mid y_4 = j_4, G = g) \tag{A5}$$

$$q_{wt}^{m_t j_t m_{t+1} j_{t+1} g}$$

$$= P(W_t = m_t \mid y_t = j_t, W_{t+1} = m_{t+1}, y_{t+1} = j_{t+1}, G = g)$$

$$t = 1, 2, 3 \tag{A6}$$

$$q_{w4}^{m_4 j_4 g} = P(W_4 = m_4 \mid y_4 = j_4, G = g) \tag{A7}$$

$g$ varie au cours de 1, 2, 3 et 4; $l_t$ et $j_t$, $t = 1, 2, 3, 4$, varient dans le cas des catégories O, C et H, $m_t$, $t = 1, 2, 3, 4$ varient dans le cas des catégories «emploi» et «pas d'emploi».

(2) Contraintes imposées aux probabilités conditionnelles

$$q_{yt}^{l_{t+1} j_t l_{t+1} j_{t+1} g}$$

$$= P(Y_t = l_{t+1} \mid y_t = j_t, y_{t+1} = l_{t+1}, y_{t+1} = j_{t+1}, G = g) = 1$$

$$\text{lorsque } l_{t+1} \neq j_{t+1}\ t = 1, 2, 3 \tag{A8}$$

$$q_{yt}^{l_t j_t l_{t+1} j_{t+1} g} = P(Y_t = l_t \mid y_t = j_t, G = g)$$

$$\text{lorsque } l_{t+1} = j_{t+1} \text{ et } t = 1, 2, 3 \tag{A9}$$

$$q_{wt}^{m_{t+1} j_t m_{t+1} j_{t+1} g}$$

$$= P(W_t = m_{t+1} \mid y_t = j_t, W_{t+1} = m_{t+1}, y_{t+1} = j_{t+1}, G = g) = 1$$

$$\text{lorsque } m_{t+1} \neq j_{t+1} \text{ et } t = 1, 2, 3 \tag{A10}$$

$$q_{wt}^{m_t j_t m_{t+1} j_{t+1} g} = P(W_t = m_t \mid y_t = j_t, G = g)$$

$$\text{lorsque } m_{t+1} = j_{t+1} \text{ et } t = 1, 2, 3 \tag{A11}$$

$$q_{t=4}^{g=1} = q_{t=3}^{g=4} = q_{t=2}^{g=3} = q_{t=1}^{g=2} \tag{A12}$$

$$q_{t=4}^{g=2} = q_{t=3}^{g=1} = q_{t=2}^{g=4} = q_{t=1}^{g=3} \tag{A13}$$

$$q_{t=4}^{g=3} = q_{t=3}^{g=2} = q_{t=2}^{g=1} = q_{t=1}^{g=4} \tag{A14}$$

$$q_{t=4}^{g=4} = q_{t=3}^{g=3} = q_{t=2}^{g=2} = q_{t=1}^{g=1} \tag{A15}$$

## ANNEXE B

**Spécification finale de modèle pour les données de la SIPP, en ce qui a trait à la décomposition de modèle de base et au patron des restrictions imposées aux paramètres**

1) Décomposition de modèle de base

$$\pi_t^{j_1} = P(y_1 = j_1) \tag{B1}$$

$$\pi_t^{j_t j_{t-1}} = P(y_t = j_t \mid y_{t-1} = j_{t-1})$$

$$t = 2, 3, 4, 5 \tag{B2}$$

$$q_{yt}^{l_t j_t j_{t+1}} = P(Y_t = l_t \mid y_t = j_t, y_{t+1} = j_{t+1})$$

$$t = 2, 3, 4, 5 \tag{B3}$$

$$q_{y6}^{l_6 j_6} = P(Y_6 = l_6 \mid y_6 = j_6) \tag{B4}$$

$$q_{wt}^{m_1 j_1 j_2} = P(W_1 = m_1 \mid y_1 = j_1, y_2 = j_2) \tag{B5}$$

$$q_{w2}^{m_2 j_2} = P(W_2 = m_2 \mid y_2 = j_2) \tag{B6}$$

$j_t$, $l_t$ et $m_t$ dans le cas de O, C et H

2) Patron des restrictions imposées aux probabilités de réponse

**Tableau B.1**
Mois d'observation

| Probabilité d'observer un état étant donnée une transition latente | Mars 91 | Juin 91 & Sept. 91 | Déc. 90 & Déc. 91 |
|---|---|---|---|
| O\|oo | 1 | 1 | 1 |
| C\|oo | 2 | 2 | 2 |
| H\|oo | 3 | 3 | 3 |
| O\|oc | L | L | L |
| C\|oc | L | L | L |
| H\|oc | * | * | * |
| O\|oh | L | L | L |
| C\|oh | * | * | * |
| H\|oh | L | L | L |
| O\|co | L | L | L |
| C\|co | L | L | L |
| H\|co | * | * | * |
| O\|cc | 4 | 4 | 4 |
| C\|cc | 5 | 5 | 5 |
| H\|cc | 6 | 6 | 6 |
| O\|ch | * | * | * |
| C\|ch | L | L | L |
| H\|ch | L | L | L |
| O\|ho | L | L | L |
| C\|ho | L | L | L |
| H\|ho | L | L | L |
| O\|hc | * | * | * |
| C\|hc | L | L | L |
| H\|hc | L | L | L |
| O\|hh | 7 | 7 | 7 |
| C\|hh | 8 | 8 | 8 |
| H\|hh | 9 | 9 | 9 |

Des chiffres égaux indiquent que les probabilités de réponse sont considérées comme étant égales.
* indique une probabilité établie à 0.
L désigne un paramètre libre.

## BIBLIOGRAPHIE


ABOWD, J.M., et ZELLNER A. (1985). Estimating gross labour force flows. *Journal of Business and Economics Statistics*, 3, 254-283.

BASSI, F., CROON, M., HAGENAARS, J., et VERMUNT, J. (1995). Estimating Latent Turnover Tables When Data are Affected by Correlated and Uncorrelated Classification Errors. WORC PAPER 95.12.25/7, Tilburg University.

BERNARD, H.R., KILLWORTH, P., KRONENFELD, D., et SAILER, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13, 495-517.

CHAKRABARTRY, R.P., et WILLIAMS, T.R. (1989). Time-in-sample biases in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 309-314.

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. (1993). *The Future of the SIPP*. Washington, D.C.: National Academy Press.

DUNCAN, G., et KALTON, G. (1987). Issues of design and analysis of surveys across time. Revue Internationale de Statistique, 55, 97-117.

GOODMAN, L.A. (1973). The analysis of a multidimensional contingency table when some variables are posterior to the others. *Biometrika*, 60, 179-192.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 79, 1178-1259.

GOODMAN, L.A. (1981). Three elementary views of log-linear models for the analysis of cross-classifications having ordered categories. Dans *Sociological Methodology*, (Éd. S.Leinhardt), 193-293. San Francisco: Jossey Bass.

HABERMAN, S.J. (1978). *Analysis of Qualitative Data. Vol.1. Introductory Topics*. New York: Academic Press.

HAGENAARS, J.A. (1988). Latent structure models with direct effects between the indicators, local dependence models. *Sociological Methods and Research*, 16, 379-405.

HAGENAARS, J.A. (1990). *Categorical Longitudinal Data: Log-linear, Panel, Trend and Cohort Analysis*. Newbury Park: Sage.

HAGENAARS, J.A. (1997). Categorical Causal Modeling: Directed Loglinear Models With Latent Variables. WORC PAPER 97.04.002/7, Tilburg University.

HUBBLE, D.L., et JUDKINS, D.R. (1989). Measuring the Bias in Gross Flows in the Presence of Autocorrelated Response Errors. SIPP Document de travail no. 8712, U.S. Bureau of the Census.

HU.S.SMAN, R., MEHRAN, F., et VERMA, M. (1990). *Surveys of Economically Active Population, Employment and Underemployment: an ILO Manual on Concepts and Definitions*. Geneva: ILO.

JORESKÖG, K.G., et SÖRBOM, D. (1988). *Lisrel 7: A Guide to the Program and Applications*. Chicago: SPSS INC.

KALTON, G., et CITRO, C.F. (1993). Enquête par panel: ajout d'une quatrième dimension. *Techniques d'enquête*, 19, 217-227.

KALTON, G., et MILLER, M.W. (1991). The seam effect with social security income in the SIPP. *Journal of Official Statistics*, 7, 235-245.

LANGEHEINE, R., PANNEKOEK, J., et van de POL, F. (1995). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, 24, 492-516.

LAZARSFELD, P.F., et HENRY, N.W. (1968). *Latent Structure Analysis*. New York: Houghton Mufflin.

MAGNAC, T., et VISSER, M. (1995). Transition Models With Measurement Errors. Document de travail, Institut National de la Recherche Agronomique (INRA), Paris.

MARQUIS, K.H., et MOORE, J.C. (1989). Some response errors in SIPP with thoughts about their effects and remedies. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 381-386.

MARTINI, A. (1988). Retrospective versus panel data in estimating labour force gross flows: comparing SIPP and CPS. *Proceedings of the Social Science Section, American Statistical Association*, 109-114.

MARTINI, A. (1989). Seam effect, recall bias, and the estimation of labour force transition rates from SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 387-392.

MENG, X.L., et RUBIN, D.B. (1993). Maximum likelihood estimation via ECM algorithm: a general framework. *Biometrika*, 80, 267-278.

McCORMICK, M., BUTLER, D., et SINGH, R. (1992). Investigating time-in-sample effects for the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 554-559.

O'MUIRCHEARTAIGH, C. (1996). Measurement errors in panel surveys: implications for survey design and for survey instruments. *Proceedings of the Scientific Reunion of the Italian Statistical Society*, 1, 207-218. Rimini: Maggioli.

PFEFFERMAN, D., SKINNER, C.J., et HUMPHREYS, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, Series A, 161, 13-32.

POTERBA, J.M., et SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.

SINGH, A.C., et RAO, J.N.K. (1995). On the adjustment of gross flows estimates for classification errors with application to data from the Canadian Labor Force Survey. *Journal of the American Statistical Association*, 90, 1-11.

SKINNER, C.J., et TORELLI, N. (1993). Measurement error and the estimation of gross flows from longitudinal economic data. *Statistica*, 3, 391-405.

U.S. DEPARTMENT OF COMMERCE (1991). *SIPP User's Guide*. Washington D.C.

van de POL, F., et LANGEHEINE, R. (1990). Mixed Markov latent class models. Dans *Sociological Methodology*, (Éd. C.Clogg), 213-247. Oxford: Blackwell.

van de POL, F., et LANGEHEINE, R. (1992). Analysing Measurement Error in Quasi-experimental Data: An Application of Latent Class Models to Labour Market Data. Document de travail de European Scientific Network on household Panel Studies, 57, Colchester, University of Essex.

van de POL, F., et LANGEHEINE, R. (1997). Separating change and measurement error in panel surveys with an application to labour market data. Dans *Survey Measurement and Process Quality*, (Éd. L. Lyberg *et al.*), 671-688. New York: Wiley.

VERMUNT, J.K. (1993). Log-linear and event history analysis with missing data using the EM algorithm. WORC PAPER 93.09.015/7, Tilburg University.

YOUNG, N. (1989). Wave seam effects in the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 393-398.

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 133-140
Statistique Canada



# Estimation des flux bruts de la population active provenant d'enquêtes donnant lieu à une non-réponse dont il faut tenir compte au niveau du ménage

PAUL S. CLARKE et RAY L. CHAMBERS[1]

RÉSUMÉ

La mesure des flux bruts de la population active est un objectif important des enquêtes continues sur la population active effectuées par un grand nombre d'offices nationaux de la statistique. Cependant, il est bien connu que l'estimation de ces flux peut être compliquée par une non-réponse, des erreurs de mesure, un renouvellement de l'échantillon et des effets complexes du plan de sondage. Le présent article, inspiré par des modèles de non-réponse dans les enquêtes sur les ménages, porte sur l'estimation des flux bruts tout en apportant des ajustements en fonction de la non-réponse dont il faut tenir compte. Les approches antérieures basées sur un modèle en ce qui concerne l'estimation des flux bruts supposaient que la non-réponse était un processus au niveau de la personne. Nous proposons une catégorie de modèles qui permettent une non-réponse dont il faut tenir compte au niveau du ménage. On a recours à une étude en simulation pour démontrer que les estimations des flux bruts de la population active au niveau de la personne provenant des données d'enquêtes sur les ménages peuvent être biaisées et que les estimations en fonction de modèles au niveau du ménage peuvent permettre de réduire ce biais.

MOTS CLÉS: Flux bruts; enquêtes sur les ménages; non-réponse dont il faut tenir compte.

## 1. INTRODUCTION

On définit de façon générale les flux bruts de la population active comme étant des transitions dans le temps entre les trois grands états de la population active, personnes occupées, personnes sans emploi et personnes économiquement inactives. Les estimations des flux bruts sont un outil important dans l'étude de la dynamique de la population active (par exemple, voir Vanski 1985). Les enquêtes continues à grande échelle telle la Labour Force Survey en Grande-Bretagne et la Current Population Survey aux États-Unis fournissent des données pour l'estimation des flux bruts. Cependant, la non-réponse, l'erreur de mesure, le renouvellement de l'échantillon et les effets complexes du plan de sondage ont une incidence sur l'estimation des flux bruts de ces enquêtes. Hogue (1985) traite de ces facteurs et d'autres qui ont une incidence sur l'estimation des flux bruts. Dans le présent article, nous nous concentrons sur le problème de la non-réponse.

Nous supposons qu'un mécanisme de non-réponse a pour conséquence que les données observées sont incomplètes. Si la probabilité de la non-réponse dépend des données manquantes, alors le mécanisme de non-réponse est un mécanisme dont il faut tenir compte (Rubin 1976). L'approche basée sur des modèles pour analyser des données incomplètes d'enquête est détaillée dans Little (1982). Les approches basées sur des modèles relativement à l'estimation des flux bruts de la population active font intervenir la modélisation des flux de la population active et du mécanisme de non-réponse tout en assurant l'ajustement des deux modèles aux données incomplètes. Des exemples de ces modèles sont données dans Stasny et Fienberg (1985), Stasny (1986) et, dans le cas de la non-réponse dont il faut tenir compte, dans Little (1985). Nous disons de ces modèles qu'ils sont au niveau de la personne parce que les personnes sont modélisées comme répondant ou ne répondant pas, indépendamment des autres personnes échantillonnées.

Tant la Labour Force Survey que la Current Population Survey sont des exemples d'enquêtes sur les ménages, c'est-à-dire des enquêtes se fondant sur un échantillon aléatoire des ménages plutôt que sur des personnes. Les enquêtes sur les ménages peuvent donner lieu à un comportement corrélé de non-réponse au sein des ménages. Par exemple, dans la Current Population Survey, un seul membre du ménage (habituellement le chef du ménage) agit en tant que représentant des membres du ménage; ainsi, si le membre choisi du ménage ne répond pas, les autres membres du ménage ne répondent pas non plus. Il s'ensuit que, en raison du comportement corrélé de non-réponse au sein du ménage, les modèles de non-réponse au niveau de la personne ne conviennent pas à l'estimation des flux bruts de la population active qui utilise des données d'enquêtes sur les ménages.

Dans le présent article, nous proposons une catégorie de modèles pour les flux de la population active au niveau de la personne et une non-réponse au niveau du ménage qui tient compte du comportement corrélé de non-réponse au sein du ménage. On présente également un certain nombre de modèles plausibles de non-réponse qui sont estimables à partir des données observées, tant pour ce qui est de la non-réponse dont on n'a pas à tenir compte que pour la non-

[1] Paul S. Clarke et Ray L. Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom.

réponse dont il faut tenir compte. Nous simulons alors des données d'enquêtes sur les ménages à l'aide de ces modèles au niveau du ménage afin de démontrer l'utilité éventuelle de notre approche: premièrement, on démontre que les estimations des flux bruts de la population active au niveau de la personne sont biaisées lorsqu'elles sont ajustées en fonction des données d'enquêtes sur les ménages; deuxièmement, on compare le biais des estimations des flux bruts au niveau de la personne à celui au niveau du ménage afin de démontrer les avantages d'ajuster les modèles au niveau des ménages aux données d'enquêtes sur les ménages. Enfin, nous résumons les conclusions de nos études en simulation et nous présentons des orientations pour des études ultérieures.

## 2. UN MODÈLE POUR LA NON-RÉPONSE AU NIVEAU DU MÉNAGE

### 2.1 Les données

Un flux brut est la probabilité ou la fréquence de personnes au sein de la population qui font une transition d'états entre deux points dans le temps, $t_1$ et $t_2 (t_1 < t_2)$. Les flux bruts de la population active renvoient aux transitions entre les trois grands états de la population active: 1 = «personne occupée», 2 = «personne sans emploi» et 3 = «hors de la population active», la dernière catégorie faisant référence aux personnes économiquement inactives telles les personnes à la retraite et les étudiants. Soit $S$ qui représente un échantillon aléatoire simple de ménages, indexé par $h$. Au sein du ménage $h$, il y a $n_h$ personnes admissibles, dont $n_h(ab)$ ont un flux de population active $(a, b)$ entre $t_1$ et $t_2$, où $\sum_{a,b} n_h(ab) = n_h$ et $a, b = 1, 2, 3$. Nous disons que $\{n_h(ab)\}$ sont les données complètes, c'est-à-dire les fréquences qui seraient observées en l'absence de non-réponse.

Le tableau 1 illustre les données des flux de la population active complètes pour le ménage $h$ comme étant un tableau de contingence de $3 \times 3$. Si $h$ répond les deux fois, les données observées sont les cellules de ce tableau à double entrée. Cependant, si le ménage ne répond pas à $t_1$ ou $t_2$, les données observées correspondent aux marges du tableau: $n_h(1+), n_h(2+), n_h(3+)$ sont les données observées si $h$ répond à $t_1$ mais ne répond pas à $t_2$; et $n_h(+1)$, $n_h(+2), n_h(+3)$ sont les données observées si $h$ répond à $t_2$ mais ne répond pas à $t_1$. (Un indice remplacé par «+» représente la somme de tous les niveaux de cet indice.) En outre, si $h$ ne répond pas à $t_1$ et $t_2$, les données observées sont la taille du ménage, $n_h$, que nous supposons connues et fixes entre $t_1$ et $t_2$.

**Tableau 1**
Données complètes des flux de la population active pour le ménage $h$

| État | | $t_2$ 1 | 2 | 3 | |
|---|---|---|---|---|---|
| $t_1$ | 1 | $n_h(11)$ | $n_h(12)$ | $n_h(13)$ | $n_h(1+)$ |
| | 2 | $n_h(21)$ | $n_h(22)$ | $n_h(23)$ | $n_h(2+)$ |
| | 3 | $n_h(31)$ | $n_h(32)$ | $n_h(33)$ | $n_h(3+)$ |
| | | $n_h(+1)$ | $n_h(+2)$ | $n_h(+3)$ | $n_h$ |

### 2.2 Spécification du modèle

Il ne convient pas de traiter le comportement de non-réponse de personnes au sein des ménages comme étant indépendant dans les enquêtes sur les ménages. Dans la Labour Force Survey, par exemple, un membre admissible du ménage détermine si le ménage peut être interviewé. Par conséquent, s'il n'y a aucune personne admissible que l'on peut contacter, chaque personne du ménage ne répond pas. Pour construire un modèle de la non-réponse au niveau du ménage, nous prenons les idées à l'origine de la non-réponse au niveau de la personne et nous les étendons au ménage en considérant qu'un ménage est une entité ayant son propre flux de non-réponse entre $t_1$ et $t_2$. Pour permettre une non-réponse dont il faut tenir compte, la probabilité d'un flux de non-réponse au sein d'un ménage est modélisée en tant qu'une fonction de ses flux individuels de la population active, que nous devons maintenant décrire.

Soit $\boldsymbol{N}_h = (N_h(11), N_h(21), ..., N_h(33))$ le vecteur aléatoire des fréquences des flux de la population active pour le ménage $h$, où $N_h(ab)$ est la variable aléatoire dont l'extrant correspond au nombre de personnes du flux de la population active $(a, b), a, b = 1, 2, 3$. En outre, désignons le vecteur aléatoire pour le flux de non-réponse du ménage $h$ par $\boldsymbol{R}_h = (R_{h1}, R_{h2})$, où

$$R_{hj} = \begin{cases} 1, & \text{si ménage répond à } t_j \\ 0, & \text{sinon} \end{cases}$$

est la variable aléatoire de l'état de non-réponse pour $h$ à $t_j$, $j = 1, 2$. Les réalisations de ces quantités aléatoires sont désignées par $\boldsymbol{n}_h$ et $\boldsymbol{r}_h$. Nous supposons maintenant que $\boldsymbol{n}_h$ et $\boldsymbol{r}_h$ sont connus et formulons la probabilité conjointe de $\boldsymbol{N}_h$ et $\boldsymbol{R}_h$ comme

$$\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h, \boldsymbol{R}_h = \boldsymbol{r}_h) = \Pr(\boldsymbol{N}_h = \boldsymbol{n}_h)\Pr(\boldsymbol{R}_h = \boldsymbol{r}_h \mid \boldsymbol{N}_h = \boldsymbol{n}_h),$$

où $\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h)$ est le modèle des flux de la population active et $\Pr(\boldsymbol{R}_h = \boldsymbol{r}_h \mid \boldsymbol{N}_h = \boldsymbol{n}_h)$ est appelé le modèle des flux de non-réponse.

Le modèle des flux de la population active est multinomial avec une fonction de probabilité

$$\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h; \omega) = n_h! \prod_{a,b} \frac{\omega(ab)^{n_h(ab)}}{n_h(ab)!}, \quad (1)$$

où $\omega(ab) > 0$ est la probabilité pour une personne d'avoir un flux de la population active $(a, b)$ et $\sum_{a,b}\omega(ab) = 1$. Le vecteur des paramètres des flux de la population active est désigné par $\omega = (\omega(11), \omega(21), ..., \omega(33))$, dont 8 sont libres. L'hypothèse de l'échantillonnage multinomial dans (1) sous-entend que le comportement des flux de la population active des personnes est indépendant au sein des ménages et que les ménages sont homogènes pour ce qui est du comportement de leurs flux de la population active. Ces hypothèses ne sont pas réalistes, mais (1) peut facilement être étendu à un modèle plus réaliste des flux de la population active, tel qu'il est indiqué à la section 4.

La probabilité pour le ménage $h$ d'avoir un flux de non-réponse $(u, v)$ est

$$\pi(uv \mid \boldsymbol{n}_h) = \Pr(\boldsymbol{R}_h = (u, v) \mid \boldsymbol{N}_h = \boldsymbol{n}_h; \psi) = \frac{1}{n_h} \sum_{a,b} n_h(ab)\psi(uv \mid ab), \quad (2)$$

lorsque $u, v = 0, 1$, à savoir une moyenne pondérée des paramètres du modèle de non-réponse. En établissant $n_h = 1$, on peut constater que $\psi(uv \mid ab) > 0$ est la probabilité d'un ménage d'une taille un (c'est-à-dire une personne) qui a un flux de non-réponse $(u, v)$ s'il a un flux de population active $(a, b)$. Ainsi, $\sum_{u,v}\psi(uv \mid ab) = 1$ et $\psi = (\psi(11 \mid 11), \psi(01 \mid 11), ..., \psi(00 \mid 33))$ est le vecteur des paramètres de non-réponse, dont 27 sont libres.

Avant de définir la fonction de vraisemblance des données complètes, partageons $S$ en 4 sous-ensembles exhaustifs et mutuellement exclusifs

$$S = S_{11} \cup S_{01} \cup S_{10} \cup S_{00},$$

où $S_{uv} = \{h : \boldsymbol{r}_h = (u, v)\}$ est le sous-ensemble des ménages ayant un flux de non-réponse $(u, v)$. Ainsi, étant donné que $S$ est un échantillon aléatoire simple des ménages, la fonction de vraisemblance des données complètes est

$$L(\omega, \psi; \{\boldsymbol{n}_h, \boldsymbol{r}_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} L_h(\omega, \psi; \boldsymbol{n}_h, (u, v)), \quad (3)$$

où $L_h(\omega, \psi; \boldsymbol{n}_h, (u, v))$ est la contribution du ménage $h \in S_{uv}$ à la vraisemblance, le produit de (1) et (2).

### 2.3 Ajustement des modèles

#### 2.3.1 Estimation de vraisemblance maximale

Étant donné que les données complètes ne sont pas disponibles, (3) doit être modifié de façon à donner la vraisemblance fondée sur les données observées. Représentons les données observées par $\{\boldsymbol{n}_h^*\}$. Tel qu'on l'a indiqué à la section 2.1, les données observées pour les ménages qui répondent à $t_1$ et $t_2$ sont la classification croisée complète du tableau 1, à savoir $\boldsymbol{n}_h^* = \boldsymbol{n}_h$. De même, si $h \in S_{10}$ alors $\boldsymbol{n}_h^* = (n_h(1+), n_h(2+), n_h(3+))$; si $h \in S_{01}$ alors $\boldsymbol{n}_h^* = (n_h(+1), n_h(+2), n_h(+3))$; et si $h \in S_{00}$ alors $\boldsymbol{n}_h^* = n_h$.

On obtient la contribution du ménage $h \in S_{uv}$ à la vraisemblance des données observées en faisant le total de $L_h(\omega, \psi; \boldsymbol{n}_h, (u, v))$ pour toutes les valeurs possibles que la classification croisée complète $3 \times 3$ des flux de la population active peut prendre compte tenu de la marge observée. Représentant cet ensemble de tableaux par $\boldsymbol{n}_h : \boldsymbol{n}_h^*$, la vraisemblance des données observées pour $S$ est

$$L(\omega, \psi; \{\boldsymbol{n}_h^*, \boldsymbol{r}_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} \sum_{\boldsymbol{n}_h : \boldsymbol{n}_h^*} L_h(\omega, \psi; \boldsymbol{n}_h, (u, v)). \quad (4)$$

L'ajustement du modèle nécessite le calcul de (4) à chaque étape d'un processus d'optimisation itérative. Sur le plan des calculs, ceci est exigent parce que la fonction de vraisemblance des données complètes doit être additionnée explicitement pour les données manquantes. Par exemple, les données observées pour $h \in S_{10}$ est $\boldsymbol{n}_h^* = (n_h(1+), n_h(2+), n_h(3+))$ et la contribution de vraisemblance de ce ménage à la vraisemblance des données observées est

$$\sum_{\boldsymbol{n}_h : \boldsymbol{n}_h^*} L_h(\omega, \psi; \boldsymbol{n}_h, (1, 0)).$$

Pour calculer explicitement cette contribution, chaque tableau des données complètes $3 \times 3$ $\boldsymbol{n}_h$ pour $\boldsymbol{n}_h^*$ fixe est produit et $L_h(\omega, \psi; \boldsymbol{n}_h, (1, 0))$ est évaluée pour chacun. Dans le cas d'un ménage de taille $n_h = 5$, il y a au moins 21 tableaux possibles et au plus 108 tableaux possibles, selon les valeurs de la marge fixe; lorsque $n_h = 15$, un ménage d'une très grande taille, les nombres respectifs sont de 136 et de 9 261. Une procédure semblable est utilisée pour $h \in S_{01}$, sauf qu'ici $\boldsymbol{n}_h^* = (n_h(+1), n_h(+2), n_h(+3))$ est la marge fixe. Si $h \in S_{00}$ alors aucune donnée n'est observée au sujet de la situation vis-à-vis de la population active, seulement la taille du ménage $n_h$. Donc chaque tableau $3 \times 3$ ayant le total $n_h$ doit être produit et la fonction de vraisemblance être calculée pour chacun: lorsque $n_h = 5$ il y a 1 287 tableaux et lorsque $n_h = 15$ il y a 490 314 tableaux. Il est impossible, en ce qui concerne le temps d'exécution machine, de calculer ces sommes directement. Le nombre de calculs explicites peut être réduit si l'on reconnaît que chaque ménage est défini uniquement par ses fréquences des flux de la population active observés et son flux de non-réponse. Ainsi, la somme des données manquantes ne doit être effectuée qu'une fois pour un ménage qui a des fréquences de flux de la population active et un flux de non-réponse donnés; la contribution de ce ménage à la vraisemblance est alors élevée à la puissance du nombre de ménages définis de façon semblable dans $S$.

### 2.3.2 Estimabilité des paramètres

Si nous posons $n_h = 1$ pour tous les $h$, les données complètes n'ont aucune structure de ménage et constituent un tableau à entrée quadruple recoupé par la situation vis-à-vis de la population active et l'état de non-réponse à $t_1$ et $t_2$. Le logarithme du rapport de vraisemblance (4) des données observées est maintenant l'équivalent de celui des modèles au niveau de la personne dans Stasny et Fienberg (1985), Little (1985) et Stasny (1986). Pour ces modèles, l'estimabilité nécessite que le nombre de paramètres des modèles ne soit pas supérieur à 15 (un pour chaque cellule de tableau observée moins une pour la contrainte de l'échantillonnage multinomial). En conséquence, $(\omega, \psi)$ sont impossibles à estimer parce qu'il y a $8 + 27 = 35$ paramètres libres. Étant donné que l'intérêt est concentré sur les probabilités des flux bruts de la population active, $\omega$, il est nécessaire de limiter $\psi$ afin d'assurer l'estimabilité.

Lorsque $n_h > 1$, déterminer l'estimabilité des paramètres est plus difficile parce que (4) a une expression en forme analytique fermée compliquée. Fitzmaurice, Laird et Zahner (1996) utilisent une méthode numérique pour déterminer l'estimabilité qui implique une démonstration que la matrice d'information est régulière dans le voisinage de l'estimation de vraisemblance maximale. Cependant, ce n'est non seulement peu pratique pour les problèmes d'une dimension élevée, mais évaluer la matrice d'information pour le modèle au niveau des ménages est particulièrement difficile dans ce cas. Au lieu, nous adoptons une approche pragmatique pour déterminer l'estimabilité des paramètres: tout d'abord, nous restreignons l'attention aux modèles qui remplissent la condition nécessaire pour l'estimabilité lorsque $n_h = 1$; et deuxièmement, on utilise différentes valeurs de départ pour chaque ajustement. Si les valeurs de départ différentes révèlent une estimation de vraisemblance maximale non unique, ou si les estimations des paramètres demeurent inchangées par rapport à leurs valeurs de départ, alors on suppose que les paramètres du modèle ne peuvent être estimés.

## 2.4 Modèles de non-réponse

Pour permettre d'obtenir les estimations des paramètres à partir des données observées, $\theta$ et $\psi$ doivent être limités conformément aux hypothèses au sujet du mécanisme de non-réponse. Les paramètres de non-réponse sont interprétés comme des probabilités de non-réponse individuelles, mais en ce qui concerne le cadre de ménages établi jusqu'à maintenant, il n'est pas approprié de parler de personnes qui ne répondent pas. Cependant, en réalité ce sont les personnes au sein des ménages qui déterminent un flux de non-réponse d'un ménage, et non le ménage lui-même. Par conséquent, les contraintes sont placées sur les paramètres de non-réponse au niveau de la personne qui s'appliquent au niveau du ménage par la dépendance fonctionnelle de $\pi(uv \mid \boldsymbol{n}_h)$ sur $\psi$ dans (2). Par exemple, si les paramètres de non-réponse sont limités de sorte que $\psi(uv \mid ab) = \psi(uv)$ pour la totalité de $a,b$, alors on n'a pas à tenir compte du mécanisme de non-réponse du ménage parce que les flux de non-réponse sont indépendants des flux de la population active.

Nous présentons maintenant quatre modèles pour le mécanisme de non-réponse, deux dont on doit tenir compte et deux dont on n'a pas à tenir compte.

- Modèles dont on n'a pas à tenir compte.
  - Modèle $I_A$: Probabilité de non-réponse constante,

    $$\psi(uv \mid ab) = \lambda^{1-u}(1-\lambda)^u \times \lambda^{1-v}(1-\lambda)^v,$$

    qui a 1 paramètre, $\lambda$, la probabilité d'une personne qui ne répond pas;
  - Modèle $I_B$: Indépendant de la situation vis-à-vis de la population active, mais probabilités de non-réponse différentes à $t_1$ et $t_2$,

    $$\psi(uv \mid ab) = \lambda^{1-u}(1-\lambda)^u \times \theta^{1-v}(1-\theta)^v,$$

    qui a 2 paramètres, $\lambda, \theta$, soit les probabilités de non-réponse à $t_1$ et $t_2$, respectivement.
- Modèles dont il faut tenir compte.
  - Modèle $N_A$: Les distributions de non-réponse à $t_1$ et $t_2$ sont indépendantes, mais dépendent de la situation vis-à-vis de la population active à $t_1$ et $t_2$, respectivement,

    $$\psi(uv \mid ab) = \lambda(a)^{1-u}(1-\lambda(a))^u \times \theta(b)^{1-v}(1-\theta(b))^v$$

    qui a 6 paramètres, $\boldsymbol{\lambda} = (\lambda(1), \lambda(2), \lambda(3))$ et $\boldsymbol{\theta} = (\theta(1), \theta(2), \theta(3))$, où $\lambda(a)$ est la probabilité de non-réponse à $t_1$ si on a la situation vis-à-vis de la population active $a$ à $t_1$, et $\theta(b)$ à $t_2$ si on a la situation vis-à-vis de la population active $b$ à $t_2$;
  - Modèle $N_B$: Les distributions de non-réponse à $t_1$ et $t_2$ dépendent de la situation vis-à-vis de la population active à $t_1$ et $t_2$ respectivement, c.-à-d. un processus de Markov d'ordre un. Contrairement à $N_A$, les distributions de non-réponse à $t_1$ et $t_2$ sont dépendantes: si l'état de non-réponse à $t_1$ est 1, alors la distribution de non-réponse à $t_2$ est la même qu'à $t_1$; mais si l'état de non-réponse à $t_1$ est 0, les distributions de non-réponse sont distinctes,

    $$\psi(uv \mid ab) = \lambda(a)^{1-u}(1-\lambda(a))^u \times \begin{cases} \lambda(b)^{1-v}(1-\lambda(b))^v, & \text{if} \quad u = 1, \\ \theta(b)^{1-v}(1-\theta(b))^v, & \text{if} \quad u = 0, \end{cases}$$

lorsque $a, b = 1, 2, 3$ et $u, v = 0, 1$. En vertu du modèle $I_A$, il y a un total de $8 + 1 = 9$ paramètres libres, remplissant la condition nécessaire pour l'estimabilité d'un modèle au niveau de la personne. Les modèles $I_B$, $N_A$ et $N_B$ ont 10, 14 et 14 paramètres libres, respectivement, et remplissent donc aussi la condition nécessaire pour l'estimabilité.

## 3. ÉTUDE EN SIMULATION

### 3.1 Procédure de simulation

Nous avons eu recours à une étude en simulation pour examiner les conséquences de ne pas tenir compte de la structure du ménage dans le cas des données d'enquêtes sur les ménages pour comparer les estimations des flux bruts de la population active pour les modèles au niveau de la personne et au niveau du ménage. À cette fin, on a produit des données d'enquêtes sur les ménages à l'aide de l'échantillonnage de Monte Carlo. Chaque ensemble de données de l'échantillon se composait de 10 000 personnes réparties dans des ménages de taille $n_h = k$ pour la totalité des $h$. Au sein de chaque ménage, on a produit les flux de la population active à partir de (1) et on a produit le flux de non-réponse à partir de (2) en vertu d'un des modèles $N_A$ ou $N_B$. On a rendu les données incomplètes en regroupant chaque tableau des flux de la population active des données complètes pour qu'il soit uniforme avec le flux de non-réponse du ménage. Au total, 1 000 ensembles de données indépendantes ont été produits de cette façon.

Les paramètres de la population utilisés pour produire les flux de la population active sont illustrés dans le tableau suivant:

| | | b | | |
|---|---|---|---|---|
| $\omega(ab)$ | | 1 | 2 | 3 |
| | 1 | 0.43 | 0.245 | 0.035 |
| a | 2 | 0.02 | 0.16 | 0.01 |
| | 3 | 0.015 | 0.035 | 0.05 |

Il s'agit de toute évidence d'une population en récession étant donné que la probabilité de passer de personne employée à personne sans emploi est très grande ($\omega(12) = 0.245$). En vertu des modèles $N_A$ et $N_B$ les paramètres de la population sont

| | i | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| $\lambda(i)$ | 0.2 | 0.8 | 0.5 |
| $\theta(i)$ | 0.5 | 0.2 | 0.8 |

On doit remarquer que ces valeurs des paramètres ne représentent pas un comportement des flux de non-réponse réaliste, elles ont été choisies dans le but d'illustrer la présente méthodologie. Cependant, cela n'influe pas sur les conclusions générales de l'article, qui sont également pertinentes pour des valeurs réalistes des véritables probabilités de non-réponse.

### 3.2 Résultats de la simulation

On obtient les estimations pour les modèles au niveau de la personne en corrigeant (4) avec $n_h = 1$ pour chaque ensemble de données incomplètes. La figure 1 résume les distributions d'échantillonnage de l'estimation de la vraisemblance maximale au niveau de la personne de $\omega(12)$, $\hat{\omega}(12)$, pour les modèles de non-réponse $I_A$, $I_B$, $N_A$ et $N_B$ (les estimations pour les modèles dont on n'a pas à tenir compte $I_A$ et $I_B$ sont incluses parce que les deux donnent les mêmes estimations des flux de la population active). Les traits verticaux représentent les intervalles entre le percentile 2,5 et le percentile 97,5 de chaque distribution de l'échantillonnage de l'estimation, et les points en caractères gras représentent ses médianes. On obtient trois distributions pour chaque estimation au niveau de la personne: la distribution la plus à gauche survient lorsque la taille du ménage est $k = 1$, les données simulées n'ont aucune structure de ménage; et si on lit de gauche à droite, les deux distributions suivantes sont celles que l'on obtient lorsque la taille du ménage est $k = 2$ et $k = 5$, respectivement. Le trait vertical plein désigne la véritable probabilité de flux, $\omega(12) = 0{,}0245$. Le comportement de la distribution d'échantillonnage de $\hat{\omega}(12)$ dans la présente étude reflète celui des autres estimations des flux bruts de la population active.

La figure 1a) en résume les distributions d'échantillonnage lorsque $N_A$ est le véritable modèle. Si le modèle ajusté au niveau de la personne est $I_A$, $I_B$ ou $N_B$, les estimations des flux bruts de la population active ont d'importants biais, quelle que soit la taille du ménage. Comme on devrait s'y attendre, l'estimation de la médiane pour le modèle correct $N_A$ n'est pas biaisée si $k = 1$ et a un petit biais pour $k = 2$ et $k = 5$ (quoique ce biais soit plus faible dans le cas de $k = 5$ que dans celui de $k = 2$). La diminution du biais lorsque la valeur de $k$ augmente est également apparente dans le cas des estimations au niveau de la personne $I_A$, $I_B$ et $N_B$. Ce comportement n'est pas prévu étant donné qu'il semble naturel de s'attendre à ce que le biais des estimations au niveau de la personne augmente avec la taille du ménage. Les résultats sont légèrement différents dans la figure 1b) lorsque $N_B$ est vrai. Dans le présent cas, l'estimation pour le modèle $N_B$ au niveau de la personne devient plus biaisée à mesure que la valeur de $k$ augmente, mais le biais diminue pour les modèles mal spécifiés $I_A$, $I_B$ et $N_B$ au niveau de la personne. En outre, les estimations mal spécifiées pour $I_A$ et $I_B$ comportent un petit biais lorsqu'on les compare à celles du modèle mal spécifié $N_A$. On traite de ces résultats à la section 3.3.

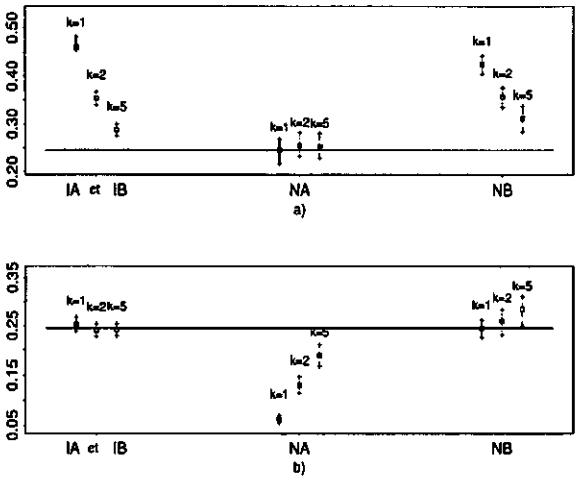



**Figure 1.** Distribution d'échantillonnage de $\hat{\omega}(12)$ pour les modèles au niveau de la personne, $I_A, I_B, N_A$ et $N_B$ lorsque le véritable modèle de non-réponse est a) $N_A$ et b) $N_B$ et la taille du ménage est $k = 1, 2, 5$.

Une comparaison des estimations de la médiane de $\omega(12)$ pour les modèles corrigés au niveau de la personne et au niveau du ménage lorsque $N_B$ est vrai est illustrée à la figure 2. Il y a quatre distributions d'échantillonnage associées à chaque modèle: les deux premières représentent celles dérivées d'un modèle de non-réponse au niveau de la personne et un modèle de non-réponse au niveau du ménage lorsque la taille du ménage est $k = 2$; et de même, les deux distributions suivantes sont celles que l'on obtient lorsque la taille du ménage est 5.

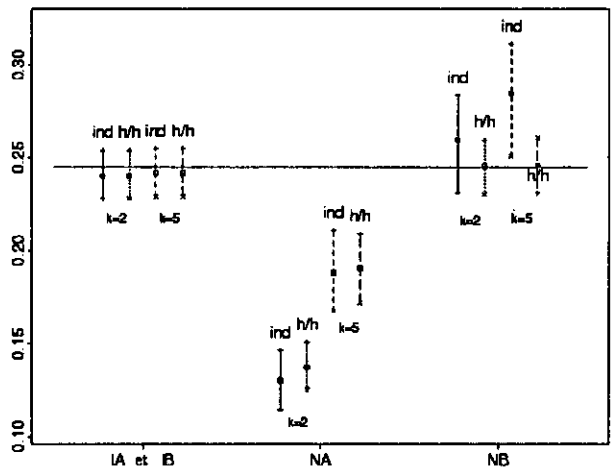



**Figure 2.** Les distributions d'échantillonnage de $\hat{\omega}$ (12) pour les modèles au niveau de la personne et au niveau du ménage $I_A, I_B, N_A$ et $N_B$ lorsque le véritable modèle de non-réponse est $N_B$ et la taille du ménage est $k = 2, 5$.

Pour une paire donnée de distributions d'échantillonnage au niveau de la personne et au niveau du ménage, on peut constater que l'estimation au niveau du ménage est moins biaisée que celle au niveau de la personne, et l'étendue de chaque distribution d'échantillonnage au niveau du ménage est plus faible. L'exception à cet énoncé est lorsque l'on corrige le modèle $I_A$, où les distributions au niveau de la personne et au niveau du ménage sont identiques. Cette égalité survient parce que la vraisemblance des données observées pour les modèles au niveau de la personne et au niveau du ménage sont équivalentes lorsque l'on n'a pas à tenir compte du modèle de non-réponse. Une autre caractéristique est que si le modèle de non-réponse est correctement spécifié, les estimations au niveau du ménage ne sont pas biaisées.

### 3.3 Résumé

Les estimations des flux bruts de la population active en vertu des modèles au niveau de la personne ne sont jamais moins biaisées que celles des modèles au niveau du ménage lorsqu'ils sont corrigés en fonction des données d'enquêtes sur les ménages dans notre étude. On doit remarquer que si l'on n'a pas à tenir compte du véritable modèle de non-réponse, il n'est pas nécessaire d'utiliser un modèle de non-réponse au niveau du ménage parce que les modèles au niveau de la personne et au niveau du ménage sont équivalents. Par exemple, si $I_A$ est vrai, (2) réduit à $\lambda^{u+v}(1-\lambda)^{1-u-v}$ et (4) se factorise en deux composantes dépendantes de $\omega$ seulement et $\lambda$ seulement; on peut démontrer que le facteur dépendant de $\omega$ est l'équivalent de celui du modèle au niveau de la personne et que par conséquent les estimations des flux de la population active sont les mêmes.

Il semble, à mesure qu'augmente la taille du ménage, que le biais des estimations des flux de la population active diminue si on doit tenir compte du modèle véritable. En fait, ce résultat survient parce que nous utilisons (1) pour produire les flux de la population active et non parce que les estimations du modèle ne sont pas biaisées pour une valeur importante $n_h$. Pour en connaître les raisons, prenons en considération le processus de formation du ménage utilisé pour produire chaque échantillon de Monte Carlo: à mesure que $n_h$ augmente, chaque fréquence du ménage tend vers la même valeur, c'est-à-dire, $n_h(ab)$ converge vers $n_h\omega(ab)$; par conséquent,

$$\pi(uv \mid \boldsymbol{n}_h) \to \frac{1}{n_h}\sum_{a,b} n_h\omega(a,b)\psi(uv \mid ab)$$

$$= \sum_{a,b} \omega(ab)\psi(uv \mid ab),$$

qui est indépendant de $\boldsymbol{n}_h$, c'est-à-dire que l'on n'a pas à tenir compte du mécanisme de non-réponse du ménage simulé. Par conséquent, les estimations des flux de la population active ne sont pas biaisées du fait que la correction des modèles dont il faut tenir compte en fonction des données simulées donne des estimations de paramètres qui sont conformes à la non-réponse dont on n'a pas à tenir

compte. Pour produire la non-réponse au niveau du ménage dont il faut tenir compte, il est nécessaire de prévenir $n_h(ab) \rightarrow n_h\omega(ab)$ en étendant (1) de façon à permettre des flux de la population active différentiels entre les ménages. Ces extensions du modèle des flux de la population active font l'objet de la section 4.

La figure 1b) illustre deux résultats anormaux qui contredisent l'explication ci-dessus lorsque $N_B$ est le modèle véritable. Tout d'abord, le biais de l'estimation du modèle $N_B$ au niveau de la personne augmente à mesure que $n_h$ augmente. Cependant, d'autres simulations avec un ménage d'une taille $n_h = 10$ ont révélé que le biais de l'estimation au niveau de la personne est zéro. Ainsi, la non-réponse asymptotique dont on n'a pas à tenir compte est également évidente lorsque $N_B$ est vrai, mais $n_h$ doit être importante avant que son incidence devienne apparente pour le modèle $N_B$. Deuxièmement, le biais des estimations du modèle dont on n'a pas à tenir compte au niveau de la personne est faible, presque zéro, lorsque $N_B$ est vrai. Ce faible biais diminue encore plus à mesure que $n_h$ augmente, conformément à l'ignorabilité asymptotique, mais nous devons encore parvenir à une explication satisfaisante quant à savoir pourquoi les modèles dont on peut ne pas tenir compte se comparent si bien dans cette situation. Des études plus approfondies sont nécessaires pour examiner cette constatation.

## 4. DISCUSSION

Dans les sections 3 et 4, on a démontré grâce à une étude en simulation que la modélisation de la non-réponse dont il faut tenir compte au niveau du ménage lors de l'estimation des flux bruts de la population active provenant d'enquêtes sur les ménages entraîne une réduction du biais dans les estimations des flux, par rapport aux flux provenant des modèles au niveau de la personne. S'il s'agit d'un modèle de non-réponse dont on n'a pas à tenir compte, il n'est pas nécessaire de recourir à des modèles au niveau du ménage parce que les modèles au niveau du ménage et au niveau de la personne sont équivalents. En outre, on a démontré que le contrôle de la non-réponse au niveau du ménage n'élimine pas nécessairement tout le biais provenant des estimations des flux de la population active. La spécification correcte du modèle de non-réponse est toujours considérée impérative, quoique le fait de tenir compte de la structure des données du ménage peut entraîner un peaufinage des estimations des flux si le modèle de non-réponse est mal spécifié. En particulier, nous démontrons que les estimations au niveau du ménage sont moins biaisées que leurs estimations équivalentes au niveau de la personne.

Notre modèle de non-réponse est une extension de l'idée qui veut que la non-réponse peut dépendre des caractéristiques d'une unité, dans le présent cas les flux de la population active des membres du ménage. La non-réponse dans les enquêtes sur les ménages peut survenir pour plusieurs raisons, par ex., un refus, un non-contact, un déménagement ou un renouvellement de l'échantillon. Le modèle actuel peut facilement être étendu pour modéliser des tendances plus complexes de non-réponse en précisant l'indicateur de non-réponse comme étant une variable polytome et en paramétrant le modèle de non-réponse conformément aux tendances complexes de non-réponse. Il est également à remarquer que nous ne supposons pas que le modèle au niveau du ménage est une représentation exacte du comportement de non-réponse du ménage; plutôt, nous supposons que le modèle au niveau de la personne offre une approximation de la dynamique de non-réponse au sein du ménage.

Un problème important, mis en évidence par les résultats provenant de l'étude en simulation, est notre hypothèse que le comportement des flux de la population active individuels est homogène au sein des ménages. De toute évidence, il s'agit d'une hypothèse qui n'est pas réaliste. Le modèle est facilement étendu en précisant les flux de la population active et les probabilités de flux de non-réponse comme modèles de régression pour tenir compte des renseignements liés à la covariable au niveau de la personne, au niveau du ménage ou à un niveau plus élevé. Par exemple, les probabilités de flux de la population active pourraient être précisées comme étant une régression multinomiale-logistique:

$$\log\left(\frac{\omega_{hi}(ab)}{\omega_{hi}(11)}\right) = \beta_0^{(ab)} + \boldsymbol{\beta}_1^{(ab)} \boldsymbol{x}_{hi}^T,$$

où $\omega_{hi}(ab)$ désigne la probabilité d'une personne $i$ dans un ménage $h$ du flux de la population active $(a, b)$, $\boldsymbol{x}_{hi}$ est un vecteur de covariable (rangée), et $(\beta_0^{(ab)}, \boldsymbol{\beta}_1^{(ab)})$ sont les coefficients de régression pour le mutinomial-logit $(a, b)$. Cependant, l'ajustement de ces modèles nécessite que l'on émette des hypothèses d'indépendance conditionnelles au sujet de la relation entre les distributions des covariables, les flux de la population active et les flux de non-réponse parce que les renseignements sur les covariables peuvent être manquants dans le cas des ménages qui ne répondent pas. Une autre solution est de permettre des covariables hétérogènes entre les flux de la population active du ménage, à l'aide d'effets aléatoires, en adoptant des hypothèses relativement à la distribution des différences entre les ménages. L'ajustement de ces modèles est également compliquée et nécessiterait, par exemple, une méthode de Monte Carlo de chaîne de Markov pour effectuer l'intégration nécessaire. Si $S$ n'est pas un échantillon aléatoire simple, on peut incorporer des variables auxiliaire du plan dans le processus d'ajustement en utilisant le cadre de régression que nous venons de décrire.

## REMERCIEMENTS

Le travail de Paul Clarke dans le présent article était financé grâce à une bourse d'études du Conseil de la recherche économique et sociale (prix n° R00429614273); le travail de Ray Chambers était financé grâce au contrat conclu entre l'Office for National Statistics et l'Université de Southampton pour la prestation de services de recherches en méthodologie statistique. Les deux auteurs aimeraient remercier les examinateurs, dont les observations et les conseils pratiques ont aidé à rendre la version définitive du manuscrit considérablement plus compréhensible.

## BIBLIOGRAPHIE

FITZMAURICE, G.M., LAIRD, N.M., et ZAHNER, G.E.P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91, 99-107.

HOGUE, C.R. (1985). History of the problems encountered in estimating gross flows. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 1-8.

LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin de l'Institut International de Statistique*, 15, 1-15.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

STASNY, E.A. (1986). Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

STASNY, E.A., et FIENBERG, S.E. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 25-39.

VANSKI, J.E. (1985). Uses of gross change data in assessing demographic labor market dynamics. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 9-12.

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 141-149
Statistique Canada



# Analyse longitudinale de données de l'enquête suisse sur la population active par régression logistique multidimensionnelle

PAUL-ANDRÉ SALAMIN[1]

RÉSUMÉ

Dans le cadre d'enquêtes longitudinales, de simples estimations de variations, comme des différences de pourcentages, peuvent ne pas être toujours suffisamment efficaces pour déceler des variations ayant de l' importance sur le plan pratique, notamment dans le cas de sous-populations. Le recours à des modèles, qui peuvent représenter la structure de dépendance de l'enquête longitudinale, peut aider à résoudre ce problème. Une des principales caractéristiques observées dans le cadre de l'Enquête suisse sur la population active (ESPA) est la situation d'activité. Étant donné que cette enquête comporte un plan d'échantillonnage à renouvellement de panel, les données qui en sont tirées sont des données nominales multidimensionnelles, et une grande partie des profils de réponse est absente conformément à ce plan. Le modèle logistique multidimensionnel, qui a été introduit par Glonek et McCullagh (1995) comme généralisation de la régression logistique présente de l'intérêt dans ce contexte parce qu'il autorise des observations répétées dépendantes et des profiles de réponse incomplets. Nous montrons qu'en utilisant la régression logistique multidimensionnelle, on peut représenter la structure de dépendance complexe de l'ESPA à l'aide d'un petit nombre de paramètres et obtenir des estimations plus efficaces d'une variation.

MOTS CLÉS: Données binaires longitudinales; modèle logistique multidimensionnel; enquête sur la population active.

## 1. INTRODUCTION

Un des principaux objectifs de l'Enquête suisse sur la population active (ESPA) est la production d'estimations des variations des pourcentages de population relatifs à diverses situations d'activité. Normalement, de simples estimations d'une variation, comme la différence entre les pourcentages observés d'une année à une autre au sujet des personnes qui occupent un emploi, sont calculées pour l'ensemble de la population, ainsi que pour un nombre important de sous-populations. En général, ce résultat est insatisfaisant, car les estimations relatives aux sous-populations peuvent ne pas être toujours suffisamment efficaces pour déceler des variations qui ont une importance sur le plan pratique. Le travail que nous présentons ici répond à la question de savoir si le recours à des modèles qui permettent de représenter la structure de dépendance de l'enquête précitée pouvait contribuer à résoudre ce problème.

Étant donné que l'ESPA est fondée sur un plan d'échantillonnage à renouvellement de panel, nous avons affaire à des données longitudinales nominales dont une grande partie des profils de réponse sont laissés incomplets conformément au plan. Le point central de notre étude est la modélisation des probabilités marginales, et notamment des probabilités d'être dans une situation d'activité donnée, en tant que fonction du temps et d'autres covariables qui définissent des sous-populations. Si les observations répétées de la situation d'activité étaient indépendantes, une façon normale de procéder consisterait à utiliser la régression logistique. Le modèle logistique multidimensionnel introduit par Glonek et McCullagh (1995) comme généralisation de la régression logistique est intéressant dans ce contexte car il autorise des observations répétées dépendantes et des profils de réponse incomplets.

Le but du présent article est de montrer que le recours à la régression logistique multidimensionnelle pour modéliser la structure de dépendance complexe de l'ESPA permet d'obtenir des estimateurs de variation plus efficaces. Bien que nous présentions la méthode en l'appliquant à des données de l'ESPA, il est clair qu'elle peut être appliquée à un éventail de données plus large.

De nombreuses questions importantes ne sont pas abordées dans le présent article. Parce que les données de l'ESPA sont tirées d'une enquête complexe, on peut soutenir que toute analyse devrait prendre en considération les poids d'échantillonnage (Pfeffermann 1993). Ici, nous utilisons uniquement les données non pondérées. Cependant, on peut montrer, à l'aide de la méthode de pseudo-vraisemblance de Binder (1983), que la régression logistique multidimensionnelle peut être élargie au cas en question (Salamin 1998). La non-réponse est toujours une préoccupation importante dans le cas des enquêtes sur échantillon. Ici, nous considérons uniquement les profils de réponse incomplets qui résultent du renouvellement du panel. Dans ce cas, l'hypothèse de valeurs manquantes complètement aléatoire est raisonnable. Il est à noter, cependant, que la régression logistique multidimensionnelle présente suffisamment de souplesse pour pouvoir intégrer des paramètres supplémentaires en considération des profils incomplets résultant de la perte de personnes faisant partie du panel. Ainsi, les personnes qui ont quitté le panel

[1] Paul-André Salamin, Statistical Methods Unit, Swiss Federal Statistical Office, Espace de l'Europe 10, CH-2010 Neuchâtel, Switzerland.

auraient pu être incluses également dans l'analyse. Enfin, il est bien connu que des erreurs de classification peuvent introduire des biais importants dans les probabilités relatives aux profils de réponse observées (voir, par exemple, Pfeffermann, Skinner et Keith 1998). Il serait certainement souhaitable d'étudier l'incidence de ces biais sur les estimations de paramètres de la régression logistique multidimensionnelle, qui comportent des interprétations ayant trait à des moments marginaux.

Les modèles log-linéaires et les modèles marginaux sont étroitement liés à la régression logistique multidimensionnelle; nous y revenons dans la section 3. Nous présentons brièvement ci-après les modèles de transition, les modèles d'effets aléatoires et l'analyse de survie, dans le contexte de l'ESPA. Dans le cas d'un modèle de transition (voir, par ex., Diggle, Liang et Zeger 1994, chap. 10, ou Zeger et Liang 1992), les observations répétées de la situation d'activité sont corrélées parce que les situations d'activité passées ont une incidence sur la situation d'activité actuelle. Le point d'intérêt réside dans les probabilités de transition entre les diverses situations d'emploi; par exemple, la probabilité d'occuper un emploi peut dépendre du fait d'avoir été sans emploi dans le passé. Dans le contexte de la régression, les réponses antérieures sont traitées comme des variables explicatives supplémentaires. Un aspect important est la détermination du nombre de réponses antérieures à inclure comme variables prédictives. Si le modèle relatif aux probabilités de transition est correctement spécifié, on peut traiter des transitions répétées relatives à une personne en particulier comme des événements indépendants et utiliser des méthodes statistiques courantes, comme la régression logistique. Dans le cadre d'un modèle d'effets aléatoires (voir, par ex., Diggle et coll. 1994, chap. 9), la probabilité d'être dans une situation d'activité donnée est une fonction de variables explicatives et les coefficients de régression varient d'un individu à l'autre. Cette variabilité des coefficients de régression reflète l'hétérogénéité naturelle des individus qui est due à des facteurs non mesurés. Compte tenu des coefficients de régression, on présuppose que les observations répétées de la situation d'activité sont indépendantes. La corrélation entre les observations répétées survient uniquement parce que nous ne sommes pas en mesure d'observer les coefficients de régression réels. Cette façon de procéder présente la plus grande utilité lorsque le point d'intérêt réside dans l'inférence relative à des individus plutôt que dans des moyennes de population. Dans le cas de l'analyse de survie, que l'on désigne également par le terme «analyse des antécédents» dans les textes qui traitent d'économétrie (Lancaster 1990), le point d'intérêt est la modélisation des transitions entre les diverses situations d'activité au fil du temps, considérées comme une fonction de variables explicatives. Dans ce cas, le moment précis où a lieu la transition est important. Dans l'ESPA, la situation d'activité est observée une fois par an. Les changements qui ont eu lieu durant l'année qui a précédé l'interview peuvent être reconstitués. Cependant, étant donné que cette reconstitution est fondée sur l'auto-évaluation des répondants, il peut y avoir des imprécisions en ce qui a trait aux situations antérieures et aux moments où les changement ont eu lieu. On peut trouver une analyse des données de l'ESPA fondée sur cette méthode dans Gerfin (1996).

Le présent article est organisé de la manière décrite ci-après. Dans la section 2, nous décrivons les données, un sous-ensemble d'environ 5000 individus tiré de l'ESPA, qui sont utilisées dans les exemples présentés dans les sections 4 et 5. Dans la section 3, nous présentons la régression logistique multidimensionnelle en la comparant avec les modèles log-linéaires et marginaux. Dans la section 4, nous montrons comment on peut utiliser la régression logistique multidimensionnelle pour représenter la structure de dépendance complexe des données de l'ESPA au moyen d'un faible nombre de paramètres. Dans la section 5, nous considérons l'estimation de la variation en comparant la régression logistique multidimensionnelle avec un simple estimateur de variation. Nous montrons que l'utilisation de la régression logistique multidimensionnelle aboutit à un gain d'efficacité. Enfin, dans la section 6, nous présentons nos conclusions et des orientations pour des études ultérieures.

## 2. DONNÉES DE L'ENQUÊTE SUISSE SUR LA POPULATION ACTIVE

On peut trouver une description détaillée du plan d'échantillonnage et de la méthode de pondération de l'ESPA dans Hulliger, Ries, Comment et Bender (1997). Nous rappelons ici uniquement certains des aspects importants de cette enquête. L'ESPA recueille des renseignements sur la situation d'activité de résidents de la Suisse âgés de quinze ans et plus. Depuis le deuxième trimestre de 1991, un échantillon d'environ 16 000 personnes fait l'objet d'interviews annuelles. Cette enquête est fondée sur un plan d'échantillonnage à renouvellement de panel, avec une durée de présence dans l'échantillon de cinq ans. Durant la phase de démarrage, c'est-à-dire entre 1992 et 1996, environ un cinquième de l'échantillon d'origine a été remplacé par un échantillon de renouvellement chaque année. Les unités de l'échantillon de renouvellement vont rester dans le panel pour une période complète de cinq ans.

Dans les exemples que nous donnons dans les sections 4 et 5, nous utilisons les observations relatives à la situation d'activité pour les années 1992 à 1995 qui ont été obtenues auprès des personnes faisant partie de l'échantillon du canton de Vaud. Le tableau 1 indique la structure des données ainsi que la taille des échantillons longitudinal et transversal. En raison du plan d'échantillonnage qui a été choisi, certains des profils de réponse sont incomplets. Par exemple, dans le cas des personnes qui ont été choisies en

1991 et qui ont été remplacées en 1994, la période d'observation, (notée (1)234, va de 1991 à 1994. Nous utilisons la notation (1)234 pour indiquer le fait que nous n'avons pas eu recours aux observations qui ont été effectuées en 1991.

**Tableau 1**
Structure des données, taille des échantillons longitudinal et transversal, canton de Vaud, 1992-1995

| Première année dans l'échantillon | Époques d'observation des diverses parties de l'échantillon | | | | Période d'observation | |
|---|---|---|---|---|---|---|
| 91 | 92 | | | | (1)2 | 622 |
| | 92 | 93 | | | (1)23 | 412 |
| | 92 | 93 | 94 | | (1)234 | 527 |
| | 92 | 93 | 94 | 95 | (1)2345 | 481 |
| 92 | 92 | 93 | 94 | 95 | 2345 | 612 |
| 93 | | 93 | 94 | 95 | 345 | 722 |
| 94 | | | 94 | 95 | 45 | 728 |
| 95 | | | | 95 | 5 | 877 |
| | 2 654 | 2 754 | 3 070 | 3 420 | | 4 981 |

La situation d'activité est une variable nominale qui comprend les trois catégories suivantes: «personne occupée», «sans emploi» et «hors de la population active». Dans les exemples que nous donnons dans les sections 4 et 5, nous utilisons une variable binaire qui prend la valeur 1 si la personne sondée occupe un emploi, et la valeur 2 si elle est sans emploi ou hors de la population active. Nous procédons de cette façon uniquement pour simplifier la présentation des modèles logistiques multidimensionnels. Étant donné que la méthode que nous utilisons permet de traiter un nombre arbitraire de catégories, il serait préférable de ne pas regrouper les situations d'activité dans le cas d'une analyse réelle. Dans le cas où il serait néanmoins nécessaire de regrouper certaines des situations d'activité, il faut faire preuve de prudence, étant donné que l'hétérogénéité des situations peut introduire des biais.

## 3. MODÈLES LOGISTIQUES MULTIDIMENSIONNELS

Le modèle logistique multidimensionnel qui a été introduit par Glonek et McCullagh (1995) permet de traiter des réponses multidimensionnelles de type nominal ou ordinal ainsi que des variables explicatives discrètes ou continues. Nous allons considérer ici uniquement des réponses binaires multidimensionnelles et des variables prédictives discrètes. Le modèle logistique multidimensionnel est un exemple de modèle linéaire généralisé (voir McCullagh et Nelder 1989). Sa fonction de liaison, appelée également transformation logistique multidimensionnelle, exprime la distribution conjointe des profils de réponse sous forme de moments marginaux par ordre croissant, les deux premiers moments étant des logits marginaux et des risques relatifs logarithmiques marginaux. La fonction de liaison a une caractéristique, désignée par le terme «reproductibilité», qui fait que le modèle logistique multidimensionnel s'applique à tout sous-ensemble du vecteur de réponse. Cette caractéristique assure l'égalité des interprétations des paramètres, quel que soit le nombre de variables de réponse, et que des paramètres d'un ordre supérieur soient inclus ou non. Cette propriété rend la régression logistique multidimensionnelle particulièrement intéressante pour l'analyse de données longitudinales, où les observations répétées d'un résultat ont lieu sur un pied d'égalité et où le nombre de ces observations répétées peut varier d'un individu à l'autre. La reproductibilité est également l'élément clé qui permet d'utiliser le modèle pour traiter des réponses incomplètes. Cependant, il est à noter qu'il faut présupposer que l'absence complète de données est aléatoire, si les mêmes paramètres doivent être utilisés pour modéliser aussi bien les réponses complètes que les réponses incomplètes. Les estimations de paramètres sont déterminées selon la méthode du maximum de vraisemblance. Une opération clé est l'inversion de la transformation logistique multidimensionnelle. Dans le cas de plus de trois réponses, il se peut que cette opération ne puisse pas toujours être effectuée, car dans un tel cas il existe des contraintes qui touchent les paramètres (Glonek et McCullagh 1995, Liang, Zeger et Qaqish 1992). En outre, la présence de cellules vides peut limiter l'ordre des paramètres qui peuvent être ajustés.

Le modèle log-linéaire est largement utilisé pour modéliser des données binaires multidimensionnelles. Dans le modèle log-linéaire saturé (voir, par ex., Liang et coll. 1992), le paramètre canonique associé à un sous-ensemble des variables a une interprétation sous forme de probabilités conditionnelles, compte tenu des autres variables; par exemple, les paramètres de premier et deuxième ordre sont des logits et des risques relatifs logarithmiques qui dépendent de toutes les autres réponses. Il s'ensuit que le modèle log-linéaire n'est pas reproductible, ce qui en fait un choix moins intéressant que la régression logistique multidimensionnelle pour l'analyse des données longitudinales. Il est néanmoins possible d'élaborer des modèles log-linéaires qui, à l'instar du modèle logistique multidimensionnel, comportent des logits marginaux comme paramètres. Cela mène aux modèles marginaux (Diggle et coll, 1994, chap. 8). Dans ces modèles, la dépendance des probabilités marginales à l'égard de variables explicatives est modélisée de manière distincte de la corrélation existant au sein des unités. Avec cette approche, les paramètres ne sont pas estimés par la méthode du maximum de vraisemblance; on spécifie plutôt uniquement la structure de la corrélation existant entre les observations répétées d'un résultat, et les paramètres sont estimés par la résolution d'équations d'estimation généralisées (EEG), un analogue multidimensionnel de la quasi-vraisemblance (McCullagh et Nelder 1989). Un certain nombre de spécifications de la structure de corrélation ont été proposées; par exemple, Liang et coll. (1992) ont utilisé les risques relatifs marginaux comme

Glonek et McCullagh (1995). Nous avons effectué certaines comparaisons entre la régression logistique multidimensionnelle et le progiciel PROC GENMOD de SAS (version 6.12). Ce progiciel permet d'ajuster des modèles de réponses corrélées d'après la méthode des EEG. Nous avons obtenu des estimations des logits marginaux très semblables. La méthode des EEG s'est avérée légèrement moins efficace que la régression logistique multidimensionnelle. Un des inconvénients de cette méthode est qu'elle ne permet pas d'obtenir des estimations de toutes les probabilités de profils de réponse, mais uniquement les probabilités marginales. Le modèle logistique multidimensionnel ne présente pas cet inconvénient parce que ses paramètres sont estimés selon la méthode du maximum de vraisemblance.

En suivant Glonek et McCullagh (1995), nous décrivons, dans la section 3.1, la transformation logistique multidimensionnelle, et nous donnons, dans la section 3.2, l'algorithme relatif au maximum de vraisemblance.

### 3.1 Transformation logistique multidimensionnelle

Soient $Y_1, Y_2, ..., Y_d$ un nombre $d$ d'observations répétées de la même variable binaire, effectuées aux époques $t_1 < t_2 < ... < t_d$, et soit

$$\pi_{i_1 i_2 ... i_d} = P(Y_1 = i_1, Y_2 = i_2, ..., Y_d = i_d),$$

où $i_1, i_2, ..., i_d$ ont comme valeur 1 ou 2, les probabilités conjointes des variables aléatoires $Y_1, Y_2, ..., Y_d$. Dans le modèle logistique multidimensionnel, les probabilités conjointes de $Y_1, Y_2, ..., Y_d$ sont paramétrées sous forme de logits marginaux, de risques relatifs logarithmiques marginaux et de contrastes de risques relatifs logarithmiques marginaux. Ce paramétrage peut être exprimé de la manière suivante: $\eta = C^T \log(L\pi)$, où $\pi$ est le vecteur de dimension $q = 2^d$

$$\pi = (\pi_{11...11}, \pi_{11...12}, ..., \pi_{22...21}, \pi_{22...22})^T,$$

et où les matrices $L$ et $C$ sont les produits tensoriels, respectivement, d'un indicateur marginal et de matrices de contraste appropriés. Les matrices $L$ et $C$, qui dépendent de la durée $d$ de la période d'observation, sont définies dans une suite récurrente, en commençant par $L_0 = C_0 = 1$, avec

$$L_d = \begin{bmatrix} L_{d-1} \otimes 1_2^T \\ L_{d-1} \otimes \tilde{L} \end{bmatrix}$$

et

$$C_d = \begin{bmatrix} C_{d-1} & 0 \\ 0 & C_{d-1} \otimes \tilde{C} \end{bmatrix},$$

où $1_2^T = (1, 1)$, $\tilde{L}$ la matrice unité deux par deux et $\tilde{C} = (1, -1)^T$ (Glonek et McCullagh 1994).

À titre d'exemple, nous considérons des périodes d'observation d'une durée $d = 1, 2, 3, 4$. Dans le cas de $d = 1$, $\pi = (\pi_1, \pi_2)^T$ et $\eta = (\eta_0, \eta_1)^T = (\log \pi_+, \text{logit } Y_1)^T$, où l'indice positif indique la sommation et où logit $Y_1$ est défini comme suit:

$$\text{logit } Y_1 = \log \frac{P(Y_1 = 1)}{P(Y_1 = 2)} = \log \frac{\pi_1}{1 - \pi_1} = \log \frac{\pi_1}{\pi_2}.$$

Dans ce cas, la transformation logistique multidimensionnelle est équivalente à la transformation logistique habituelle. Il est à noter que même si le paramètre $\eta_0 = \log \pi_+ = 0$ est tout à fait superflu, il est utile de le conserver comme moyen permettant de s'assurer que l'application $\pi \to \eta$ est de plein niveau et exprime également l'exigence selon laquelle $\pi_+ = 1$.

Lorsque $d = 2$, $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^T$ et

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12})^T =$$

$$(\log \pi_{++}, \text{logit } Y_1, \text{logit } Y_2, \log OR(Y_1, Y_2))^T$$

où

$$RR(Y_1, Y_2) =$$

$$\frac{P(Y_1 = 1, Y_2 = 1) P(Y_1 = 2, Y_2 = 2)}{P(Y_1 = 1, Y_2 = 2) P(Y_1 = 2, Y_2 = 1)} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

est le risque relatif, une grandeur qui mesure l'association entre les variables $Y_1$ et $Y_2$. Les paramètres $\eta_1$ et $\eta_2$ sont les logits marginaux aux époques $t_1$ et $t_2$ ; par exemple:

$$\eta_1 = \text{logit } Y_1 = \log \frac{\pi_{1+}}{(1 - \pi_{1+})}.$$

Lorsque $d = 3$, $\pi = (\pi_{111}, \pi_{112}, ..., \pi_{221}, \pi_{222})^T$ et

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123})^T.$$

Les paramètres $\eta_1, \eta_2$ et $\eta_3$ sont les logits marginaux aux époques $t_1, t_2$ et $t_3$. Les paramètres $\eta_{12}, \eta_{13}$ et $\eta_{23}$ sont les risques relatifs logarithmiques des tableaux marginaux bidimensionnels correspondants; par exemple:

$$\eta_{23} = \log RR(Y_2, Y_3) = \log \frac{\pi_{+11}\pi_{+22}}{\pi_{+12}\pi_{+21}}.$$

Le paramètre $\eta_{123}$ est un contraste des risques relatifs qui est donné par

$$\eta_{123} = \log RR(Y_1, Y_2 \mid Y_3 = 1) - \log RR(Y_1, Y_2 \mid Y_3 = 2)$$
$$= \log \frac{\pi_{111}\pi_{221}}{\pi_{121}\pi_{211}} - \log \frac{\pi_{112}\pi_{222}}{\pi_{122}\pi_{212}}.$$

Lorsque $d = 4$, $\pi = (\pi_{1111}, \pi_{1112}, ..., \pi_{2221}, \pi_{2222})^T$ et

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123},$$
$$\eta_4, \eta_{14}, \eta_{24}, \eta_{124}, \eta_{34}, \eta_{134}, \eta_{234}, \eta_{1234})^T.$$

Les paramètres $\eta_i, \eta_{ij}$ et $\eta_{ijk}$, où $1 \le i < j < k \le 4$, sont définis comme ci-dessus à l'aide des tableaux marginaux appropriés. Le paramètre $\eta_{1234}$ est un contraste des risques relatifs logarithmiques qui est donné par

$$\eta_{1234} = \log RR(Y_1, Y_2 \mid Y_3 = 1, Y_4 = 1)$$
$$- \log RR(Y_1, Y_2 \mid Y_3 = 1, Y_4 = 2)$$
$$- \log RR(Y_1, Y_2 \mid Y_3 = 2, Y_4 = 1)$$
$$+ \log RR(Y_1, Y_2 \mid Y_3 = 2, Y_4 = 2).$$

Une opération clé de l'estimation du maximum de vraisemblance est le calcul de l'inverse de la transformation logistique multidimensionnelle. Pour faire en sorte que $\pi > 0$, nous utilisons $\pi = \exp \nu$, c'est-à-dire que nous cherchons à trouver $\nu$ dans l'équation $\eta = C^T \log(L \exp \nu)$. En général, on ne dispose pas d'une solution explicite, de sorte qu'il faut utiliser une méthode itérative. On peut notamment appliquer les itérations de Newton-Raphson, comme nous l'expliquons ci-après. Pour plus de clarté, nous définissons les deux fonctions $\varphi(\pi) = C^T \log(L\pi)$ et $\psi(\nu) = \varphi(\exp \nu)$.

i) Commencer avec une approximation initiale $\nu_0$.
ii) Prendre ensuite $\nu_{k+1} = \nu_k - [D\psi(\nu_k)]^{-1}(\varphi(\exp \nu_k) - \eta)$, où $D\psi(\nu)$ est la matrice jacobienne de la fonction $\psi(\nu)$, et itérer jusqu'à convergence.

Les matrices jacobiennes des fonctions $\varphi(\pi)$ et $\psi(\nu)$ sont données respectivement par $D\varphi(\pi) = C^T(\operatorname{diag} L\pi)^{-1} L$ et $D\psi(\nu) = (\exp \nu) \cdot \operatorname{diag}(\exp \nu)$.

### 3.2 Estimation du maximum de vraisemblance

Dans le cas d'une variable de réponse binaire observée à $d$ époques il y a $q = 2^d$ profils de réponse possibles, où $i_1, i_2, ..., i_d$ ont comme valeur soit 1, soit 2. Dans le cas de chaque profil $i = (i_1, ..., i_d)$, nous définissons la variable indicatrice $Y_{i_1 ... i_d}$, qui est égale à 1 si l'on a observé le profil $i$, et à 0 dans le cas contraire. On obtient alors

$$P(Y_{i_1 ... i_d} = 1) = P(Y_1 = i_1, ..., Y_d = i_d) = \pi_{i_1 ... i_d}.$$

Les vecteurs de grandeur $q$ étant définis comme suit:

$$Y = (Y_{11...11}, Y_{11...12}, ..., Y_{22...21}, Y_{22...22})^T$$

et

$$\pi = (\pi_{11...11}, \pi_{11...12}, ..., \pi_{22...21}, \pi_{22...22})^T,$$

nous pouvons écrire $Y \sim M(1, \pi)$, c'est-à-dire que $Y$ est un vecteur multinomial avec $q = 2^d$ catégories dont les probabilités sont données par le vecteur $\pi$.

Les modèles de régression logistique multidimensionnelle sont donc définis pour être ceux de la forme $\eta = X\beta$, où $X$ est une matrice $q \times p$ de variables explicatives, $\beta$ un vecteur de grandeur $p$ de paramètres inconnus, et où $\eta = C^T \log(L\pi) = \varphi(\pi)$.

Si nous prenons $y$ comme observation du vecteur aléatoire $Y$, nous pouvons alors exprimer le noyau de la fonction de vraisemblance logarithmique comme suit: $l(\beta; y) = y^T \log \pi(\beta)$, où, en utilisant l'inverse de la transformation logistique multidimensionnelle, nous pouvons exprimer les probabilités conjointes $\pi$ comme une fonction du paramètre inconnu $\beta$, sous la forme $\pi(\beta) = \varphi^{-1}(X\beta)$. Le vecteur de valeur numérique est donné par

$$s(\beta) = s(\beta, y, X) = D\pi(\beta)^T (\operatorname{diag} \pi(\beta))^{-1} y,$$

où $D\pi(\beta)$, c'est-à-dire la matrice jacobienne de la fonction $\pi(\beta)$ qui lie le paramètre $\beta$ au vecteur de probabilités $\pi$, est donnée par $D\pi(\beta) = [D\varphi(\varphi^{-1}(X\beta))]^{-1} X$, et où $D\varphi(\pi) = C^T(\operatorname{diag} L\pi)^{-1} L$ est la matrice jacobienne de la fonction de liaison. La matrice d'information est définie comme suit: $\Im(\beta) = Es(\beta)s(\beta)^T$. Il s'ensuit, d'après la présupposition relative à la distribution de $Y$, que $E(YY^T) = \operatorname{diag} \pi$, d'où nous pouvons déduire que

$$\Im(\beta) = \Im(\beta, X) = D\pi(\beta)^T (\operatorname{diag} \pi(\beta))^{-1} D\pi(\beta).$$

Si nous avons un nombre $n$ d'observations indépendantes $y_k \sim M(1, \pi_k)$, $k = 1, ..., n$, où $\eta_k = C^T \log(L\pi_k) = X_k\beta$, le vecteur de valeur numérique et la matrice d'information sont donnés alors par $s(\beta) = \sum_{k=1}^n s(\beta, y_k, X_k)$ et par $\Im(\beta) = \sum_{k=1}^n \Im(\beta, X_k)$.

L'estimateur du maximum de vraisemblance de $\beta$ est la solution de $s(\beta) = 0$, qui peut être trouvée à l'aide de l'algorithme de cotation de Fisher qui, à partir d'une quelconque valeur initiale $\beta_0$, itère la séquence $\beta_{m+1} + \beta_m + \Im_m^{-1}(\beta_m) s(\beta_m)$ jusqu'à convergence.

Des profils de réponse incomplets peuvent être intégrés facilement à l'analyse. En particulier, lorsqu'un quelconque sous-ensemble des variables de réponse $Y_1, Y_2, ..., Y_c$ est enregistré pour une unité donnée, la distribution des probabilités sur ce tableau marginal de grandeurs $c$ est multinomiale et, en raison de la reproductibilité de la transformation logistique multidimensionnelle, un modèle de régression logistique multidimensionnelle s'applique au tableau des probabilités. En outre, la matrice de plan

d'échantillonnage qui lie les probabilités marginales à $\beta$ est établie en choisissant les rangs appropriée de la matrice de plan d'échantillonnage complète qui serait utilisée si l'on disposait de données complètes concernant l'unité en question.

## 4. MODÈLES POUR LA DÉPENDANCE LONGITUDINALE

Dans la présente section, nous montrons, à l'aide des données de l'ESPA indiquées dans la section 2, comment on peut appliquer la régression logistique multidimensionnelle pour décrire la dépendance existant entre les observations répétées de la situation d'activité. Nous n'entendons pas effectuer une recherche exhaustive pour trouver un modèle optimal; nous allons plutôt montrer que la régression logistique multidimensionnelle permet de représenter une structure de dépendance complexe à l'aide d'un petit nombre de paramètres.

Nous considérons six modèles à complexité décroissante (voir tableau 2). Dans le cas de tous les modèles, nous avons un paramètre pour chacun des logits marginaux qui correspond à une époque d'observation donnée. Ce paramètre est désigné par la notation $\eta_i = \beta_i$. Étant donné que les époques d'observation sont les deuxièmes trimestres des années 1992 à 1995, nous prenons $i = 2, 3, 4, 5$. Ainsi, $\beta_3$, par exemple, correspond au logit de la probabilité d'occuper un emploi en 1993. De façon analogue, les indices pour les paramètres d'ordre supérieur vont de 2 à 5. Dans le cas du modèle 1, nous prenons un modèle saturé pour la dépendance longitudinale, c'est-à-dire que nous avons un paramètre pour chacune des interactions d'ordre 2, 3 ou 4 à l'intérieur de chaque période d'observation. Dans le cas des modèles 2 à 5, nous présupposons que les interactions d'ordre 3 et 4 sont toutes égales à zéro. La dépendance longitudinale est alors décrite seulement sous forme de risques relatifs logarithmiques. Dans le cas du modèle 2, nous avons recours à un modèle saturé pour les risques relatifs logarithmiques. Dans le modèle 3, nous laissons tomber la covariable période d'observation, c'est-à-dire que nous présupposons que les risques relatifs logarithmiques sont les mêmes pour toutes les périodes d'observation. Dans le modèle 4, nous utilisons des risques relatifs logarithmiques stationnaires, c'est-à-dire des risques relatifs logarithmiques qui dépendent uniquement de la différence entre les époques d'observation. Il est à noter que dans le modèle 4, le paramètre $\gamma_1$ correspond à la contrainte $\beta_{23} = \beta_{34} = \beta_{45}$ à laquelle sont assujettis les paramètres du modèle 3, et la même chose vaut pour $\gamma_2$ et $\gamma_3$. Dans le cas du modèle 5, on présuppose un modèle linéaire pour les risques relatifs logarithmiques stationnaires. Enfin, dans le modèle 6, nous présupposons que les observations effectuées à des époques différentes sont indépendantes. Il est à noter que dans ce cas, la régression logistique multidimensionnelle équivaut à la régression logistique ordinaire.

**Tableau 2**
Six modèles pour la dépendance longitudinale

| | Paramètres | | |
|---|---|---|---|
| Modèle | Logits marginaux | Risques relatifs logarithmiques | Paramètres de 3ième et 4ième ordre |
| 1 | $\eta_i = \beta_i$ | $\eta_{ij} = \beta_{ij,\text{période}}$ | $\eta_{ijk} = \beta_{ijk,\text{période}}$, $\eta_{ijkl} = \beta_{ijkl,\text{période}}$ |
| 2 | $\eta_i = \beta_i$ | $\eta_{ij} = \beta_{ij,\text{période}}$ | $\eta_{ijk} = 0$, $\eta_{ijkl} = 0$ |
| 3 | $\eta_i = \beta_i$ | $\eta_{ij} = \beta_{ij}$ | $\eta_{ijk} = 0$, $\eta_{ijkl} = 0$ |
| 4 | $\eta_i = \beta_i$ | $\eta_{ij} = \gamma_{\lvert i-j \rvert}$ | $\eta_{ijk} = 0$, $\eta_{ijkl} = 0$ |
| 5 | $\eta_i = \beta_i$ | $\eta_{ij} = \delta + \gamma . \lvert i-j \rvert$ | $\eta_{ijk} = 0$, $\eta_{ijkl} = 0$ |
| 6 | $\eta_i = \beta_i$ | $\eta_{ij} = 0$ | $\eta_{ijk} = 0$, $\eta_{ijkl} = 0$ |

Les estimations de paramètres pour les modèles 2 à 6 sont indiquées dans le tableau 3. Le nombre de paramètres et les valeurs de la fonction de vraisemblance logarithmique dans le cas des estimations du maximum de vraisemblance sont indiqués dans le tableau 4, où nous avons inclus également, à des fins de comparaison, la vraisemblance logarithmique pour le modèle entièrement saturé.

Dans l'ensemble, nous constatons que la forme présupposée de la dépendance longitudinale semble avoir peu d'incidence sur les estimations des logits marginaux. Il s'agit là d'une caractéristique souhaitable, étant donné que les logits marginaux seraient normalement les paramètres auxquels on s'intéresse. Les erreurs-types des logits marginaux sont presque les mêmes dans le cas des modèles qui tiennent compte de la dépendance longitudinale, mais ils sont gonflés dans une mesure d'environ 15 % dans le cas de la régression logistique ordinaire (modèle 6). On peut également montrer que les estimations des logits marginaux présentent entre elles une corrélation positive dans le cas de modèles qui présupposent une dépendance longitudinale, et aucune corrélation dans le cas de la régression logistique ordinaire. Dans l'exemple qui nous intéresse, nous avons constaté que la corrélation se situe entre 0,4 et 0,8. Ainsi, la modélisation de la dépendance longitudinale permet également d'obtenir des estimations plus efficaces de la différence existant entre les logits marginaux.

D'après l'ajustement du modèle 1, on peut voir que les paramètres d'interaction d'ordre 3 et 4 ne sont pas très différents de 0. Cela laisse penser que la dépendance longitudinale peut être décrite uniquement par les risques relatifs logarithmiques. Cette hypothèse est corroborée par l'écart incrémental du modèle 2 par rapport au modèle 1, qui est de 7,9 pour 12 degrés de liberté. En outre, tous les paramètres du modèle 2 sont différents de zéro de manière importante, et un examen des résidus normalisés dans le cas des probabilités ajustées relatives aux profils de réponse ne révèle aucune anomalie. Dans le cas d'applications relatives à des statistiques officielles, le modèle 2 serait le modèle privilégié, étant donné qu'il est fondé sur un nombre de présuppositions aussi faible que possible tout en permettant une réduction importante du nombre de paramètres, ce qui

diminue le risque d'avoir des tableaux clairsemés lorsqu'on considère des périodes d'observation prolongées comportant un plus grand nombre de covariables.

Les modèles 3, 4 et 5 montrent qu'il serait néanmoins possible de simplifier grandement la description de la dépendance longitudinale sans perdre trop d'informations. En passant du modèle 2 au modèle 5, nous constatons que l'écart par rapport au modèle entièrement saturé n'augmente pas beaucoup (voir tableau 4). En outre, un examen des résidus montre que les modèles 3, 4 et 5 correspondent aux données presque aussi bien que le modèle 2. D'autre part, si le modèle 2 requiert vingt paramètres pour décrire la dépendance longitudinale, le modèle 5 n'en nécessite que deux. Ce fait doit être opposé au modèle 6, qui présuppose l'indépendance entre les observations effectuées à des époques différentes; la vraisemblance logarithmique est beaucoup plus faible que dans le cas du modèle entièrement saturé (voir tableau 4) et l'ajustement aux données laisse à désirer.

**Tableau 3**
Estimations de paramètres et erreurs-types

| Paramètres | Période | Modèle 2 | Modèle 3 | Modèle 4 | Modèle 5 | Modèle 6 |
|---|---|---|---|---|---|---|
| logit 92 | | 0,6348 (0,0350) | 0,6360 (0,0352) | 0,6348 (0,0352) | 0,6347 (0,0352) | 0,6471 (0,0409) |
| logit 93 | | 0,5555 (0,0335) | 0,5570 (0,0338) | 0,5597 (0,0335) | 0,5601 (0,0335) | 0,5509 (0,0396) |
| logit 94 | | 0,5440 (0,0324) | 0,5407 (0,0325) | 0,5402 (0,0326) | 0,5397 (0,0325) | 0,5377 (0,0374) |
| logit 95 | | 0,4699 (0,0317) | 0,4711 (0,0320) | 0,4710 (0,0320) | 0,4712 (0,0320) | 0,4705 (0,0351) |
| $\beta_{23}$ | (1)23 | 4,2563 (0,3311) | 4,2579 (0,1465) | | | |
| | (1)234 | 4,2003 (0,2894) | | | | |
| | (1)2345 | 4,0859 (0,2954) | | | | |
| | 2345 | 4,4830 (0,2841) | | | | |
| $\beta_{34}$ | (1)234 | 4,0894 (0,2794) | 4,1111 (0,1310) | | | |
| | (1)2345 | 3,9611 (0,2840) | | | | |
| | 2345 | 4,0989 (0,2600) | | | | |
| | 345 | 4,2490 (0,2468) | | | | |
| $\beta_{45}$ | (1)2345 | 5,3992 (0,3854) | 4,5561 (0,1389) | | | |
| | 2345 | 3,9779 (0,2544) | | | | |
| | 345 | 4,7288 (0,2735) | | | | |
| | 45 | 4,5069 (0,2600) | | | | |
| $\beta_{24}$ | (1)234 | 3,7168 (0,2641) | 3,8371 (0,1442) | | | |
| | (1)2345 | 4,2560 (0,3059) | | | | |
| | 2345 | 3,5330 (0,2370) | | | | |
| $\beta_{35}$ | (1)2345 | 4,4000 (0,3098) | 3,7913 (0,1334) | | | |
| | 2345 | 3,6493 (0,2396) | | | | |
| | 345 | 3,6116 (0,2192) | | | | |
| $\beta_{25}$ | (1)2345 | 4,3984 (0,3173) | 3,5774 (0,1530) | | | |
| | 2345 | 3,2209 (0,2256) | | | | |
| $\gamma_1$ | | | | 4,3260 (0,0928) | | |
| $\gamma_2$ | | | | 3,8519 (0,1050) | | |
| $\gamma_3$ | | | | 3,5340 (0,1495) | | |
| $\delta$ | | | | | 4,7341 (0,1266) | |
| $\gamma$ | | | | | −0,4191 (0,0653) | |

**Tableau 4**
Nombre de paramètres et valeur de la fonction de vraisemblance logarithmique dans le cas des estimations du maximum de vraisemblance

| Modèle | Nombre de paramètres d'ordre | | | | Total | Vraisemblance logarithmique |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | |
| Modèle complet | 20 | 20 | 10 | 2 | 52 | -5 342,7 |
| 1 | 4 | 20 | 10 | 2 | 36 | -5 345,4 |
| 2 | 4 | 20 | 0 | 0 | 24 | -5 349,4 |
| 3 | 4 | 6 | 0 | 0 | 10 | -5 365,2 |
| 4 | 4 | 3 | 0 | 0 | 7 | -5 368,9 |
| 5 | 4 | 2 | 0 | 0 | 6 | -5 369,5 |
| 6 | 4 | 0 | 0 | 0 | 4 | -7 815,3 |

## 5. COMPARAISON AVEC L'ESTIMATION SIMPLE DE LA VARIATION

Dans la présente section, nous nous concentrons sur l'estimation de la différence, observée d'une année donnée à une autre, dans les probabilités d'occuper un emploi. Nous montrons que des estimations fondées sur la régression logistique multidimensionnelle sont plus efficaces que des estimations simples définies comme étant la différence existant entre les proportions de personnes qui occupent un emploi.

Le modèle qui est considéré ici est le modèle 2 de la section 4, le sexe étant ajouté comme variable explicative supplémentaire. Dans le cas de chacun des sexes, nous avons un paramètre pour chacun des logits marginaux qui correspond à une année donnée. La dépendance longitudinale est prise en considération au moyen d'un modèle saturé pour les risques relatifs logarithmiques. Les paramètres de troisième et quatrième ordre sont établis à zéro. Ce modèle comporte donc 8 paramètres pour les logits marginaux et 40 paramètres pour les risques relatifs logarithmiques (2 sexes × 20 risques relatifs à l'intérieur des périodes d'observation; voir tableau 3). En inversant la transformation logistique multidimensionnelle, on peut également calculer les estimations des probabilités d'occuper un emploi et de la variation de ces probabilités d'une année à l'autre.

Un estimateur simple de variation est donné par la différence existant entre les proportions de personnes occupant un emploi qui est observée d'une année donnée à une autre. La variance, qui tient compte du chevauchement des deux échantillons, est donnée par l'expression suivante:

$$\frac{1}{n+r}\pi_{1+}(1-\pi_{1+}) + \frac{1}{n+c}\pi_{+1}(1-\pi_{+1}) - 2\frac{n}{(n+r)(n+c)}(\pi_{11}-\pi_{1+}\pi_{+1}),$$

où $n$ est le nombre de cas au sujet desquels on dispose d'observations pour les deux années, $r$ et $c$ le nombre de cas au sujet desquels on dispose d'observations relatives à un an, $\pi_{11}$ la probabilité d'occuper un emploi durant les deux ans, et $\pi_{1+}$ et $\pi_{+1}$ les probabilités marginales d'occuper un emploi.

Concernant les données de l'ESPA de la section 2, le tableau 5 indique les estimations de la différence de la probabilité d'occuper un emploi obtenues avec les deux méthodes. On remarquera que ces méthodes donnent des estimations de la variation similaires. Les erreurs-types des estimations simples sont, en moyenne, 30 % plus importants que dans le cas de la régression logistique multidimensionnelle. L'efficacité relative moyenne de la régression logistique multidimensionnelle par rapport aux estimations simples est de 1,7. Par comparaison, l'efficacité relative moyenne de la régression logistique multidimensionnelle par rapport à la régression logistique ordinaire est de 3,2.

**Tableau 5**
Variation de la probabilité d'occuper un emploi, canton de Vaud, 1992-1995

| | Comparaison | Régression logistique multidimensionnelle | Estimation simple |
|---|---|---|---|
| Femme | 92 *vs.* 93 | 0,0138 (0,0090) | 0,0136 (0,0115) |
| | 92 *vs.* 94 | 0,0184 (0,0102) | 0,0168 (0,0134) |
| | 92 *vs.* 95 | 0,0375 (0,0109) | 0,0356 (0,0149) |
| | 93 *vs.* 94 | 0,0047 (0,0087) | 0,0031 (0,0107) |
| | 93 *vs.* 95 | 0,0238 (0,0095) | 0,0219 (0,0128) |
| | 94 *vs.* 95 | 0,0191 (0,0076) | 0,0188 (0,0100) |
| Homme | 92 *vs.* 93 | 0,0220 (0,0095) | 0,0283 (0,0116) |
| | 92 *vs.* 94 | 0,0245 (0,0102) | 0,0334 (0,0133) |
| | 92 *vs.* 95 | 0,0387 (0,0106) | 0,0452 (0,0144) |
| | 93 *vs.* 94 | 0,0024 (0,0092) | 0,0052 (0,0111) |
| | 93 *vs.* 95 | 0,0167 (0,0098) | 0,0169 (0,0130) |
| | 94 *vs.* 95 | 0,0143 (0,0080) | 0,0117 (0,0102) |

## 6. CONCLUSIONS

Les analyses des données de l'ESPA que nous avons décrites dans le présent article montrent l'utilité de la régression logistique multidimensionnelle. La modélisation de la dépendance longitudinale est nécessaire pour obtenir un ajustement satisfaisant des probabilités relatives aux profils de réponse qui sont observés. En ne tenant pas compte de la dépendance longitudinale, nous obtenons quand même des estimations ponctuelles acceptables des logits marginaux, mais les informations sur la structure détaillée des données sont perdues. La modélisation de la dépendance longitudinale permet également d'obtenir des

estimations plus efficaces des paramètres marginaux et de la variation, comparativement à la régression logistique ordinaire et à l'estimation simple de la variation. Enfin, nous avons décrit la possibilité qu'offre la régression logistique multidimensionnelle de représenter une structure de dépendance complexe à l'aide d'un petit nombre de paramètres.

En utilisant les résultats obtenus par Glonek et McCullagh (1995), il est possible d'appliquer les exemples que nous avons présentés ici à des réponses multidimensionnelles de type nominal ou ordinal, avec des variables explicatives discrètes ou continues. La méthode que nous avons décrite permet également de tenir compte des poids d'échantillonnage (Salamin 1998). Dans le cas de l'ESPA, nous avons constaté que les poids d'échantillonnage ont peu d'incidence sur les estimations de paramètres du modèle logistique multidimensionnel. L'erreur-type des estimations de paramètres était gonflé d'environ 15 %. Cette augmentation modérée de la variabilité de ces estimations, qui est due aux poids d'échantillonnage, est plausible. En effet, étant donné qu'on choisit une seule personne par ménage dans le cadre de l'ESPA, nous ne nous attendions pas à un effet de grappe important.

Outre que pour les types d'analyse que nous avons présentés ici, la régression logistique multidimensionnelle peut être utilisée également pour modéliser des probabilités de non-réponse dans le cas d'études longitudinales. De tels modèles pourraient être utiles lorsqu'il faut corriger des poids d'échantillonnage pour tenir compte de la non-réponse. La possibilité qu'offre la régression logistique multidimensionnelle d'obtenir un modèle parcimonieux des données pourrait avoir de l'intérêt dans le cas de l'estimation relative à de petites région. Des estimateurs pour une région géographique donnée pourraient notamment être fondés sur des modèles relatifs à une région plus vaste choisie de manière appropriée.

Bien que nous n'ayons pas rencontré des difficultés importantes dans les exemples que nous avons présentés ici, des études ultérieures pourraient être nécessaires en ce qui a trait au problème des tableaux clairsemés. L'inversion de la transformation logistique multidimensionnelle est une opération importante lorsqu'il y a un grand nombre de cellules vides. La méthode proposée par Lang (1996), dans laquelle l'inversion de la fonction de liaison est évitée en spécifiant les modèles par l'entremise de contraintes, pourrait être intéressante dans un tel contexte. Un autre aspect à étudier est l'incidence des erreurs de classification sur les estimations de paramètres du modèle logistique multidimensionnel.

## REMERCIEMENTS

Pour la rédaction du présent article, nous avons bénéficié d'entretiens avec des collègues de l'Office fédéral de la statistique, parmi lesquels méritent d'être mentionnés spécialement Beat Hulliger et Philippe Eichenberger. L'aide d'Ariane Bender, de la section responsable de l'Enquête suisse sur la population active, a été extrêmement utile. L'auteur remercie également le rédacteur et deux examinateurs pour leurs excellents conseils, qui ont permis d'améliorer cet article de manière importante.

## BIBLIOGRAPHIE

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

DIGGLE, P.J., LIANG, K.-Y., et ZEGER, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

GERFIN, M. (1996). *Entwicklung von ökonometrischen Modellen zur Analyse der Dynamik auf dem schweizerischen Arbeitsmarkt*. SLFS-News, Swiss Federal Statistical Office, Berne.

GLONEK, G.F.V., et McCULLAGH, P. (1994). Multivariate Logistic Models. Rapport technique 94-31, School of Information Science and Technology, Flinders University of South Australia, Adelaide.

GLONEK, G.F.V., et McCULLAGH, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society*, B, 57, 533-546.

HULLIGER, B., RIES, A., COMMENT, T., et BENDER, A. (1997). Weighting the Swiss Labour Force Survey. (Éds. C. Malaguerra, S. Morgenthaler et E. Ronchetti ). Dans *Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini's Birth, Monte Verità, Switerland*, Basel: Birkhäuser Verlag.

LANCASTER, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.

LANG, J.B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Annals of Statistics*, 24, 726-752.

LIANG, K.-Y., ZEGER, S.L., et QAQISH, B. (1992). Multivariate Regression Analysis for Categorical Data. *Journal of the Royal Statistical Society*, B, 54, 3-40.

McCULLAGH, P., et NELDER, J.A. (1989). *Generalized Linear Models*, (2ième édition). London: Chapman and Hall.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

PFEFFERMANN, D., SKINNER, C., et KEITH, H. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, A, 161, Part 1, 13-32.

SALAMIN, P.-A. (1998). Multivariate logistic regression for data from complex surveys. À paraître *Recueil: Symposium '98, Analyse longitudinale pour les enquêtes complexes*, Statistique Canada, Mai 1998.

ZEGER, S.L., et LIANG, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11, 1825-1839.

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 151-158
Statistique Canada



# Les enquêtes sur les indices de prix en tant qu'études quasi-longitudinales

ALAN H. DORFMAN[1]

RÉSUMÉ

Pour calculer les indices de prix, il faut recueillir des données relatives à un «même article» (en fait, un ensemble d'articles définis avec précision) durant diverses périodes. La question qu'on se pose est celle de savoir si de telles données «quasi-longitudinales» peuvent être modélisées de manière à expliquer ce qu'est un indice des prix. Des chercheurs de pointe spécialisés dans les questions relatives aux indices de prix ont émis des doutes quant à la possibilité d'utiliser la modélisation statistique pour caractériser de tels indices. Dans la présente communication, on propose un simple modèle à espace d'états relatif aux données sur les prix qui donne un indice des prix à la consommation exprimé d'après les paramètres du modèle.

MOTS CLÉS: Modèle à marche aléatoire et bruit; modèle à espace d'états; indice de Laspeyres; indice de Paasche; indice des prix géométrique.

## 1. INTRODUCTION

L'échantillonnage d'enquête utilisé pour calculer un indice des prix à la consommation comporte la surveillance d'un article donné au fil du temps, afin de déterminer ses prix à diverses époques. Seulement voilà, normalement, on ne suit pas exactement le même article (en effet, ce n'est pas le prix de la boîte de soupe aux tomates de marque Y que l'on trouve au point de vente Z qui est vérifié à plusieurs reprises, car il est probable que la boîte en question aura été vendue et consommée entre deux visites de la personne chargée de l'enquête), mais plutôt d'une succession d'articles qui correspondent tous à la même description («Boîte de 8 oz de soupe aux tomates avec harengs de marque Y, vendue au point de vente Z») et dont le prix est noté à diverses époques. Autrement dit, il s'agit essentiellement d'un groupe d'articles qui correspondent à une description précise suivie au fil du temps. C'est pourquoi les enquêtes qui portent sur les indices de prix peuvent être qualifiées de «quasi-longitudinales», par opposition aux enquêtes longitudinales, dans lesquelles on suit des articles distincts au fil du temps. Il est néanmoins raisonnable d'espérer que le fait d'effectuer des relevés à plusieurs reprises au fil du temps puisse mener à des méthodes d'estimation qui pourraient tirer avantage de l'aspect chronologique des enquêtes de ce type.

Compte tenu de cet espoir, la présente communication se penche sur une question qui a été généralement ignorée par les statisticiens et les économistes, ou qui tout au plus à reçu une réponse négative. Cette question est la suivante: est-ce qu'on peut traiter un indice des prix à la consommation (IPC) d'un point de vue statistique? Autrement dit, est-ce que le paramètre qui caractérise la «variation du coût de la vie» d'une période à une autre, et que les enquêtes sur les indices de prix essaient d'estimer, peut être défini d'après un modèle stochastique?

Aldrich (1992) donne une interprétation historique des premières tentatives, entreprises par Jevons, et surtout par Edgeworth, d'intégrer des hypothèses distributionnelles aux indices des prix à la consommation. Les communications récentes ayant pour objet la modélisation stochastique sont celles de Balk (1980), Clements et Izan (1981, 1987), Bryan et Cechetti (1993), Kott (1984), et Selvanathan et Rao (1994). Diewert (1995) passe en revue et critique ces tentatives en invoquant un argument de Keynes (1930) comme motif déterminant pour rejeter l'approche stochastique.

Dans la présente communication, nous proposons une approche précise en ce qui a trait à la modélisation de l'indice des prix à l'aide de modèles à espace d'états, et nous soumettons provisoirement un tel modèle. Celui-ci est appliqué à des données de balayage afin de montrer la faisabilité d'un indice fondé sur ces mêmes données. La méthode que nous considérons contourne la critique de Keynes d'une façon fondamentale et offre la perspective des nombreux avantages que peut apporter une solide modélisation statistique, y compris, peut-être, une simplification du processus d'échantillonnage d'enquête.

Ci-après, nous allons d'abord passer brièvement en revue la définition d'un indice des prix ainsi que les deux méthodes (non stochastiques) qui sous-tendent de manière prédominante le choix des indices (section 2). Nous examinons notamment l'exemple de modèle statistique pour indices des prix qui est présenté par Bryan et Cecchetti (1993), ainsi que la formulation de l'objection de Keynes par Diewert (section 3). Nous présentons par la suite une méthode de modélisation d'un indice des prix à la consommation qui contourne les difficultés mentionnées

[1] Alan H. Dorfman, U.S. Bureau of Labor Statistics, Room 4915, 2 Massachusetts Ave. N.E., Washington, D.C., 20212-0001, U.S.A.; courrier électronique: dorfman_a@bls.gov.

par Keynes et Diewert, ce qui mène naturellement à l'utilisation de modèles à espace d'états (section 4). Nous présentons notamment les résultats de l'application d'un modèle relativement simple à marche aléatoire et bruit à des données de balayage tirées de l'Academic Data Base d'A.C. Nielsen (section 5). Nous évaluons le nouvel indice dans la section 6, en mentionnant les travaux de recherche ultérieurs qui pourraient être utiles.

## 2. HISTORIQUE DE LA QUESTION

Par indice des prix à la consommation (IPC) on entend un nombre unique qui indique la variation du pouvoir d'achat des consommateurs entre une période $t'$ et une période $t$. Les éléments de base bruts de cet indice sont les prix des divers articles que l'on peut se procurer (au moins) aux deux époques

$$p_\tau = (p_{\tau 1}, \ldots, p_{\tau N}), \tau = t', t$$

ainsi que les quantités des articles vendus

$$q_\tau = (q_{\tau 1}, \ldots, q_{\tau N}), \tau = t', t.$$

(Cependant, dans la pratique, les données quantitatives relatives aux deux époques en question ne sont pas disponibles, et il faut alors recourir à des données de remplacement d'un type ou d'un autre). L'IPC est obtenu à l'aide d'une «formule» dans laquelle on utilise les éléments bruts suivants:

$$I_{t't} = f(p_{t'}, p_t, q_{t'}, q_t),$$

où $f(\cdot)$ est une fonction de l'une des nombreuses expressions possibles. La plupart de ces expressions existent depuis longtemps et ont été décrites de façon approfondie dans les textes spécialisés qui portent sur les indices de prix.

À titre d'exemple, nous mentionnons ici l'indice de Laspeyres

$$L_{t't} = \frac{\sum_{i=1}^{N} q_{t'i} p_{ti}}{\sum_{i=1}^{N} q_{t'i} p_{t'i}} = \sum_{i=1}^{N} f_{t'i} r_{t't},$$

où $f_{t'i} = q_{t'i} p_{t'i} / \sum_{i=1}^{N} q_{t'i} p_{t'i}$ représente les «dépenses relatives» et $r_{t'i} = p_{ti}/p_{t'i}$ les «rapports de prix». L'indice de Laspeyres fait appel aux quantités relatives à la période qui vient en premier comme base fixe pour la comparaison des prix antérieurs et des prix qui suivent. L'indice Laspeyres (ou une variante proche) a tendance à être l'indice le plus ciblé par les gouvernements, en raison de sa simplicité et de sa transparence aux yeux du profane.

Le pendant naturel de l'indice de Laspeyres est l'indice de Paasche

$$P_{t't} = \frac{\sum_{i=1}^{N} q_{ti} p_{ti}}{\sum_{i=1}^{N} q_{ti} p_{t'i}}$$

qui étalonne les prix d'après les quantités de la période subséquente. La plupart des autres indices qui sont fondés sur d'autres formules se situent généralement entre les indices de Paasche et de Laspeyres.

À titre de référence ultérieure dans la présente communication, nous mentionnons un indice fondé sur la moyenne géométrique, où des poids non négatifs fixes $f_i$, donnent au total 1:

$$G_{t't} = \prod_{i=1}^{N} \left( \frac{p_{ti}}{p_{t'i}} \right)^{f_i}.$$

On utilise le terme «moyenne géométrique» pour référer à cet indice.

Fisher (1922) analyse ces formules, ainsi que de nombreuses autres formules d'indices. Il introduit notamment l'approche dite «des tests» pour effectuer un choix parmi les diverses possibilités de formule $f(\cdot)$; dans le cas de cette méthode, on décrit les propriétés («tests») qu'un indice raisonnable devrait nécessiter, puis on détermine dans quelle mesure chaque formule d'indice satisfait à ces critères.

Un des tests est la condition de réversibilité dans le temps: $I_{t't} I_{tt'} = 1$. Deux indices qui sont toujours utilisés dans le monde mais qui ne passent pas ce test sont l'indice Carli-Sauerbach $C_{t't} = \sum_{i=1}^{N} f_i p_{ti}/p_{t'i}$ et la moyenne géométrique $\tilde{G}_{t't} = \prod_{i=1}^{N} (p_{ti}/p_{t'i})^{f_{t'i}}$ qui fait appel aux dépenses de la période antérieure plutôt qu'à des poids fixes. On peut montrer facilement que $C_{t't} C_{tt'} \geq 1$, en utilisant l'inégalité de Cauchy-Schwartz, ce qui laisse penser que cet indice sera trop élevé.

Si une augmentation des prix de l'article $i$ a tendance à produire une augmentation des dépenses, alors $\tilde{G}_{t't} \tilde{G}_{tt'} \leq 1$, de sorte que dans de telles conditions, la moyenne géométrique fondée sur la période antérieure tend à être trop basse. D'autre part, si une augmentation des prix de l'article $i$ tend à produire une diminution des dépenses, alors $\tilde{G}_{t't}$ donne des indications trop élevées. En général on peut donc s'attendre à ce que cet indice soit plutôt irrégulier.

Cette constatation suggère l'axiome suivant: les indices de prix qui sont exprimés sous la forme d'une moyenne géométrique ne devraient pas lier des poids aux prix observés dans une des périodes comparées; les indices qui prennent la forme d'une moyenne arithmétique ne devraient pas comporter des poids indépendants des prix en question.

De façon contrastante par rapport à $\tilde{G}_{t't}$, la moyenne géométrique $G_{t't} = \prod_{i=1}^{N} (p_{ti}/p_{t'i})^{f_i}$ qui comporte des poids

fixes, est le seul indice qui satisfait aux cinq axiomes relatifs aux indices de prix qui sont énoncés dans Balk (1995) ainsi qu'au «test de circularité», selon lequel $I_{t't} = I_{t't^*} I_{t^*t}$ lorsque $t' < t^* < t$. L'inversion du temps est une conséquence immédiate.

Les indices qui passent la plupart des tests sont généralement ceux qui intègrent des données quantitatives tirées des deux périodes; c'est le cas, par exemple, de l'indice de Fisher

$$F_{t't} = \left(L_{t't} P_{t't}\right)^{1/2}$$

et de l'indice de Törnqvist

$$T_{t't} = \prod_{i=1}^{N} \left( \frac{p_{ti}}{p_{t'i}} \right)^{f_{t'ti}},$$

où $f_{t'ti} = (f_{t'i} + f_{ti})/2$. Souvent, on ne peut pratiquement pas distinguer l'un de l'autre les indices de Fisher et de Törnqvist. On peut trouver d'autres analyses de la méthode par tests dans Balk (1995), Diewert (1987), ainsi que dans Eichhorn et Voeller (1976).

La deuxième façon d'évaluer les formules relatives aux indices est la méthode «économique». Celle-ci définit un indice générique qui prend la forme suivante:

$$I_{t't} = \frac{C(p_t, U)}{C(p_{t'}, U)},$$

où $U = U(q_1, ..., q_N)$ est une «fonction utilitaire» bien définie et $C(p_t, U)$ est le coût minimal aux prix $p_t$, de l'atteinte du niveau de vie ou «utilité» $U$. Dans le cas d'une fonction utilitaire en particulier, on cherche à savoir si une formule donnée peut être considérée comme une bonne approximation de l'indice du coût de la vie correspondant. Comme la méthode des tests, cette deuxième façon de procéder a tendance à donner des indices qui comprennent des données quantitatives relatives aux deux périodes. Voir Diewert (1987) pour plus détails.

## 3. LA MÉTHODE STOCHASTIQUE

Aldrich (1992) présente l'historique des premiers essais de modélisation des rapports de prix, ou des logarithmes de rapports de prix, à l'aide d'un paramètre courant qui représente le taux global de croissance des prix. Une idée fondamentale contenue dans l'étude de cet auteur est que l'application de la méthode stochastique aux indices de prix, bien qu'elle soit un exemple d'une première application de la statistique à des questions économiques, est morte de sa belle mort. Diewert (1995) analyse lui aussi ces exemples, ainsi que des exemples plus récents de la modélisation statistique de rapports de prix. La difficulté que Diewert, à l'instar de Keynes (1930), voit dans ce genre de modélisation est illustrée par une modèle de Clements et Izan (1987).

La période comprise entre $t'$ et $t$ est divisée en segments de même durée, ce qui donne des intervalles relativement courts représentés de manière générique comme étant compris entre $t - 1$ et $t$. Le logarithme des rapports de prix pour une telle «micro-période» est donné par l'expression suivante:

$$\log\left( \frac{p_{ti}}{p_{t-1,i}} \right) = \pi_t + \beta_i + \varepsilon_{ti}, \qquad (1)$$

où $\varepsilon_{ti} \sim (0, \sigma_i^2 / f_i)$. Dans le modèle de ces auteurs, les $f_i$ représentent la part de dépenses relative à l'article $i$, au cours de la période comprise entre $t'$ et $t$. À des fins d'identification, on présume que $\sum_{i=1}^{N} f_i \beta_i = 0$. Ces présuppositions mènent à un estimateur du maximum de vraisemblance

$$\hat{\pi}_t = \sum_{i=1}^{N} f_i \log\left( \frac{p_{ti}}{p_{t'i}} \right),$$

ce qui donne un estimateur du maximum de vraisemblance (EMV) de la tendance des prix durant la période courte

$$\exp(\hat{\pi}_t) = \prod_{i=1}^{N} \left( \frac{p_{ti}}{p_{t'i}} \right)^{f_i};$$

cela veut donc dire que, d'après le modèle stochastique de ces auteurs, on obtient un indice géométrique, avec des poids $f_i$, semblable à l'indice de Törnqvist.

On peut également obtenir des estimations de $\beta_i$ et de $\sigma^2$, ainsi que des estimations de la précision, par exemple, de la variance de $\hat{\pi}$. Ainsi, une nouvelle fondation statistique semble être placée sous un vieil estimateur.

Diewert (1995) formule plusieurs objections, dont aucune ne peut être prise à la légère. L'objection principale est la suivante:

> «...l'objection fondamentale de Keynes (Keynes 1930, p. 78): 'La variation du niveau des prix [$\exp(\pi_t)$] qui devrait avoir eu lieu s'il n'y avait pas eu de variation dans les prix relatifs, n'est plus pertinente si les prix relatifs ont effectivement subi une variation, car cette variation a eu elle-même une incidence sur le niveau des prix'.» (Traduction).

Si, par exemple, le prix du pain varie par rapport au prix des voitures, cette même variation entraîne une variation du niveau global des prix.

L'objection de Keynes n'est pas tout à fait claire. Pourquoi ne peut-il y avoir deux aspects de la variation des prix, un aspect global et un aspect particulier? Cependant, il n'est pas difficile de reconnaître que les tendances des prix considérés individuellement sont primaires, une tendance des prix considérés globalement ne pouvant être

qu'une somme pondérée des tendances individuelles. Diewert propose la traduction suivante, sous forme d'un modèle, de l'objection de Keynes: étant donné que nous devons déterminer la tendance globale des prix par l'expression

$$\pi_t^* = \sum_{i=1}^{N} f_i \beta_{ti}^*,$$

le modèle (1) doit être remplacé par

$$\log\left(\frac{p_{ti}}{p_{t-1,i}}\right) = \pi_t + \beta_{ti} + \varepsilon_{ti}, \qquad (2)$$

où $\beta_{ti} = \pi_t - \beta_{ti}^*$ et $\sum_{i=1}^{N} f_i \beta_{ti} = 0$. La différence déterminante entre cette équation et l'équation (1) est que, maintenant, les paramètres d'article $\beta_{ti}$ sont indexés en fonction du temps. Mais «alors, le modèle qui est obtenu comporte trop de paramètres à identifier». Ce fait devrait suffire à invalider la méthode.

Diewert (1995) n'aborde pas le modèle chronologique, beaucoup plus compliqué, de Bryan et Cechetti (1993). Parmi les communications précédentes, celle-ci est probablement celle qui se rapproche le plus de la nôtre. Elle porte sur l'utilisation d'un modèle complexe à espace d'états et du filtre de Kalman. À l'instar des autres communications examinées par Diewert, elle fait l'objet de l'objection de Keynes.

## 4. RÉEXAMEN DES INDICES DE PRIX

### 4.1 Présuppositions courantes

La modélisation stochastique du comportement des prix décrite dans la section précédente, qu'elle soit effectuée à l'aide des équations (1) ou (2), ou d'une équation similaire, présente trois caractéristiques dignes de mention; la modélisation est

1. complète: elle vise directement un «taux d'inflation» global qui comprend tous les articles;
2. atomistique: chaque article est modélisé individuellement et comporte son «propre» paramètre, son propre taux d'inflation $[\exp(\pi_t + \beta_i)]$, de manière distincte de tous les autres articles;
3. à périodes multiples: la modélisation pour la période comprise entre $t-1$ et $t$ est distincte de celle s'appliquant à la période qui va de $t-2$ à $t-1$ *etc.*

C'est la combinaison de ces présuppositions qui donne lieu à l'argument de Diewert concernant le trop grand nombre de paramètres. La critique de Keynes vise avant tout la première caractéristique : un taux d'inflation global, pour lequel une augmentation ou une baisse du coût de la vie doit être un mélange pondéré de plusieurs tendances des prix. On peut concéder cela sans aller jusqu'à inclure le point 2. Ce point est accepté tacitement dans le cas de presque toutes les constructions (non stochastiques) d'indices de prix. Cependant, il n'est pas évident du tout que chaque article a sa propre tendance en ce qui a trait au prix. En effet, il est probable que le prix d'articles différents (par exemple, la crème glacée de la marque X dans plusieurs supermarchés) a tendance à augmenter et à diminuer simultanément (du moins à long terme). Il existe des degrés d'homogénéité entre les articles. En tout cas, aucune de ces présuppositions n'est une composante nécessaire d'une approche stochastique à l'égard des indices de prix.

### 4.2 Un modèle élémentaire à espace d'états

Nous divisons la période comprise entre $t'$ et $t$ en sous-périodes $t'$, $t'+1, \ldots, t-1, t$, et l'ensemble d'articles hétérogènes en sous-groupes homogènes, où la caractéristique qui définit l'homogénéité est la tendance à une similitude de comportement en ce qui a trait aux variations de prix. Nous présupposons deux choses:

1. $I_{t't}$ est un mélange d'indices $I_{gt't}$ «homogènes»;
2. $I_{gt't}$ peut être obtenu par enchaînement: $I_{gt't} = \prod_{\tau} I_{g\tau-1,\tau}$, où $\tau = t'+1, \ldots, t$.

Nous nous concentrons sur un seul indice de groupe $I_{gt't}$, et laissons tomber l'indice inférieur $g$, afin de simplifier la notation. Par conséquent, pour le reste de la présente communication, nous nous concentrons sur le «sous-indice» $I_{t't} \equiv I_{gt't}$.

Nous allons maintenant créer un modèle élémentaire à espace d'états (Harvey 1990, chapitre 3) pour les logarithmes des rapports de prix à l'intérieur du groupe. Supposons que le groupe comprend un nombre $n$ d'articles. Dans le cas de $i = 1, \ldots, n$, soient $r_{ti} \equiv p_{ti}/p_{t-1,i}$ les rapports de prix relatifs à la micro-période, et $y_{ti} \equiv \log(p_{ti}/p_{t-1,i}) = \log(p_{ti}) - \log(p_{t-1,i})$, leurs logarithmes. La raison pour laquelle nous utilisons des logarithmes est que d'importants travaux empiriques, en premier lieu ceux d'Edgeworth (voir Diewert (1995)) laissent penser que les logarithmes des rapports de prix seront beaucoup plus susceptibles d'avoir une distribution normale que les rapports de prix eux-mêmes, qui peuvent être considérablement dissymétriques. La distribution normale des erreurs est une présupposition standard dans le cas des modèles à espace d'états. Soit $y_t \equiv (y_{t1}, \ldots, y_{tn})$ et $\mathbf{1}$ un vecteur de un de longueur $n$.

Considérons le modèle multidimensionnel à marche aléatoire et bruit (MMAB):

$$y_t = \mathbf{1}\mu_t + \varepsilon_t,\ \varepsilon_t \sim \mathrm{MVN}\left(0, \sum_{\varepsilon\varepsilon}\right)$$

$$\mu_t = \mu_{t-1} + \eta_t,\ \eta_t \sim N\left(0, \sigma_{\eta\eta}\right) \qquad (3)$$

où $\varepsilon_\tau, \eta_\tau, \tau \in (t', t'+1, \ldots, t-1, t)$ sont mutuellement indépendants. Le modèle laisse entendre que le montant de l'augmentation (ou de la baisse) de l'ensemble des prix du

groupe dans une micro-période tend à osciller autour du montant de l'augmentation (ou de la baisse) observé généralement au cours de la micro-période précédente. C'est une question d'observation courante: si la hausse des prix au cours d'un mois donné tend à être forte (ou faible), au cours du mois suivant, elle sera forte (ou faible) d'une manière correspondante. Étant donné que nous considérons un ensemble d'articles homogène, il est normal que leurs logarithmes de rapport de prix aient une moyenne commune. Nous renvoyons à des travaux ultérieurs la question de savoir comment intégrer des sous-indices à un indice global.

Le modèle (3) comporte le modèle (plus simple) unidimensionnel à marche aléatoire et bruit:

$$\bar{y}_t = \mu_t + \bar{\varepsilon}_t,\ \bar{\varepsilon}_t \sim N(0, \sigma_{\bar{\varepsilon}\bar{\varepsilon}})$$

$$\mu_t = \mu_{t-1} + \eta_t,\ \eta_t \sim N(0, \sigma_{\eta\eta}) \qquad (4)$$

où $\bar{y}_t = n^{-1} 1' y_t, \bar{\varepsilon}_t = n^{-1} 1' \varepsilon_t$, et $\bar{\sigma}_{\varepsilon\varepsilon} = n^{-1} 1' \sum_{\varepsilon\varepsilon} 1$. Certains renseignements sont perdus lorsqu'on utilise l'équation (4); en revanche, la présupposition concernant la normalité est encore plus susceptible d'être juste. Pour plus de commodité, les calculs de l'étude décrite dans la section 5 étaient fondés sur le modèle unidimensionnel.

Le filtre de Kalman (Harvey 1990, section 3.2) peut être utilisé pour obtenir les estimations $\hat{\mu}_\tau$, et $\hat{\sigma}_{\bar{\varepsilon}\bar{\varepsilon}}, \hat{\sigma}_{\eta\eta}$ des paramètres d'état $\mu_\tau$ et des variances $\sigma_{\bar{\varepsilon}\bar{\varepsilon}}, \sigma_{\eta\eta}$ respectivement.

Nous définissons ensuite $I_{t't} \equiv E(G_{t't} \mid S_t)$, où $G_{t't} = \prod_i (p_{it}/p_{t'i})^{f_i}$ est une moyenne géométrique qui dépend des parts fixes $f_i$, et où $S_t$ représente la totalité des paramètres d'état $\mu_\tau$ au cours de la période $t$, et également les «hyperparamètres» $\sigma_{\bar{\varepsilon}\bar{\varepsilon}}, \sigma_{\eta\eta}$. Autrement dit, nous déterminons ce que nous considérons être le processus sous-jacent au cours de la période $t$. Il s'ensuit que

$$I_{t't} = \exp\left(\mu_t + \mu_{t-1} + \cdots \mu_{t'+1} + \frac{1}{2}\nu\right), \qquad (5)$$

où $\nu = (t - t')\sum_i \sum_{i'} \sigma_{ii'} f_i f_{i'}$, avec $\sigma_{ii'}$ étant la covariance de $\varepsilon_{it}$ et $\varepsilon_{i't}$, sera habituellement d'un ordre inférieur à celui des paramètres d'état $\mu_{ti}$. L'estimateur naturel de $I_{t't}$ est $\hat{I}_{t't} \equiv \exp(\hat{\mu}_t + \hat{\mu}_{t-1} + \ldots \hat{\mu}_{t'+1})$; alors

$$E\left(\hat{I}_{t't} \mid S_t\right) = \exp\left(\mu_t + \mu_{t-1} + \cdots \mu_{t'+1} + \frac{1}{2}\tilde{\nu}\right), \qquad (6)$$

où $\tilde{\nu}$, donné dans l'annexe, n'est généralement pas égal à $\nu$, mais est souvent proche de cette valeur et, en tout cas, présente le même ordre de grandeur. La différence $\Delta(\nu) = \tilde{\nu} - \nu$ peut être estimée par, disons, $\hat{\Delta}(\nu)$, ce qui donne un estimateur corrigé pour le biais $\check{I}_{t't} \equiv \hat{I}_{t't} \exp(-1/2\hat{\Delta}(\nu))$. Les expressions pour $\nu$ et $\tilde{\nu}$, ainsi qu'une proposition pour un $\hat{\Delta}(\nu)$ sont donnés dans l'annexe. On peut signaler que $\hat{\Delta}(\nu)$, et donc $\check{I}_{t't}$, dépend des poids $f_i$, mais pas $\hat{I}_{t't}$.

## 5. ÉTUDE EMPIRIQUE

Afin de déterminer la faisabilité du calcul d'indices de prix à l'aide du modèle MMAB, et pour avoir une certaine idée du comportement de l'indice fondé sur ce modèle, on a effectué une petite étude empirique en utilisant des données sur les prix et les quantités du thon en boîte tirées de l'Academic Database d'A.C. Nielsen. Le thon en boîte présente un comportement quelque peu instable en ce qui a trait aux prix et aux quantités, en raison de ventes fréquentes, qui s'effectuent parfois à des prix réduits de façon marquée.

Cette étude couvrait le nord-est des États-Unis et les 104 semaines des années 1992 et 1993. L'ensemble de données initial était plutôt volumineux. Afin de faciliter l'enquête, les données hebdomadaires ont été regroupées par périodes de quatre semaines, ce qui a donné 26 périodes réparties sur deux ans. Ainsi, aux fins de l'étude en question, les données qui ont été utilisées étaient des quantités cumulatives et des prix moyens pondérés d'après la quantité, pour des périodes de quatre semaines.

Les groupes homogènes ont été définis d'après la marque et le type de produit, de la façon suivante: 3 marques, désignées ici par les lettres A, B et C, de thon de «première qualité» dans l'eau, les trois mêmes marques de thon «léger» dans l'huile, et de nouveau les trois mêmes marques de thon «léger» dans l'eau, pour un total de 9 groupes.

L'étude était centrée sur 83 points de vente qui présentaient des quantités positives durant la plupart des périodes de 4 semaines, dans le cas de chacun des 9 groupes.

L'indice $\hat{I}_{t't}$ fondé sur le modèle MMAB et l'indice corrigé $\check{I}_{t't}$ fondé sur le même modèle ont été calculés pour quatre intervalles. Dans chaque cas, la période finale $t = 26$, et la période initiale a été prise successivement comme étant $t' = 3, 6, 10, 14$. À des fins de comparaison, nous avons également calculé les indices correspondants de Laspeyres et de Paasche. Ces deux indices standard servent également de base pour une comparaison indirecte avec les indices de Fisher et de Törnqvist, qui se situent environ à moitié chemin entre les deux indices précédents.

Les figures 1 et 2 (thon de première qualité et thon léger respectivement) donnent les valeurs des quatre indices pour les quatre intervalles, les points indiquant les indices d'espace d'états et les lignes les indices de Laspeyres et de Paasche. L'indice corrigé $\check{I}_{t't}$ est invariablement plus élevé que l'indice non corrigé $\hat{I}_{t't}$. Il est à noter que puisque c'est la première période que nous faisons varier lorsque le tracé des indices est monotone vers le haut, cela semble indiquer une tendance à la baisse du prix du thon dans le groupe visé (et vice versa).

Nous constatons que les nouveaux indices ne sont pas aberrants par rapport aux indices classiques, se situant souvent entre les indices de Laspeyres et de Paasche, mais ils ont tendance à être nettement plus stables à mesure que

$t'$ varie, ce qui laisse penser que les indices classiques réagissent à un «bruit» présent dans les données et qu'en fait, très peu de variation a lieu au fond durant cette période de deux ans. On remarquera également, dans la figure 2, que le thon léger dans l'huile et le thon léger dans l'eau d'une même marque ont des comportements semblables, ce qui laisse penser que nous aurions peut-être dû prendre un groupe «homogène» plus large.

## 6. TRAVAUX ULTÉRIEURS

La recherche décrite dans la présente communication suggère plusieurs sujets pour des travaux de recherche ultérieurs.

Il faut notamment mettre au point des mesures de la précision et des estimations des indices fondés sur le modèle MMAB, sous forme de variances ou d'intervalles de confiance fondés sur le modèle à espace d'états. Même ceux qui doutent de la viabilité de l'application d'une méthode stochastique aux indices de prix trouvent la possibilité d'avoir une mesure de la précision attrayante (Diewert 1995). Il serait également utile d'obtenir des mesures de la précision d'autres indices standard, d'après le modèle à espace d'états.

Il est également souhaitable que du travail empirique soit effectué afin de déterminer avec plus de précision quels groupes d'articles pourraient répondre le mieux au critère d'«homogénéité». Il faudrait également étudier des modèles qui seraient plus sophistiqués que le simple modèle à marche aléatoire et bruit. À cet égard, l'utilisation de données de balayage sera très utile car ces données renseignent de manière très détaillée tant sur les quantités que sur les prix.

La méthode fondée sur des modèles à espace d'états permet de composer avec l'absence de données (Harvey 1990, section 3.4.7). Une question importante que l'on se pose est celle de savoir dans quelle mesure ces modèles permettront de tenir compte de données manquantes dans l'estimation d'indices de prix. Étant donné que, dans la pratique, la plupart des données qui sont utilisée pour calculer des indices de prix portent sur un petit échantillon d'articles offerts, il faut connaître la robustesse des indices fondés sur des modèles à espace d'états à l'égard de l'absence de données.

Les algorithmes de lissage et de prévision des modèles à espace d'états sont bien connus. Leur utilisation dans la révision et la prévision d'indices pourrait être d'un grand intérêt. Enfin, dans la présente communication, nous avons abordé uniquement l'obtention d'un indice pour un seul groupe homogène. Il serait utile de créer un modèle à espace d'états qui réunit plusieurs groupes et qui permet d'obtenir une mesure globale du pouvoir d'achat.

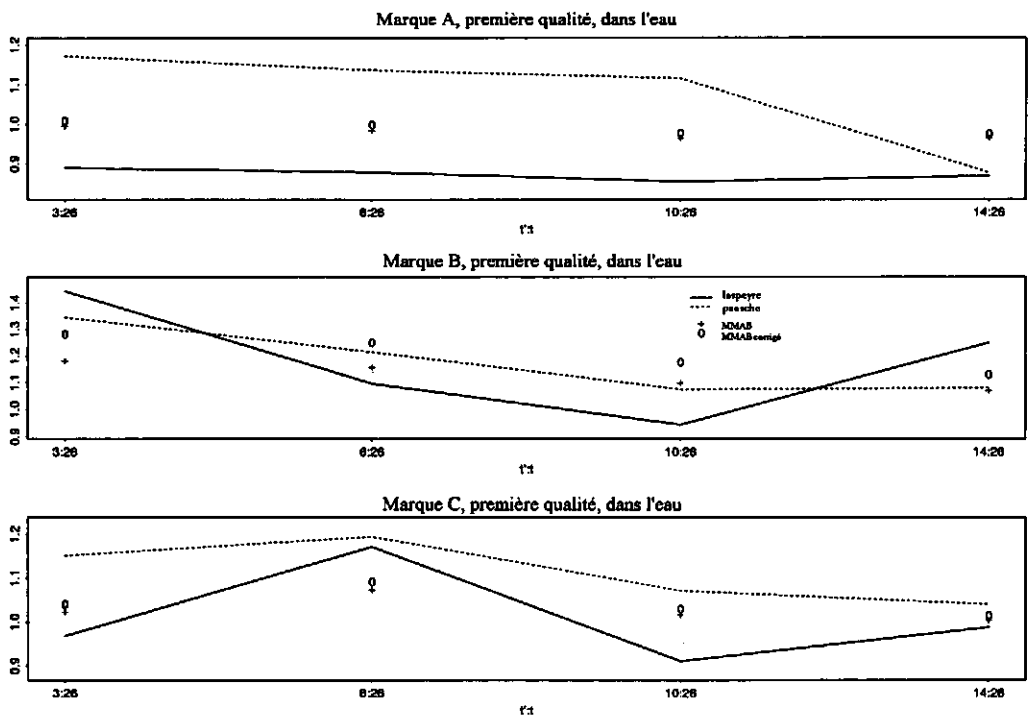



**Figure 1.** Quatre indices de prix pour quatre périodes de temps, thon de première qualité

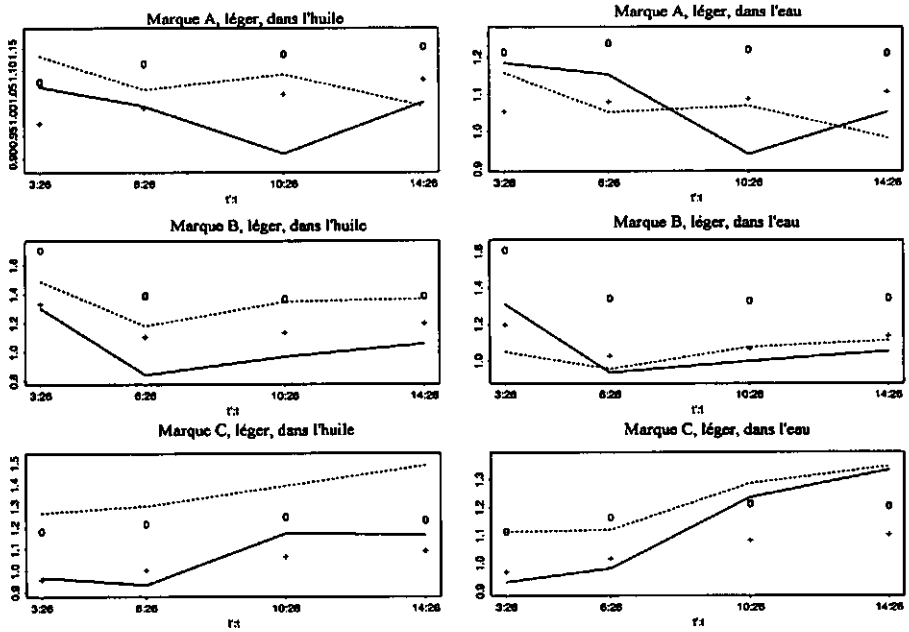



**Figure 2.** Quatre indices de prix pour quatre périodes de temps, thon léger


## REMERCIEMENTS

L'auteur remercie B. Moulton, S. Scott, M. Reinsdorf, R. Tiller, B. Balk, et J. Aldrich pour des entretiens discussions portant sur les idées contenues dans la présente communication.


## ANNEXE

### Détails des équations (5) et (6).

Nous avons

$$G_{t't} = \prod_i \left( \frac{p_{ti}}{p_{t-1,i}} \frac{p_{t-1,i}}{p_{t-2,i}} \cdots \frac{p_{t'+1,i}}{p_{t'i}} \right)^{f_i}$$

$$= \prod_i \left( r_{ti} r_{t-1,i} \cdots r_{t'+1,i} \right)^{f_i},$$

et lorsque

$$H_{t't} = \log(G_{t't}) = \sum_i f_i \log(p_{ti}/p_{t'i}),$$

nous obtenons

$$H_{t't} = \sum_i f_i \log\left(r_{ti} r_{t-1,i} \cdots r_{t'+1,i}\right)$$

$$= \sum_i f_i \left(y_{ti} + y_{t-1,i} + \cdots y_{t'+1,i}\right)$$

et également

$$I_{t't} = E(G_{t't}) = \exp(E(H_{t't}) + 1/2 \operatorname{var}(H_{t't})),$$

où les moments sont calculés en tenant compte de l'état $S_t$, comme dans la section 4.3. Soit $\nu = \operatorname{var}(H_{t't})$. Alors

$$E(H_{t't}) \equiv E(H_{t't} \mid S_t) =$$

$$\sum_i f_i(\mu_t + \mu_{t-1} + \cdots \mu_{t'+1}) = \mu_t + \mu_{t-1} + \cdots \mu_{t'+1}$$

et

$$\nu = \operatorname{var}(H_{t't}) \equiv \operatorname{var}(H_{t't} \mid S_t) =$$

$$\operatorname{var}\left( \sum_{\tau=t'+1}^{t} \sum_i f_i \varepsilon_{\tau i} \mid S_t \right) = (t - t') \sum_i \sum_{i'} \sigma_{ii'} f_i f_{i'},$$

où $\sigma_{ii'}$ représente la covariance de $\varepsilon_{ti}$ et $\varepsilon_{ti'}$. Nous constatons que $\nu = (t - t') \sum_i f_i^2 \sigma_{ee}$, dans le cas particulier où les erreurs $\varepsilon_{ti}$ sont indépendantes et réparties de manière identique dans chaque période.

Nous considérons maintenant l'estimateur $\hat{I}_{t't} \equiv \exp(\hat{\mu}_t + \hat{\mu}_{t-1} + \cdots \hat{\mu}_{t'+1})$. Nous constatons que $E(\hat{I}_{t't}) = \exp(\mu_t + \mu_{t-1} + \cdots \mu_{t'+1} + 1/2\, \tilde{\nu})$, où

$$\tilde{\nu} \equiv \operatorname{var}\left( \sum_{t'+1}^{t} \hat{\mu}_\tau \mid S_t \right) = \left\{ \sum_{t'+1}^{t} \gamma_\tau^2 \right\} \operatorname{var}(\bar{y}_t \mid S_t) +$$

$$\gamma_{t'}^{*2} \operatorname{var}(\hat{\mu}_{t'} \mid S_t) = \left\{ \sum_{t'+1}^{t} \gamma_\tau^2 + \gamma_{t'}^{*2} p_{t'-1} \right\} \sigma_{\bar{e}\bar{e}}$$

avec

$$\gamma_\tau = k_\tau(1 + \sum_{v=\tau+1}^{t} \prod_{u=\tau+1}^{v} (1 - k_u))$$

et

$$\gamma^*_{t'} = \sum_{v=t'+1}^{t} \prod_{u=t'+1}^{v} (1 - k_u),$$

où

$$k_\tau = p_{\tau|\tau-1} / (p_{\tau|\tau-1} + 1),$$

et $p_{\tau|\tau-1}$, $p_\tau$ représentent les erreurs quadratiques moyennes de $\hat{\mu}_\tau$, dans le cas de données jusqu'à $\tau - 1$, $\tau$ respectivement, et ceux-ci sont estimés à l'aide du filtre de Kalman.

Ce résultat découle des équations qui sont utilisées pour estimer $\mu_\tau$:

$$\hat{\mu}_t = k_t \bar{y}_t + (1 - k_t)\hat{\mu}_{t-1}$$

$$\hat{\mu}_{t-1} = k_{t-1} \bar{y}_{t-1} + (1 - k_{t-1})\hat{\mu}_{t-2}$$

$$\vdots$$

$$\hat{\mu}_{t'+1} = k_{t'+1} \bar{y}_{t'+1} + (1 - k_{t'+1})\hat{\mu}_{t'}$$

(*cf.* Harvey 1990, équation 3.2.8), en exprimant chaque $\hat{\mu}_\tau$ par $\bar{y}_\tau, \bar{y}_{\tau-1}, \ldots, \bar{y}_{t'+1}, \hat{\mu}_{t'}$.

En comparant $v$ et $\tilde{v}$, nous constatons de façon empirique que

$$\sum_{t'+1}^{t} \gamma_\tau^2 + \gamma_{t'}^{*2} p_{t'-1} \approx t - t'.$$

Nous considérons ici le cas simple où $\text{var}(\varepsilon_{ti}) = \sigma_{\varepsilon\varepsilon}$ et $\text{cov}(\varepsilon_{ti}, \varepsilon_{ti'}) = \rho\sigma_{\varepsilon\varepsilon}$, avec $\rho \geq 0$, dans le cas de $i' \neq i$, c'est-à-dire où non seulement les variances, mais toutes les covariances aussi sont égales et non négatives. On peut alors montrer que

$$\sigma_{\bar{\varepsilon}\bar{\varepsilon}} = n^{-2} \sum_i \sum_{i'} \sigma_{ii'} \leq \sum_i \sum_{i'} \sigma_{ii'} f_i f_{i'} \leq n \sum_i f_i^2 \sigma_{\bar{\varepsilon}\bar{\varepsilon}},$$

où $n$ représente le nombre d'articles présents du groupe. La borne inférieure est obtenue dans le cas $f_i = 1/n$, et la borne supérieure dans le cas $\rho = 0$. Dans le premier cas, aucune correction de biais n'est nécessaire; dans le deuxième cas, nous prendrions $\hat{\Delta}(v) = \tilde{v} - \hat{v}$, où $\hat{v} = (t - t')n \sum_i f_i^2 \hat{\sigma}_{\bar{\varepsilon}\bar{\varepsilon}}$ et $\hat{\tilde{v}} = \{\sum_{t'+1}^{t} \gamma_\tau^2 + \gamma_{t'}^{*2} p_{t'-1}\} \hat{\sigma}_{\bar{\varepsilon}\bar{\varepsilon}}$. Ces valeurs correspondraient respectivement à $\hat{I}_{t't}$ et à $\tilde{I}_{t't}$.

## BIBLIOGRAPHIE


ALDRICH, J. (1992). Probability and depreciation: a history of the stochastic approach to index numbers. *History of Political Economy*, 24, 657-87.

BALK, B.M. (1980). A method for constructing price indexes for seasonal commodities. *Journal of the Royal Statistical Society, A*, 143, 68-75.

BALK, B.M. (1995). Axiomatic price theory: a survey. *Revue Internationale de Statistique*, 63, 69-93.

BRYAN, M.F., et CECCHETTI, S.G. (1993). The consumer price index as a measure of inflation. *Economic Review, Federal Reserve Bank of Cleveland*, 29, 15-24.

CLEMENTS, K.W., et IZAN, H.Y. (1981). A note on estimating Divisia index numbers. *International Economical Review*, 22, 745-747.

CLEMENTS, K.W., et IZAN, H.Y. (1987). The measurement of inflation: a stochastic approach. *Journal of Business and Economic Statistics*, 5, 339-350.

DIEWERT, W.E. (1987). Index numbers. Dans *The New Palgrave: A Dictionary of Economics*, (Éds. J. Eatwell, M. Milgate, et P. Newman). London: MacMillan.

DIEWERT, W.E. (1995). On the Stochastic Approach to Index Numbers. Document de discussion no. DP 95-31, Department of Economics, University of British Columbia.

EICHHORN, W., et VOELLER, J. (1976). *Theory of the Price Index*. Berlin: Springer-Verlag.

FISHER, I. (1922). *The Making of Index Numbers*. Boston: Houghton Mifflin.

HARVEY, A.C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

KEYNES, J.M. (1930). *A Treatise on Money*. New York: Harcourt, Brace and Company.

KOTT, P.S. (1984). A superpopulation approach to the design of price index estimators with small sampling biases. *Journal of Business and Economic Statistics*, 2, 83-90.

SELVANATHAN, E.A., et RAO, D.S.P. (1994). *Index Numbers: A Stochastic Approach*. Ann Arbor: The University of Michigan Press.

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 159-169
Statistique Canada



# Traitement de la non-réponse du cycle deux de l'enquête nationale sur la santé de la population

JEAN-LOUIS TAMBAY, IOANA ŞCHIOPU-KRATINA, JACQUELINE MAYDA, DIANA STUKEL et SYLVAIN NADON[1]

RÉSUMÉ

L'Enquête nationale sur la santé de la population (ENSP) est l'une des trois principales enquêtes-ménages longitudinales que mène Statistique Canada à une grande échelle auprès de la population canadienne. Depuis vingt ans, tous les deux ans, on a suivi un panel constitué d'environ 17 000 personnes. Les données provenant de l'enquête sont utilisées pour des analyses longitudinales, même si l'un des objectifs important est la production d'estimations transversales. Pour chaque cycle, les panélistes fournissent des renseignements détaillés sur leur santé (S) pendant qu'au même moment, pour augmenter l'échantillon transversal, des données socio-démographiques et quelques renseignements sur la santé sont recueillis (G) auprès de tous les membres des ménages. Cette stratégie de collecte présente différents schémas de réponse pour les panélistes après deux cycles: GS-GS, GS-G*, GS-**, G*-GS, G*-G* et G*-**, où «*» indique une portion de données manquantes. Le présent article explique la méthodologie élaborée pour traiter ces types de non-réponse longitudinale de même que la non-réponse d'une perspective transversale. L'utilisation de facteurs de pondération pour la non-réponse et la création de cellules d'ajustement pour la pondération à l'aide de l'algorithme CHAID sont expliquées ici.

MOTS CLÉS: Enquêtes longitudinales; traitement de la non-réponse; algorithmes CHAID.

## 1. INTRODUCTION

En 1996-1997, Statistique Canada a terminé la collecte de données pour le cycle 2 de l'Enquête nationale sur la santé de la population (ENSP). Cette enquête longitudinale a été lancée en 1994 afin d'obtenir des renseignements complets sur l'état de santé de la population canadienne ainsi que sur les facteurs déterminants en matière de santé. La population de l'enquête comprend des résidents de ménages et d'établissements de santé de l'ensemble du pays. Dans les provinces, le questionnaire rempli par les ménages se compose de deux principaux volets qui sont administrés à l'aide d'interviews assistées par ordinateur. Le volet Généralités sert à recueillir des données socio-démographiques et des renseignements de base sur la santé pour chaque membre du ménage. Le volet Santé sert à obtenir des données plus détaillées sur la santé du membre du ménage choisi pour participer au groupe longitudinal.

Même si l'ENSP est une enquête longitudinale, ses objectifs incluent la production d'estimations transversales périodiques (Catlin et Will 1992). La méthodologie de collecte des données reflète les besoins à la fois sur les plans longitudinal et transversal. Les panélistes choisis pour le cycle 1, sont suivis tous les deux ans pendant une période de 20 ans. Les personnes résidant avec les panélistes durant ces périodes fournissent des données pour le volet Généralités qui sont utilisées pour l'estimation transversale. Le champ d'observation de la composante transversale de l'enquête se rétrécissant avec le temps, il faut procéder régulièrement à une «bonification» de l'échantillon. La première bonification est prévue pour le cycle 3, en 1998.

Le présent document explique la méthodologie utilisée dans le Cycle 2 pour traiter la non-réponse au niveau des ménages et des personnes (avec indicateur pour la non-réponse à une question). La méthodologie se fonde sur la repondération des répondants au sein des sous-populations appelées cellules de pondération afin de tenir compte de la non-réponse. La repondération est une méthode couramment utilisée pour le traitement de la non-réponse à une question. Le biais et la variance découlant de cette méthode ont été étudiés par Thomsen (1973), Oh et Scheuren (1983), Kalton et Kasprzyk (1986) et Little (1986), entre autres. Si les cellules de pondération sont définies de telle sorte que la non-réponse survient presque totalement de façon aléatoire au sein de chaque cellule, alors le biais dû à la non-réponse peut devenir négligeable. Dans la même veine, David, Little, Samuhel et Triest (1983) ont étendu à la non-réponse la théorie élaborée par Rosenbaum et Rubin (1983) dans le contexte du jumelage du degré de propension. Leurs résultats supposent que la repondération peut rajuster le biais de la non-réponse lorsque les cellules de pondération sont composées en fonction de la propension à répondre.

On trouve à la section 2, une vue d'ensemble du plan d'échantillonnage et des données de sortie pour les deux premiers cycles. La section 3 explique les stratégies de traitement de la non-réponse et leurs résultats. Quant à la section 4, on y trouve les remarques finales. Il faut noter que la méthodologie présentée se rapporte aux échantillons de ménages des provinces; cette méthodologie ne couvre par les échantillons des territoires et des établissements.

[1] Jean-Louis Tambay, Ioana Şchiopu-Kratina, Jacqueline Mayda, Diana Stukel et Sylvain Nadon, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, 16ième étage, Immeuble R.H. Coats, Parc Tunney, Ottawa, (Ontario), Canada, K1A 0T6.

## 2. VUE D'ENSEMBLE DU PLAN D'ÉCHANTILLONNAGE DE L'ENSP ET DES DONNÉES DE SORTIE

### 2.1 Plan d'échantillonnage du cycle 1

Le premier échantillon de ménages a été constitué en 1994 à l'aide de l'outil de sélection d'un échantillon élaboré pour l'Enquête sur la population active (EPA) et, dans la province de Québec, en utilisant les logements qui ont participé à une enquête sur la santé menée par Santé Québec l'année précédente. Dans les deux cas, les ménages ou les logements ont été choisis au hasard dans des échantillons stratifiés de groupes sélectionnés selon la méthode de la probabilité proportionnelle à la taille. Les groupes ont été organisés en échantillons répétés et d'une période de collecte pour tenir compte du caractère saisonnier et pour permettre l'estimation de la variance. Il y a eu deux périodes de collecte estivale (juin et août) et deux périodes de collecte hivernale (novembre et mars 1995).

La figure 1 illustre le mécanisme de sélection du panel appliqué à l'extérieur du Québec. Les ménages de l'échantillon ont été désignés aléatoirement comme des ménages «adultes» ou «enfants», et comme admissibles ou non au filtrage, avant la collecte. Le filtrage a provoqué une augmentation dans le panel du nombre d'habitants provenant de ménages plus nombreux qui seraient sous-représentés si l'on choisissait seulement un membre par ménage, particulièrement les enfants et les adolescents. Les ménages admissibles au filtrage ont été rejetés de l'échantillon si aucun de leur membre n'avait moins de 25 ans. Le filtrage n'a pas été utilisé au Québec puisque les données provenant de l'enquête provinciale sur la santé permettaient l'application de différents taux de sous-échantillonnage par type de ménage et par taille.

| Type d'unité de sondage | Caractéristique du ménage | Sélection du panel limitée à: |
|---|---|---|
| Ménages «enfants» admissibles au filtrage (AF) | Aucun membre < 25 ans | s.o. – ménage rejeté |
| | Pas d'enfant, quelques membres < 25 ans | N'importe quel membre |
| | Présence d'enfants | Enfants |
| Ménages «enfants» non AF | Pas d'enfants présents | N'importe quel membre |
| | Présence d'enfants | Enfants |
| Ménages «adultes» | Tous | Membres de plus de 12 ans |

**Figure 1.** Mécanisme de sélection du panel à l'extérieur du Québec

La classification en ménages «adultes» et «enfants» a été effectuée pour des raisons opérationnelles: le questionnaire sur la santé, pour les enfants, ne serait pas disponible avant les périodes de collecte hivernales. Pour ce qui est des ménages «adultes», qui pouvaient être interviewés en tout temps, les enfants de moins de 12 ans n'étaient pas admissibles dans ce groupe. Les ménages «enfants», même ceux des groupes de la période estivale, ont été interviewés lors d'une période de collecte hivernale. Si des enfants étaient présents dans ces ménages, alors la sélection était limitée à eux. Pour réduire les distorsions saisonnières sur la charge de travail relative à la collecte de données et sur la représentativité du groupe découlant de ces procédures, un nombre moins grand de ménages ont été classés ménages «enfants» dans les groupes de la période estivale et, sauf une exception mineure, le filtrage a été appliqué seulement aux ménages «enfants».

Les provinces souhaitant améliorer les estimations au niveau infraprovincial pouvaient financer l'ajout d'effectifs à l'échantillon. Trois provinces se sont prévalues de cette possibilité en augmentant la taille de l'échantillon dans des régions ciblées. En Colombie-Britannique, un échantillon additionnel de 800 ménages a été constitué dans un établissement local de santé en utilisant le système de génération aléatoire de numéros de téléphone (GANT). La taille prévue de l'échantillon total dans les provinces était d'environ 23 000 ménages après le filtrage.

Ce qui précède donne une indication générale du plan d'échantillonnage de 1994, et ces renseignements sont suffisants pour les besoins du présent document. Les lecteurs désireux d'obtenir une présentation plus précise de l'échantillon de 1994 peuvent consulter Tambay et Catlin (1995) ou Statistique Canada (1995).

### 2.2 Pondération et résultats pour le cycle 1

Le principal résultat de l'ENSP consiste en un ensemble de fichiers de microdonnées à grande diffusion (FMGD) des réponses individuelles confidentielles (des versions internes de ces fichiers dont certains renseignements sont supprimés pour assurer la confidentialité sont aussi créés). Pour 1994, un FMGD pour le volet Généralités (58 400 dossiers) et un FMGD pour le volet Santé (17 600 dossiers) contenant les données recueillies auprès de chacun des membres des ménages et des membres sélectionnés dans le groupe sans enfants respectivement ont été publiés (Statistique Canada 1995).

Les poids propres à chaque dossier contenu dans un FMGD ont été calculés en appliquant une série d'ajustements à un poids de base représentant les fractions de sondage inverse (FSI) des ménages. Les FSI sont calculées en multipliant les poids des FSI originales ou des échantillons de Santé Québec par l'inverse de la fraction de sondage du sous-échantillon utilisé par l'ENSP. Pour ne pas alourdir le texte, nous décrirons seulement les principaux ajustements utilisés à l'extérieur du Québec.

Les ajustements apportés aux poids pour le FMGD du volet Généralités comprennent: 1) un ajustement pour la non-réponse des ménages; 2) un ajustement pour la procédure de filtrage; 3) un ajustement pour la non-réponse

des personnes [au sein des ménages répondants]; 4) un ajustement simple pour la stratification a postériori. L'ajustement 2) n'a été appliqué qu'aux ménages qui n'ont aucun membre de moins de 25 ans. La formule utilisée est $1/(1 - r_s)$, où $r_s$ représente la fraction de sondage du sous-échantillon pour le filtrage appliqué dans la strate. L'ajustement pour la stratification a postériori a été effectué séparément pour chaque classification croisée province-âge groupe-sexe. Les pondérations résultant de toutes les étapes précédentes sont multipliées par le rapport entre la population connue et la population estimée au sein de la classification croisée. Les chiffres relatifs à la population connue sont en fait des projections découlant du recensement.

Les ajustements apportés relativement à la non-réponse au niveau des ménages et des personnes (11,3 % et 1,4 %, respectivement) ont été appliqués aux unités déclarantes puisque les non-répondants ont été exclus des FMGD. Si $w_i$ est le poids de l'échantillon d'une unité $i$, le poids ajusté de la non-réponse, $w_{aj,i}$ se définit comme $w_{aj,i} = w_i(\sum_{tout} w_i)/(\sum_{rép} w_i)$, où les sommes s'effecturont sur toutes les unités d'échantillonnage et toutes les unités déclarantes, respectivement, à l'intérieur des cellules de pondération pour l'ajustement de la non-réponse. En raison d'un manque de renseignements de la part des ménages non répondants, les cellules de pondération pour la non-réponse des ménages ne sont que de simples classifications croisées des strates de l'ENSP et de la saison (c'est-à-dire, les groupes d'été par opposition aux groupes d'hiver). Les cellules de pondération pour la non-réponse des individus, dont le taux est très faible, sont les classifications croisées province-âge-sexe qui ont été utilisées pour l'ajustement de la stratification a postériori.

Les ajustements apportés aux poids pour le FMGD du volet Santé comprennent: 1) un ajustement pour la non-réponse des ménages; 2) un ajustement pour la procédure de filtrage; 3) un ajustement pour le sous-échantillon des ménages «enfants» et des ménages «adultes»; 4) un ajustement pour la sélection des panélistes de l'enquête longitudinale; 5) un ajustement pour la non-réponse des panélistes 6) un ajustement pour la stratification a postériori. Les deux premiers ajustements sont exactement les mêmes que pour le FMGD du volet Généralités. Comme le FMGD du volet Santé n'inclut pas les membres du groupe qui sont des enfants, l'ajustement 3) compense pour les ménages de l'échantillon où des membres non-enfants ne sont pas admissibles comme panélistes. Conséquemment, l'ajustement ne s'applique qu'aux ménages avec enfants et équivaut à $1/r$, où $r$ représente la proportion de ménages « adultes » dans l'échantillon. L'ajustement 4) est l'inverse de la probabilité que l'on ait sélectionné le panéliste. Les ajustements pour la non-réponse des panélistes (3,9 %) et pour la stratification a postériori sont similaires à ceux des FMGD pour le volet Généralités, et utilisent les mêmes classifications croisées province-âge-sexe. Même si les panélistes enfants ne sont pas inclus dans les FMGD du volet Santé, aux fins de l'enquête longitudinale, leur pondération dans l'échantillon est obtenue comme ci-dessus en utilisant $1/(1 - r)$ au lieu de $1/r$ comme à l'étape 3.

### 2.3 Plan d'échantillonnage du cycle 2

Pour le cycle 2, l'accent a davantage été mis sur l'estimation longitudinale: aucune «bonification» de l'échantillon n'a été prévue avant le cycle suivant. L'échantillon principal était donc constitué d'environ 17 000 panélistes et de leur ménage actuel. Les panélistes ont été suivis et on leur a administré les volets Généralités et Santé du questionnaire, alors que les autres membres du ménage n'ont répondu qu'au volet Généralités. Aucun suivi n'a été effectué auprès des ménages non répondants de 1994. En Alberta, au Manitoba et en Ontario d'importants échantillons supplémentaires (ne faisant pas partie de l'échantillon principal) ont été obtenus par génération aléatoire de numéros de téléphone (GANT), ce qui a permis la production d'estimations transversales au niveau infraprovincial. Pour chaque ménage contacté par GANT, un membre de plus de 12 ans a été choisi pour remplir le volet Santé du questionnaire. En Alberta et au Manitoba, dans les ménages contactés par GANT qui avaient des enfants, un enfant a aussi été choisi pour remplir le volet Santé.

Nous faisons remarquer, qu'aux fins de l'étude transversale, l'échantillon principal ne couvre pas très bien les nouveaux arrivants dans la population comme les nouveaux-nés et les immigrants récents. La population à qui on a administré le volet Généralités du questionnaire est constituée des résidents des ménages où au moins un des membres était admissible au cycle 1; les ménages composés entièrement de récents immigrants (et de leurs nouveaux-nés) sont donc exclus. La population à qui on a administré le volet Santé du questionnaire est constituée de personnes qui étaient admissibles au cycle 1: les immigrants récents et les enfants de moins de deux ans sont exclus de la population cible de l'échantillon principal (ils sont inclus dans la population cible pour la GANT). Pour les deux questionnaires, Généralités et Santé, la stratification après sélection est effectuée en utilisant des chiffres sur la population qui n'excluent pas les récents immigrants. Il en résulte que la population de récents immigrants est implicitement estimée par la population de non-immigrants parce que les dernières pondérations de l'échantillon principal sont ajustées à la hausse pour rendre compte des chiffres précédents. Il s'agit d'une contrainte qui est indiquée dans la documentation des FMGD. D'autres méthodes auraient pu être envisagées comme la stratification a postériori en utilisant uniquement la population non immigrante ou en ajustant d'une façon quelconque seulement les pondérations des immigrants moins récents (qui sont couverts) pour tenir compte des immigrants plus récents (qui ne le sont pas). Ces méthodes auraient été difficiles à appliquer lorsque, dans le cadre du questionnaire Généralités, une distinction doit être établie entre les ménages composés uniquement d'immigrants et les ménages mixtes dont certains membres sont des immigrants.

### 2.4 Données de sortie et pondération du cycle 2

La figure 2 constitue un sommaire des trois principaux fichiers de sortie prévus pour le cycle 2: un FMGD pour l'étude longitudinale; un FMGD pour le volet Santé de l'étude transversale et un FMGD pour le volet Généralités de l'étude transversale. Le FMGD de l'étude longitudinale contient des données sur les volets Généralités et Santé pour les deux cycles pour les 17 000 panélistes [nota: pour des raisons de confidentialité, il est possible que nous ne puissions publier le FMGD de l'étude longitudinale – auquel cas seul un fichier interne de microdonnées serait produit]. Le FMGD du volet Santé de l'étude transversale contient les renseignements de 1996 sur les volets Généralités et Santé pour environ 70 000 panélistes et membres sélectionnés par GANT. Le FMGD du volet Généralités de l'étude transversale contient les renseignements de 1996 sur le volet Généralités pour environ 220 000 membres de l'échantillon principal et du groupe GANT. Les procédés de pondération propres à chaque FMGD, présentés ci-après pour l'échantillon principal, sont décrits de façon plus détaillée dans l'étude de Stukel, Mohl et Tambay (1997).

| Fichier de sortie | FMGD ÉTUDE LONGITUDINALE | FMGD – SANTÉ ÉTUDE TRANSVERSALE | FMGD – GÉNÉRAL ÉTUDE TRANSVERSALE |
|---|---|---|---|
| Contenu | Généralités et Santé | Généralités et Santé | Généralités seulement |
| Échantillons | Principal seulement | Principal et GANT (3 prov.) | Principal et GANT (3 prov.) |
| Unités | Panéliste | Panélistes et GANT | Tous les membres des ménages |
| Taille | ≈ 17 000 dossiers | ≈ 70 000 dossiers | ≈ 220 000 dossiers |
| Pondération (échantillon principale) | 1. Pondération pour l'année de base<br>2. Ajust. panélistes non répondants<br>3. Stratification a postériori | 1. Pondération pour l'année de base<br>2. Ajust. panélistes non répondants<br>3. Intégration échant. principal et GANT<br>4. Stratification a postériori | 1. Pondération pour l'année de base<br>2. Ajust. ménager non répondants<br>3. Ajust. répartition du poids<br>4. Ajust. membres des ménages NR<br>5. Intégration échant. principal et GANT<br>6. Stratification a postériori |

**Figure 2.** Description des fichiers de sortie pour le cycle 2

On obtient les pondérations pour le FMGD de l'étude longitudinale en ajustant la pondération pour l'année de base d'abord pour les non-réponses du groupe de 1996, puis pour la stratification a postériori. La pondération pour l'année de base représente la fraction de sondage inverse pour 1994, y compris tous les ajustements pour le FMGD du volet Santé décrits à la section 2.2 jusqu'à l'ajustement 4) pour la sélection du panel (une correction est nécessaire pour le «retrait»des additions provinciales à l'échantillon de 1994). Le facteur de compensation de la non-réponse est traité dans la prochaine section et y sera décrit. La stratification a postériori est effectuée pour reproduire les chiffres de 1994 pour la population des provinces selon l'âge et le sexe.

Pour ce qui est du FMGD du volet Santé de l'étude transversale, le processus de pondération pour les panélistes (échantillon principal) nécessite trois ou quatre étapes. Habituellement, la pondération de l'année de base est ajustée pour la non-réponse des panélistes, selon les explications qui se trouvent dans la prochaine section, ainsi que pour la stratification a postériori (pour établir la correspondance avec les chiffres de population régionaux ou provinciaux de 1996, selon l'âge et le sexe). Dans les provinces où une partie de l'échantillon provient de la GANT, l'étape supplémentaire consiste à intégrer l'échantillon composé de répondants sélectionnés par la GANT. On obtient l'estimation intégrée par une forme d'adaptation minimale de l'estimateur double Skinner-Rao (Skinner Rao 1996).

Quant au FMGD du volet Généralités de l'étude transversale, le processus de pondération pour l'échantillon principal nécessite cinq ou six étapes. Premièrement, une fois de plus, il faut calculer la pondération de l'année de base. Il faut ensuite procéder à un ajustement pour la non-réponse au niveau des ménages, dont traitera la prochaine section. La prochaine étape consiste à appliquer la méthode de répartition du poids. Cette méthode a été décrite par Ernst (1989) et mise au point plus tard par Lavallée (1995). Le poids du panéliste divisé par le nombre de personnes de son ménage qui étaient admissibles au cycle 1 est attribué à tous les membres du ménage, y compris ceux qui n'étaient pas admissibles au cycle 1 (c'est-à-dire nouveaux-nés, immigrants). Cette méthode est non biaisée pour les estimations des totaux pour la population des ménages où au moins un membre était admissible au cycle 1. La prochaine étape consiste à établir l'ajustement pour la non-réponse d'un membre du ménage. Dans les provinces qui ont eu recours à la GANT, cette étape est suivie de l'intégration de l'échantillon principal avec l'échantillon GANT (cette fois pour tous les âges). La stratification a postériori est effectuée de la même façon que pour le FMGD du volet Santé de l'étude transversale.

## 3. STRATÉGIE RELATIVE À LA NON-RÉPONSE DANS L'ÉCHANTILLON PRINCIPAL DU CYCLE 2

La présente section explique la stratégie adoptée pour le traitement de la non-réponse du cycle 2 de l'échantillon principal (non GANT). L'ajustement pour la non-réponse a encore une fois été effectué à l'aide de la méthode de la cellule de pondération sauf que, cette fois-ci, les données du cycle 1 étaient disponibles, ce qui a permis de créer des cellules plus homogènes en regard de la propension à répondre. Dans la section 3.1, on définit les non-répondants de l'ENSP. Dans la section 3.2, on explique deux méthodes générales utilisées pour créer des cellules de pondération, dont celle choisie pour l'ENSP. La stratégie d'ajustement pour la non-réponse est expliquée dans la section 3.3 alors qu'à la section 3.4, on décrit la création des cellules de pondération de la non-réponse.

### 3.1 Définition des non-répondants et des unités non admissibles

La première étape du traitement de la non-réponse consiste en sa définition. Pour le cycle 2, les questionnaires ont été complètement remplis par 89 % de l'échantillon principal et partiellement remplis par 3 % de l'échantillon. Le reste de l'échantillon se compose de refus (3,1 %), de cas où le panéliste n'a pu être retracé (1,7 %), de décès (1,7 %) de départs du Canada ou de placements en établissement (0,4 %), ainsi que d'autres types de non-réponse comme les absences temporaires et les circonstances spéciales (0,7 %). Parmi les ménages répondants, le taux de non-réponse au niveau des personnes est très faible: 1,8 % pour le questionnaire Généralités et 1,1 % pour le questionnaire Santé. Nous définissons d'abord les cas qui sont des non-réponses aux fins des études longitudinale et transversale.

Aux fins de l'étude longitudinale, un décès est considéré comme une réponse valide à l'enquête. Les panélistes qui sont décédés avant le cycle 2 voient leur état enregistré comme tel et la portion 1996 des données qui les concernent codée comme «sans objet» dans le fichier de données de l'étude longitudinale. Les panélistes qui sont déménagés dans un établissement ou dans les territoires ont fait l'objet d'un suivi et leurs réponses ont été utilisées aux fins de l'étude longitudinale. Les panélistes qui ont quitté le pays n'ont pas fait l'objet d'un suivi, mais ils ont été traités comme des non-répondants à l'étude longitudinale, même s'il aurait été plus exact aux fins de certaines analyses de les considérer comme ne faisant plus partie du champ de l'enquête. Nous avons choisi ce traitement parce que si ces personnes revenaient au Canada, elles seraient de nouveau admissibles.

Aux fins de l'étude transversale, tous les cas présentés au paragraphe précédent ont été traités comme des situations ne faisant pas partie du champ de l'enquête. Ce choix est acceptable parce que des outils d'enquête distincts pour les établissements et pour les territoires ont été utilisés pour la couverture transversale de ces populations en particulier, qui font partie du champ de l'enquête. Les unités non admissibles ne font pas partie de l'ENSP, mais, comme elles représentent d'autres unités de ce genre, elles ont été traitées aux fins de la pondération comme des répondants à toutes les étapes de l'ajustement de la pondération, sauf celle de l'intégration et celle de la stratification a postériori.

Les refus et les cas où des questionnaires étaient manquants pour des raisons autres que celles données précédemment ont été définis comme des non-réponses. Comme nous le verrons, une distinction a été faite ultérieurement entre les non-répondants complets ou partiels afin de tenir compte de différents utilisateurs.

### 3.2 Méthode de création de cellules de pondération pour l'ajustement de la non-réponse

Il existe deux approches statistiques à la création de cellules de pondération de la réponse qui font appel à la modélisation par segmentation et à la régression logistique. On trouve un exemple de cette dernière approche dans Hirabayashi, Little et Rubin (1992). Les auteurs ont établi des estimations de l'impôt anticipé à partir de déclarations hâtives en créant des cellules de pondération pour l'ajustement à partir de niveaux de propension à remettre une déclaration hâtive. La régression logistique permettait d'estimer la propension des déclarants à soumettre précocement leur déclaration. L'Enquête longitudinale sur la dynamique du travail et du revenu (EDTR) constitue un autre exemple d'application de la régression logistique (Grondin 1996). On calcule la régression des indicateurs de réponse des unités d'échantillonnage selon des caractéristiques connues (dichotomiques). On a créé les cellules d'ajustement de la non-réponse par recoupement des unités d'échantillonnage, à l'aide de toutes les covariables significatives. Afin de tenir compte des exigences quant à la taille minimale des cellules et aux taux de réponse, on a dû fusionner des cellules, à partir de celles partageant toutes les covariables sauf les moins significatives.

Dans l'approche de modélisation des segments, on génère un arbre de décision à partir des données; pour ce faire, on divise les ensembles de données en séquence de façon que, à chaque noeud, le prédicteur le plus significatif pour la variable de la réponse serve de point de départ à la division suivante. La division se poursuit jusqu'à ce qu'il ne soit plus possible de trouver une variable significative pour effectuer une autre division, ou qu'il ne soit plus possible de respecter les exigences quant à la taille minimale de la cellule. L'une des premières applications de la méthode de modélisation des segments pour l'ajustement de la non-réponse portait sur la Panel Study of Income Dynamics (Institute for Social Research 1979). En raison de ses avantages, énoncés ci-après, on a adopté pour l'ENSP l'approche de modélisation des segments fondée sur l'algorithme CHAID (détection automatique d'interactions du chi carré), inventé par Kass (1980). L'algorithme CHAID utilise des tests du $\chi^2$ pour diviser les prédicteurs catégoriques et retenir la division la plus significative à chaque étape. La division, en deux catégories ou plus, n'est pas exécutée de la même façon pour les prédicteurs ordonnés et non ordonnés. L'algorithme CHAID a été appliqué à l'aide du logiciel Knowledge Seeker (ANGOSS Software, 1995). Il faut noter que le logiciel applique l'algorithme aux prédicteurs continus en les traduisant tout d'abord en variables discrètes ordonnées.

Diverses études traitent des avantages et des inconvénients des approches logistiques et CHAID (se reporter par exemple à Kalton et Kasprzyk 1986). Le fondement théorique de la régression logistique est bien connu de nombreux analystes, et la programmation est possible dans plusieurs logiciels d'analyse statistique normalisée. Cette méthode permet entre autres d'obtenir des estimations individuelles de la propension à répondre, que l'on applique directement pour ajuster les poids des répondants. Cependant, pour arriver à une vitesse d'exécution raisonnable du

programme, il vaut mieux limiter le nombre de variables et de termes d'interaction. La fusion des cellules peut aussi s'avérer complexe, comme ce fut le cas pour l'EDTR, déjà mentionnée. Au contraire, l'algorithme CHAID permet de traiter un grande nombre de covariables et, en raison de la structure en arbre de décision, il est facile d'établir les interactions. Qui plus est, il est facile d'incorporer les exigences quant à la taille minimale des cellules comme paramètres d'exécution de programme. Parmi les principaux inconvénients, citons le fondement théorique moins familier (le logiciel Knowledge Seeker est vendu comme un outil d'«exploration des données»), la documentation limitée et le peu de logiciels d'exécution. À ceux-là s'ajoute l'impossibilité d'incorporer le plan d'échantillonnage pour adapter les modèles logistiques aux données de l'enquête, comme c'est le cas pour certains programmes statistiques tels que SUDAAN et PC CARP. Pour pallier à ces contraintes dans l'ENSP, on a inclus comme prédicteurs des caractéristiques liées au plan d'échantillonnage (se reporter à la section 3.4).

Deux études empiriques comparant les approches logistique et CHAID pour l'analyse de la non-réponse obtiennent des résultats divergents. Ainsi, Rizzo, Kalton et Brick (1996) n'ont pas conclus à des différences notables entre les deux approches appliquées à la Survey of Income and Program Participation. Par contre, Dufour, Gagnon, Morin, Renaud et Särndal (1998), dans une étude simulée de l'EDTR, ont trouvé que l'application de l'approche CHAID pour l'ajustement de la non-réponse produisait un biais moins important.

### 3.3 Ajustement de la non-réponse dans l'échantillon principal

Il a fallu établir une méthode d'ajustement de la non-réponse pour chacun des FMGD, soit les fichiers de données longitudinales, Généralités (transversales) et Santé (transversales). Le FMGD Généralités sera traité en premier.

Comme le démontre la figure 2, la stratégie de pondération appliquée au FMGD Généralités a exigé d'ajuster séparément la non-réponse aux niveaux des ménages et des personnes. Quand on a créé des cellules d'ajustement de la non-réponse au niveau des ménages, on a considéré les caractéristiques des panélistes ainsi que celles du ménage comme étant des prédicteurs de non-réponse, pour trois raisons. Premièrement, étant donné que le panéliste repré sentait le lien avec le ménage dans le cycle 2, on considérait que ses caractéristiques avaient permis de trouver le ménage et d'obtenir une réponse (cette personne a souvent permis d'établir le premier contact). Deuxièmement, quelques caractéristiques personnelles du panéliste, telles que la race, constituent jusqu'à un certain point des caractéristiques du ménage. Enfin, l'utilisation des caractéristiques du panéliste n'était pas incompatible avec la nécessité qui nous incombait d'utiliser divers renseignements pour construire les cellules de pondération. Si les caractéristiques du panéliste ne sont pas significatives, alors l'algorithme CHAID ne les retient tout simplement pas.

Pour le volet Généralités, la non-réponse au niveau individuel indique qu'une information est disponible pour certains membres seulement d'un ménage, en raison d'un refus de l'un des membres ou d'une absence temporaire. Le taux de non-réponse au niveau individuel étant de 1,8 % seulement, il a été décidé que la création de cellules de pondération en fonction des catégories province, âge et sexe (comme dans le Cycle 1) suffirait amplement.

Contrairement au FMGD Généralités, on a pu combiner l'ajustement de la non-réponse aux niveaux des ménages et des personnes pour les FMGD longitudinal et Santé, parce qu'il est appliqué aux réponses d'une même personne (le panéliste). On a donc créé un seul ensemble de cellules d'ajustement.

Pour ce qui est du FMGD longitudinal, on a noté que les éléments de données provenaient à la fois des volets Généralités et Santé, mais que les taux de réponse différaient d'un volet à l'autre. Cet écart a entraîné des schémas de réponse différents pour les cycles 1 et 2, comme suit: GS-GS; GS-G*; G*-GS; G*-G*. À ces différences s'ajoutent les schémas de non-réponse longitudinale GS-** et G*-** (les lettres correspondent aux réponses pour les volets dans chaque cycle, et l'astérisque correspond à une non-réponse). Afin de maximiser l'utilité des données, on a décidé d'ajuster de deux façons la non-réponse longitudinale. L'une des méthodes d'ajustement est appliquée au schéma «réponse longitudinale complète» GS-GS. Autrement dit, tous les autres schémas de réponse sont considérés comme étant des non-réponses. L'autre méthode d'ajustement s'applique à la «réponse longitudinale partielle», pour les cas où on a obtenu au moins les données générales pour chacun des cycles (schémas GS-GS, GS-G*, G*-GS et G*-G*). Le groupe de données Réponse complète sera utile aux chercheurs qui souhaitent analyser un ensemble de données longitudinales complètes couvrant tout le contenu du questionnaire. Le groupe de données Réponse partielle peut servir aux chercheurs qui s'intéressent surtout aux types de variables du questionnaire général. On le constate dans le tableau ci-dessous, la Réponse longitudinale partielle dépasse de 3 % seulement la Réponse longitudinale complète.

**Tableau 1**
Schémas de la réponse longitudinale

| Type de réponse | | Schéma de réponse Cycles 1-2 | Nombre de dossiers |
|---|---|---|---|
| Complète | Partielle | | |
| ■ | ■ | GS-GS | 15 670 |
| | ■ | GS-G* | 110 |
| | ■ | G*-GS | 366 |
| | ■ | G*-G* | 22 |
| | | GS-** | 1 014 |
| | | G*-** | 94 |
| | | Total | 17 276 |

À la lumière des données du tableau, il faut construire des cellules d'ajustement pour cinq types de réponse (ou de non-réponse) dans le cycle 2, comme suit:

- FMGD Généralités – réponse des ménages
- FMGD Généralités – réponse des personnes
- FMGD Santé – réponse combinée
- FMGD longitudinal – réponse complète
- FMGD longitudinal – réponse partielle

Seulement trois groupes de cellules d'ajustement ont été créées pour ces types de réponse. Les cellules d'ajustement créées pour la réponse au niveau des ménages dans le FMGD Généralités ont aussi servi pour les réponses partielles dans le FMGD longitudinal, puisque la réponse partielle au niveau des ménages correspondait presque toujours à une réponse partielle au niveau des personnes dans le fichier longitudinal (sauf pour 53 individus). De même, les cellules d'ajustement créées pour la réponse complète dans le FMGD longitudinal ont été appliquées à la réponse dans le FMGD Santé. Bien qu'on ait obtenu 366 réponses de plus pour le dernier type (schéma G*-GS), on a considéré que le même mécanisme de réponse était à l'oeuvre. Le troisième groupe de cellules d'ajustement a été appliqué à la réponse au niveau des personnes dans le FMGD Généralités. On a utilisé les catégories province – âge – sexe, comme ce fut le cas au cycle 1.

Incidemment, bien que les mêmes cellules d'ajustement puissent être utilisées pour différents groupes de données, on effectue des calculs distincts pour chaque type de groupe de données. Ainsi, les 366 dossiers assortis d'un schéma de réponse G*-GS sont traités comme répondants dans le processus d'ajustement des poids pour le FMGD Santé, mais comme non-répondants dans le processus d'ajustement des poids pour l'analyse de la réponse complète dans le FMGD longitudinal.

### 3.4 Création de cellules d'ajustement de la pondération

On a créé des groupes distincts de cellules d'ajustement de la pondération pour chaque province. Tout d'abord, il a fallu déterminer les variables à prendre en compte. Avec l'approche CHAID, le nombre de variables à considérer importe peu, et on a tenu compte de différents types de variables. Bien entendu, on a pris en compte les caractéristiques du ménage, ou du logement, ainsi que les caractéristiques personnelles du panéliste. Afin d'être en mesure d'incorporer le plan d'enquête dans l'analyse, nous avons aussi tenu compte de certaines caractéristiques liées à ce plan ou au poids d'échantillonnage. Parmi celles-ci se trouvent les variables géographiques telles que le code Région métropolitaine de recensement (RMR) ou l'indicateur urbain/rural, les variables d'échantillonnage particulières au cycle 1, telles que les indicateurs de ménages visés par le filtrage et le type de ménage «adulte/enfant», de même que les caractéristiques liées à l'application de ces variables d'échantillonnage, telles que la présence dans un ménage d'un membre âgé de moins de 25 ans ou d'un enfant. La taille du ménage a été considérée comme étant une caractéristique du ménage, et on l'a aussi liée au poids de l'échantillon. À la lumière des expériences passées, on a décidé d'inclure, outre la caractéristique revenu du ménage, une caractéristique fictive indiquant si le revenu du ménage avait été déclaré dans le cycle 1. Les changements d'adresse pouvant conduire à une incapacité de retracer la provenance d'une non-réponse, il aurait été utile d'intégrer un identificateur de changement d'adresse. Cependant, dans certains cas de non-réponse et d'absence de contact, il était difficile de déterminer si le panéliste avait vraiment déménagé. Enfin, nous avons utilisé pour l'analyse une variable «personne ayant déménagé», qui nous permet de savoir si le panéliste avait changé de province entre les cycles. Cette modalité représente un pis-aller parce que la valeur par défaut est «non». Des caractéristiques personnelles tirées du volet Santé de l'enquête, telles que le statut fumeur/consommateur d'alcool, le niveau d'indice de santé et l'échelle de santé/souffrance morale n'ont pas été utilisées parce qu'elles étaient inconnues pour quelque 500 panélistes.

On trouve ci-après la liste des variables qui ont été utilisées. La valeur de l'indicateur de non-réponse, soit la variable dépendante, a été assignée en fonction de la définition donnée au concept de non-réponse.

#### VARIABLES D'ÉCHANTILLONNAGE/GÉOGRAPHIQUES

| | |
|---|---|
| PROVINCE | Analyse effectuée à l'échelle provinciale |
| CMA | Région métropolitaine de recensement (0 s'il ne s'agit pas d'une RMR) |
| URBAN | Indicateur urbain/rural |
| REJECT | Indicateur qu'une unité (ménage) est sujette au filtrage |
| ACFLAG | Classe d'échantillonnage Adulte/enfant pour l'unité |

#### CARACTÉRISTIQUES DU LOGEMENT/MÉNAGE

| | |
|---|---|
| DWELL | Type de logement (10 catégories) |
| OWNER | Indicateur propriétaire/locataire |
| FAMTYP | Type de famille (personne seule, famille monoparentale, couple marié, autre) |
| INC | Suffisance du revenu du ménage (5 niveaux) |
| INCNR | Indicateur de non-réponse pour le revenu |
| INCSRC | Principale source de revenu (6 catégories) |
| *HHSIZE | Taille du ménage |
| UND25 | Indicateur membre de moins de 25 ans |
| KIDS | Indicateur enfant de moins de 12 ans |

**CARACTÉRISTIQUES PERSONNELLES DU PANÉLISTE**

| | |
|---|---|
| SEX | Sexe |
| AGE | Âge en années |
| AGE16 | Indicateur personne âgée de 16 ans et plus |
| MARIT | État civil |
| FAMID | Identificateur famille à l'intérieur du ménage (A, B, C, ...) |
| RACE | Blanc, Noir, Autochtone ou autre |
| BORN | Lieu de naissance (Canada, É.-U./Mexique Amérique S./Afrique, Europe/Australie, Asie) |
| AGIMM | Âge au moment de l'immigration pour les immigrants |
| *MOVED | Indicateur changement de province (voir le texte) |
| *EDUC | Niveau de scolarité le plus élevé (12 catégories) |
| *STUDNT | Indicateur étudiant |
| MACT | Activité principale (8 catégories) |
| *NUMJOB | Nombre d'emplois occupés l'an dernier (cycle 1) |
| RESTR | Indicateur pour activité restreinte |
| *CAUSE | Cause principale de restriction (12 catégories) |
| CONSUL | Nombre de consultations auprès d'un médecin |
| INHOSP | Indicateur pour séjour de 1 nuit à l'hôpital |
| *CHRONIC | Nombre de maladies chroniques |

* Indique que la variable n'a jamais été significative lors de l'établissement de classes.

La figure 3 énonce les variables choisies par l'algorithme CHAID en vue de créer des cellules d'ajustement de la non-réponse pour la réponse au niveau des ménages et pour la réponse longitudinale complète dans chaque province. À des fins de confidentialité, on ne peut donner le détail sur les tailles des cellules individuelles et les taux de réponse (certaines des variables utilisées sont considérées comme étant sensibles et ne figurent pas dans les FMGD). Cependant, on trouvera de l'information sommaire sur la construction des cellules aux tableaux 2 et 3.

Les résultats varient d'une province à l'autre. Comme il était prévu, les provinces où l'échantillon est plus important, telles que l'Ontario, le Québec, la Colombie-Britannique et l'Alberta, ont permis de tracer des arbres de décision «plus riches». Les tailles des cellules ainsi que les taux de réponse varient aussi passablement. Dans le tableau 2, qui porte sur la réponse au niveau des ménages, une seule cellule est attribuée au Manitoba, et une seule cellule contient 88 % de l'échantillon du Nouveau-Brunswick. De même, dans le tableau 3, une grande partie de l'échantillon de Terre-Neuve est placée dans une seule des deux cellules. On constate pour quelques provinces des cellules avec des taux de non-réponse autour de 40 %.

La figure 3 montre les différences entre les caractéristiques des classes de pondération pour les différentes provinces, mais aussi pour les deux types de non-réponse à l'intérieur des provinces. Dans toutes les provinces, sauf en Alberta, l'algorithme CHAID utilise des caractéristiques distinctes pour les deux types de non-réponse, dès le premier ou le deuxième niveau d'embranchement. Quelques caractéristiques sont dominantes aux premiers niveaux d'embranchement dans de nombreux arbres, pour les deux types de non-réponse, soit le niveau de suffisance du revenu du ménage (INCNR), l'indicateur de non-réponse quant au revenu (INCNR), la race (RACE), et le lieu de naissance (BORN).

**Tableau 2**
Caractéristiques des cellules d'ajustement de la réponse (réponse au niveau des ménages)

| Prov. | $N^{bre}$ Unités | $N^{bre}$ NR | Tailles cellules | | | % par cellules NR | | |
|---|---|---|---|---|---|---|---|---|
| | | | min. | max. | moy. | min. | max. | moy. |
| T.-N. | 1 082 | 40 | 354 | 728 | 541 | 1,4 | 4,8 | 3,7 |
| Î.P.-E. | 1 037 | 51 | 81 | 478 | 259 | 3,0 | 13,6 | 4,9 |
| N-É. | 1 085 | 55 | 46 | 374 | 217 | 0,7 | 10,9 | 5,1 |
| N.-B. | 1 125 | 59 | 32 | 986 | 281 | 2,6 | 34,4 | 5,2 |
| QC | 3 000 | 133 | 123 | 2 363 | 750 | 1,8 | 12,1 | 4,4 |
| Ont. | 4 307 | 315 | 44 | 1 038 | 308 | 0,9 | 25,8 | 7,3 |
| Man. | 1 205 | 50 | 1 205 | 1 205 | 1 205 | 4,1 | 4,1 | 4,1 |
| Sask. | 1 168 | 59 | 37 | 626 | 167 | 1,6 | 35,3 | 5,1 |
| Alb. | 1 544 | 116 | 32 | 837 | 221 | 3,9 | 36,7 | 7,5 |
| C.-B. | 1 723 | 149 | 82 | 678 | 246 | 5,2 | 29,0 | 8,6 |

**Tableau 3**
Caractéristiques des cellules d'ajustement de la réponse (réponse longitudinale complète)

| Prov. | $N^{bre}$ unités | $N^{bre}$ NR | Tailles cellules | | | % par cellules NR | | |
|---|---|---|---|---|---|---|---|---|
| | | | min. | max. | moy. | min. | max. | moy. |
| T.-N. | 1 082 | 73 | 35 | 1 047 | 541 | 6,2 | 22,9 | 6,7 |
| Î.P.-E. | 1 037 | 80 | 41 | 453 | 207 | 4,1 | 26,8 | 7,7 |
| N.-É. | 1 085 | 96 | 236 | 555 | 362 | 6,5 | 14,3 | 8,8 |
| N.-B. | 1 125 | 86 | 59 | 819 | 281 | 4,8 | 16,8 | 7,6 |
| QC | 3 000 | 211 | 42 | 2 202 | 375 | 2,5 | 37,8 | 7,0 |
| Ont. | 4 307 | 470 | 34 | 619 | 196 | 0,0 | 38,0 | 10,9 |
| Man. | 1 205 | 91 | 186 | 763 | 402 | 5,6 | 15,1 | 7,6 |
| Sask. | 1 168 | 83 | 90 | 339 | 195 | 0,0 | 28,9 | 7,1 |
| Alb. | 1 544 | 148 | 41 | 866 | 172 | 1,1 | 39,0 | 9,6 |
| C.-B. | 1 723 | 192 | 33 | 408 | 191 | 4,5 | 37,3 | 11,1 |

Dans la figure 3a, le revenu du ménage et les variables liées (INCNR et INCSRC), le statut propriétaire/locataire (OWNER), la race, le lieu de naissance et le type de logement (DWELL) ont toutes été utilisées trois fois ou plus pour la création des classes de pondération de la non-réponse au niveau des ménages. On remarque aussi que dans cinq des neuf provinces, une caractéristique personnelle des panélistes

a été sélectionnée au premier embranchement par l'algorithme CHAID. Cette constatation vient confirmer qu'il a été judicieux de prendre en compte les caractéristiques personnelles pour ajuster la non-réponse au niveau des ménages.

Dans la figure 3b, Non-réponse longitudinale complète, la Région métropolitaine de recensement (CMA), l'état civil (MARIT) et le sexe, malgré leur importance moindre que les caractéristiques de revenu, de race et de lieu de naissance, ont été utilisées le plus souvent (cinq fois chacune).

Comme il a déjà été souligné, les variables d'échantillonnage telles que les indicateurs filtrage (REJECT) et ménages «Adulte/Enfant» (ACFLAG) ont été prises en compte afin que soit incorporé le plan d'enquête dans les analyses exécutées avec l'algorithme CHAID. Bien que ces variables aient été sélectionnées une fois seulement chacune, les caractéristiques du ménage utilisées dans le plan, telles que la présence d'enfants (KIDS) et de membres de moins de 25 ans (UND25) ont été sélectionnées à l'occasion. La taille du ménage n'a pas été utilisée, mais le type de famille (FAMTYP), une variable liée à la taille, a été sélectionné deux fois.

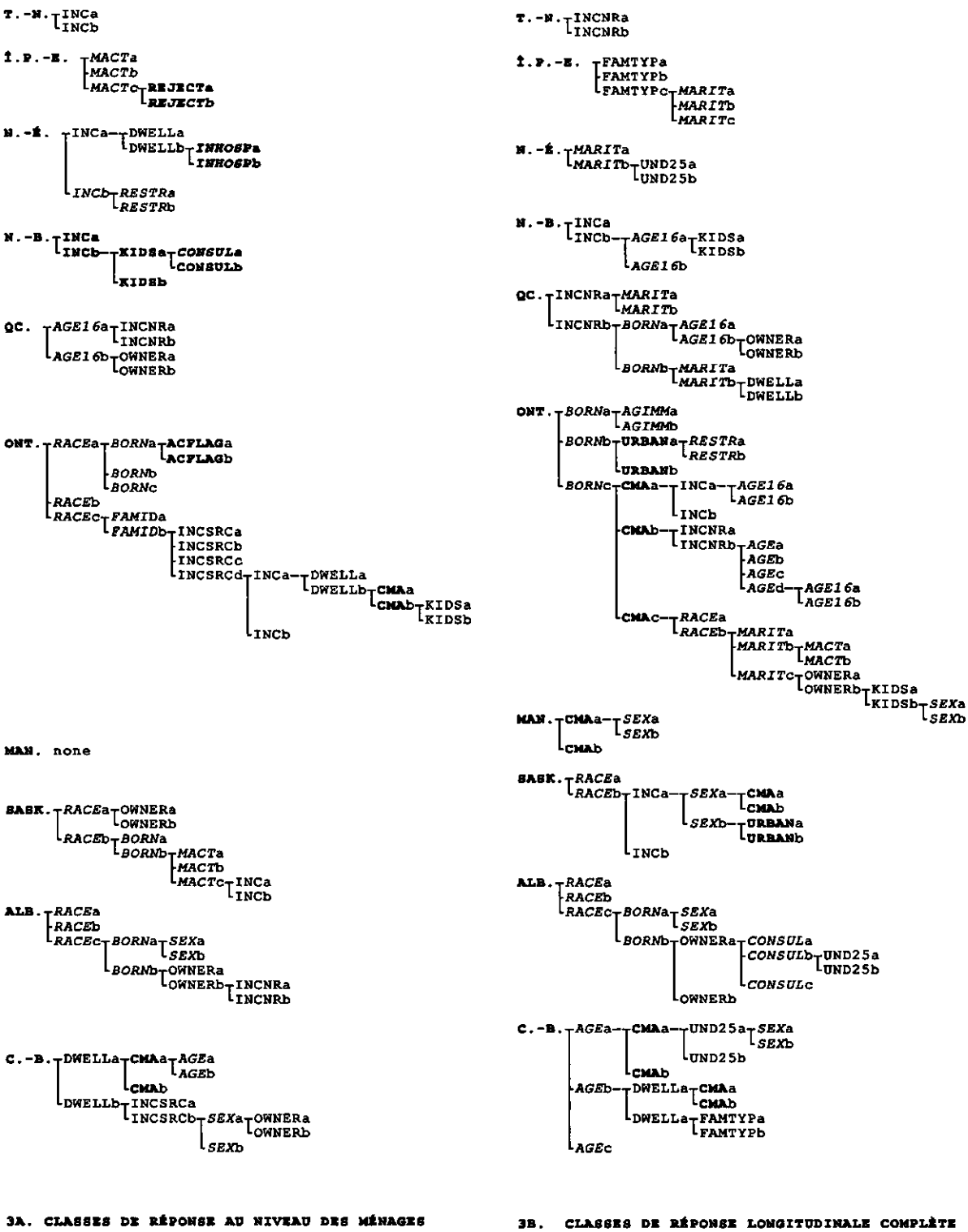
Les cellules d'ajustement produites par l'algorithme CHAID ont été révisées, mais très peu de modifications ont été apportées. À l'intérieur de chacune des cellules, le poids des unités déclarantes a été réparti au prorata, afin d'obtenir le poids total pour les unités de déclaration et de non-déclaration. L'ampleur de l'ajustement au poids des unités non déclarantes n'a jamais excédé 1,83.

## 4. CONCLUSION

Le présent document explique la stratégie utilisée pour le traitement de la non-réponse dans les volets longitudinal et transversal du cycle 2 de l'ENSP. L'approche adoptée permettait de tenir compte de considérations pratiques telles que la nécessité de se doter d'un moyen facile, valide sur le plan statistique, pour définir les cellules d'ajustement de la pondération et de produire un ensemble de données plus utile (par exemple, nous avons ajusté séparément la réponse longitudinale complète et partielle). Il fallait de plus éviter de compliquer indûment le processus; c'est pourquoi nous avons utilisé à divers escients les mêmes cellules d'ajustement de la pondération. Nous avons choisi l'algorithme CHAID au détriment de la régression logistique parce que la première approche nous donnait plus de liberté quant au nombre et à la diversité des variables pouvant être prises en compte. Ainsi, nous avons pu tenir compte de nombreuses variables d'échantillonnage et caractéristiques, pour en retenir quelques-unes seulement. Il nous semblait ainsi que, nous pourrions tirer profit de ces caractéristiques dans le traitement de la non-réponse.

Par ailleurs, le calendrier très serré auquel nous étions soumis nous a obligés à laisser de côté certaines analyses. Il aurait été intéressant d'approfondir les possibilités offertes par l'algorithme CHAID. Par exemple, comme ce dernier permet d'utiliser une variable de réponse nominale, nous aurions pu classer les unités d'échantillonnage en trois groupes, soit les déclarants vivants, les déclarants décédés ou hors du champ de l'enquête, et les non-déclarants. Nous aurions aussi aimé comparer une analyse effectuée avec l'algorithme CHAID et une autre avec la méthode de la régression logistique. Nous aurions pu utiliser les variables du volet Santé, telles que l'indice de santé ou le statut fumeur/consommateur d'alcool, pour définir les cellules d'ajustement de la pondération, bien que leur utilité aurait été amoindrie du fait de leur absence dans certaines unités (elles sont absentes des schémas de réponse G*-GS, G*-G* et G*-**). Il faudrait reconsidérer la décision d'utiliser les mêmes cellules d'ajustement de la pondération pour différents types de non-réponse. Aurait-il mieux valu utiliser les cellules d'ajustement créées pour la réponse au niveau des ménages pour l'analyse de la non-réponse transversale dans le volet Santé? Pour comparer l'efficacité des diverses stratégies d'ajustement de la non-réponse, il faudrait évaluer leur incidence sur la variance des estimateurs. Nous pourrions de la même façon évaluer l'incidence de l'ajustement de la non-réponse dans le cycle 2 sur le biais de la non-réponse, en utilisant les données du cycle 1 qui sont disponibles pour tous les panélistes. Les estimations portant sur l'ensemble de l'échantillon pourraient être comparées aux estimations ajustées en fonction de la non-réponse, générées à partir des unités déclarantes.

Le cycle 3 présentera de nouvelles difficultés. Un «aplanissement» de l'ensemble de l'échantillon est prévu pour cette année, qui influera sur la stratégie d'estimation transversale et, partant, sur le traitement de la non-réponse. Étant donné que la non-réponse dans le volet longitudinal augmentera, il faudra tenir compte des effets secondaires de l'ajustement de la pondération, tels que la création possible de valeurs aberrantes de pondération. Il deviendra ardu d'établir des groupes de pondération pour différents types d'analyses longitudinales, parce que le nombre de schémas de réponse partielle augmentera. Combien de schémas peut-on traiter de front sans difficulté, et lesquels choisir? Il faudrait évaluer la pertinence des données supplémentaires, telles que le statut «personne déménagée», pour le traitement de la non-réponse. On devra probablement recourir à l'imputation pour la non-réponse dans le cycle 3. Comment pourra-t-on concilier le procédé d'imputation avec la méthode d'ajustement de la pondération dans le traitement de la non-réponse? Il reste beaucoup de travail autour de l'ENSP. Il faut espérer que nous aurons le temps de régler certaines questions en suspens avant que ne s'achève le traitement de la réponse pour le cycle 3.




(TYPES DE CARACTÉRISTIQUES: **ÉCHANTILLONNAGE/GÉOGRAPHIQUES**, LOGEMENT/MÉNAGE, *PERSONNELLES*)

**Figure 3.** Classes de réponse à l'échelle provinciale, établies pour la non-réponse au cycle 2

## REMERCIEMENTS

Les auteurs tiennent à remercier Douglas Yeo ainsi que les examinateurs pour leurs judicieux commentaires.

## BIBLIOGRAPHIE

ANGOSS SOFTWARE (1995). *Knowledge SEEKER IV for Windows – User's Guide*. ANGOSS Software International Limited.

CATLIN, G., et WILL, P. (1992). Enquête nationale sur la santé de la population: premiers faits saillants. *Rapports sur la santé*, Statistique Canada, 4, 313-319.

CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., et RUBIN, B.R. (1992). Projecting from advance data using propensity modelling: an application to income and tax statistics. *Journal of Business & Economic Statistics*, 10, 117-131.

DAVID, M.H., LITTLE, R., SAMUHEL, M., et TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 168-173.

DUFOUR, J., GAGNON, F., MORIN, Y., RENAUD, M., et SÄRNDAL, C.-E. (1998). Measuring the impact of alternative weighting schemes for longitudinal data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. À paraître.

ERNST, L. (1989). Weighting issues for household and family estimates. Dans *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley and Sons, 135-159.

GRONDIN, C. (1996). Pondération longitudinale – Première vague de l'EDTR. Note interne, division des méthodes d'enquête sociales, Statistique Canada.

INSTITUTE FOR SOCIAL RESEARCH (1979). A Panel Study of Income Dynamics: Procedures and Tape Codes – 1978 Interviewing Year – Wave XI – A Supplement. The University of Michigan.

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

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

LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 137-157.

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

RIZZO, L., KALTON, G., et BRICK, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

ROSENBAUM, P.R., et RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

SKINNER, C.J., et RAO, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 433, 349-356.

STATISTIQUE CANADA (1995). Enquête nationale sur la santé de la population 1994-95 public use microdata file. No. 82F0001XCB au catalogue.

STUKEL, D.M., MOHL, C.A., et TAMBAY, J.L. (1997). Weighting for cycle two of Statistics Canada's National Population Health Survey. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 111-116.

TAMBAY, J.L., et CATLIN, G. (1995). Plan d'échantillonnage de l'Enquête nationale sur la santé de la population. *Rapports sur la santé*, Statistique Canada, 7, 31-42.

THOMSEN, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistik Tidskrift*, 11, 278-283.

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 171-183
Statistique Canada



# Estimations des erreurs de classification dans l'enquête sur la population active et analyse de leur incidence sur les taux de chômage publiés

MICHAEL D. SINCLAIR et JOSEPH L. GASTWIRTH[1]

RÉSUMÉ

Nous examinons dans cet article les erreurs de réponse qui surviennent lors de la Current Population Survey menée par le U.S. Bureau of the Census et évaluons l'incidence de ces erreurs sur les taux de chômage qui sont ensuite publiés par le Bureau of Labor Statistics. Les taux d'erreur sont calculés à partir des données de la réinterview, par une méthode qui se veut un prolongement de celle mise au point par Hui et Walter pour évaluer les tests de diagnostic. Contrairement aux études antérieures qui présumaient que les données conciliées de la réinterview reflétaient la situation véritable, la méthode proposée ici estime les taux d'erreur dans les deux interviews. À partir de ces estimations, nous montrons que les erreurs de classification dans l'enquête initiale ont un effet cyclique sur les estimations du taux de chômage qui sont ensuite publiées et que le degré de sous-estimation augmente tout particulièrement lorsque le taux de chômage réel est élevé. Comme nous n'avions pas suffisamment de données pour établir une distinction entre, d'une part, un modèle présumant que les taux d'erreurs de classification sont les mêmes durant tout le cycle économique et, d'autre part, un modèle qui présume que les taux varient selon que le taux de chômage est faible, modéré ou élevé, nos conclusions doivent être considérées comme provisoires. Celles-ci indiquent néanmoins qu'il y aurait lieu d'examiner plus à fond la relation entre les modèles utilisés pour évaluer l'exactitude des tests de diagnostic et les modèles servant à mesurer les taux d'erreurs de classification dans les données d'enquête.

MOTS CLÉS: Erreurs de classification; taux de chômage; test de diagnostic; conciliation; enquête de réinterview; erreurs de réponse.

## 1. INTRODUCTION

Plusieurs articles, notamment ceux de Poterba et Summers (1986 et 1995) et d'Abowd et Zellner (1985), ont utilisé les données du programme de réinterview mis en place par le U.S. Bureau of the Census, pour estimer les taux d'erreurs de classification dans la Current Population Survey (CPS) et évaluer l'incidence de ces erreurs sur les estimations des changements de situation vis-à-vis de l'activité sur le marché du travail. Ces estimations des taux d'erreurs de classification étaient basées sur l'hypothèse selon laquelle une méthode particulière de réinterview, dite interview de conciliation, donne les «vraies» valeurs. Cependant, Biemer et Forsman (1992), Forsman et Schreiner (1991) et des recherches non publiées menées par le U.S. Bureau of the Census (1963) ont mis en doute cette hypothèse. Le but du présent article est de fournir des estimations des taux d'erreurs de classification dues aux erreurs de réponse lors des interviews et des réinterviews et d'en examiner l'incidence sur les taux de chômage publiés. Contrairement aux études antérieures qui ont porté sur les mouvements bruts, nous nous intéressons ici à l'exactitude des estimations sur la main-d'oeuvre. Notre approche consiste à appliquer à des classifications trinomiales le paradigme proposé par Hui et Walter (1980) pour estimer les taux d'erreur des tests pour diagnostic médical. Un des avantages de cette méthode est qu'il n'est pas nécessaire d'avoir une interview que l'on considère parfaite.

À partir de certaines hypothèses, Hui et Walter (1980) ont mis au point une méthode pour estimer les taux d'erreur associés à un nouveau test de dépistage, en utilisant une épreuve de confirmation à faible taux d'erreur inconnu. Si l'on considère ici que la réinterview équivaut à l'épreuve de confirmation et l'enquête initiale au test de dépistage, cette méthode peut être utilisée pour estimer les taux d'erreur dans l'enquête initiale et au moment de la réinterview, ainsi que les taux de prévalence de la caractéristique à l'étude. La méthode de Hui et Walter (1980) nécessite deux sous-populations dans lesquelles le taux de prévalence de la caractéristique à l'étude diffère. Même si le taux d'erreur n'est pas nécessairement identique dans les deux tests, on présume que les taux d'erreur sont égaux dans les deux sous-populations, pour chaque test. Le modèle (décrit plus en détail en annexe) présume en outre que les erreurs, qui dépendent de la véritable situation d'activité du sujet, sont indépendantes d'un test à un autre.

La méthode de Hui et Walter a été conçue pour des résultats dichotomiques et a été adaptée à l'étude des erreurs de classification des taux d'activité de la population active, par Sinclair et Gastwirth (1996). Nous élargissons ici cette approche pour l'appliquer à trois classifications: chômeurs, population occupée et population inactive, et nous évaluons l'effet des erreurs de classification sur les taux de chômage publiés. Le modèle de base est décrit à la section deux. Les données du programme de réinterview, sur lesquelles s'appuie notre modèle, sont décrites à la

[1] Michael D. Sinclair, Senior Statistician, Mathematical Policy Research, 101 Morgan Lane, Plainsboro, N.J. 08536, U.S.A.; Joseph L. Gastwirth, Professor of Statistics and Economics, George Washington University, 2201 G Street Rm. 315, Washington, D.C. 20052, U.S.A.

section trois. La section quatre présente les taux d'erreur obtenus, ainsi que les taux de chômage «corrigés» en fonction des erreurs de classification estimées. Une mesure de l'exactitude, dite valeur prédictive, qui est utilisée dans les ouvrages sur les tests médicaux de dépistage, est également appliquée au taux de chômage, à la section quatre. Cette mesure indique que la probabilité qu'une personne classée parmi les chômeurs lors de la CPS soit véritablement en chômage varie en fonction du taux réel de chômage.

## 2. LES DONNÉES ET LE MODÈLE

Les données de réinterview de la population active consistent en des réponses trinomiales tirées à la fois de l'enquête initiale et d'une réinterview menée par la suite. Ces données pour une sous-population et une année données sont résumées dans un tableau $3 \times 3$, où la fréquence observée des sujets est représentée par $n_{ygij}$. Dans cette notation:

- $y$ représente l'année;
- $g$ représente la sous-population, $g = 1$ ou $2$;
- $i$ représente la situation d'activité du sujet selon l'enquête initiale: $i = 1$, chômeur; $i = 2$, occupé et $i = 3$, inactif, et
- $j$ représente la situation d'activité du même sujet, selon la réinterview ($j = 1, 2$ et $3$).

Pour chaque situation d'activité, $i = 1, 2$ et $3$, le taux de prévalence vrai, pour la sous-population $g$ et l'année $y$, est représenté par $\pi_{ygi}$. Tout au long du document, nous utiliserons l'expression «taux de prévalence» pour désigner la proportion de personnes dans l'une des trois situations d'activité (par ex., $\pi_{yg1}$). À noter que la fraction $\pi_{yg3}$ dans la catégorie des inactifs est égale à $(1 - \pi_{yg1} - \pi_{yg2})$ et que le taux de chômage réel durant l'année $y$, pour la sous-population $g$, est égal à $\pi_{yg1}$ divisé par $(\pi_{yg1} + \pi_{yg2})$.

Chaque taux de classification, $\beta_{ygrij}$, est défini comme la probabilité que le $r$-ième processus de collecte de données ($r = 1$ pour l'enquête initiale et $r = 2$ pour la réinterview) classe une personne de la sous-population $g$, durant l'année $y$, comme appartenant à la catégorie $i$, $i = 1, 2$ et $3$, lorsque la personne se classe en fait dans la catégorie $j$. Par exemple, $\beta_{11131}$ indique la probabilité que, durant la première année ($y = 1$), une personne de la première sous-population ($g = 1$) soit classée parmi la population inactive ($i = 3$), au moment de l'enquête initiale ($r = 1$), alors que cette personne est en fait en chômage ($j = 1$). On peut répartir les taux de classification en deux groupes, à savoir les taux associés à une classification exacte et ceux associés à une classification erronée. Pour chaque $y$, $g$ et $r$, la probabilité que la méthode d'enquête $r$, durant l'année $y$, classe correctement un chômeur de la sous-population $g$ comme étant en chômage est égale à $\beta_{ygr11} = (1 - \beta_{ygr21} - \beta_{ygr31})$. Les probabilités correspondantes pour les personnes occupées et inactives sont, respectivement, $\beta_{ygr22} = (1 - \beta_{ygr12} - \beta_{ygr32})$ et $\beta_{ygr33} = (1 - \beta_{ygr13} - \beta_{ygr23})$. Si l'on présume d'une indépendance conditionnelle entre les taux de classification selon l'enquête initiale et la réinterview, alors les fréquences observées probables, exprimées par la notation pour chacune des neuf cellules associées à une année $y$ donnée et à la sous-population $g$, sont:

$$E(n_{yg11}) = n_{yg..}(\pi_{yg1}(1-\beta_{yg121}-\beta_{yg131})(1-\beta_{yg221}-\beta_{yg231}) + \pi_{yg2}\beta_{yg112}\beta_{yg212} + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg113}\beta_{yg213})$$

$$E(n_{yg12}) = n_{yg..}(\pi_{yg1}(1-\beta_{yg121}-\beta_{yg131})\beta_{yg221} + \pi_{yg2}\beta_{yg112} * (1-\beta_{yg212}-\beta_{yg232}) + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg113}\beta_{yg223})$$

$$E(n_{yg13}) = n_{yg..}(\pi_{yg1}(1-\beta_{yg121}-\beta_{yg131})\beta_{yg231} + \pi_{yg2}\beta_{yg112}\beta_{yg232} + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg113}(1-\beta_{yg213}-\beta_{yg223}))$$

$$E(n_{yg21}) = n_{yg..}(\pi_{yg1}\beta_{yg121}(1-\beta_{yg221}-\beta_{yg231}) + \pi_{yg2}(1-\beta_{yg112}-\beta_{yg132})\beta_{yg212} + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg123}\beta_{yg213})$$

$$E(n_{yg22}) = n_{yg..}(\pi_{yg1}\beta_{yg121}\beta_{yg221} + \pi_{yg2}(1-\beta_{yg112}-\beta_{yg132}) * (1-\beta_{yg212}-\beta_{yg232}) + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg123}\beta_{yg223})$$

$$E(n_{yg23}) = n_{yg..}(\pi_{yg1}\beta_{yg121}\beta_{yg231} + \pi_{yg2}(1-\beta_{yg112}-\beta_{yg132})\beta_{yg232} + (1-\pi_{yg1}-\pi_{yg2})\beta_{yg123}(1-\beta_{yg213}-\beta_{yg223}))$$

$$E(n_{yg31}) = n_{yg..}(\pi_{yg1}\beta_{yg131}(1-\beta_{yg221}-\beta_{yg231}) + \pi_{yg2}\beta_{yg132}\beta_{yg212} + (1-\pi_{yg1}-\pi_{yg2})(1-\beta_{yg123}-\beta_{yg113})\beta_{yg213})$$

$$E(n_{yg32}) = n_{yg..}(\pi_{yg1}\beta_{yg131}\beta_{yg221} + \pi_{yg2}\beta_{yg132}(1-\beta_{yg212}-\beta_{yg232}) + (1-\pi_{yg1}-\pi_{yg2})(1-\beta_{yg123}-\beta_{yg113})\beta_{yg223})$$

$$E(n_{yg33}) = n_{yg..}(\pi_{yg1}\beta_{yg131}\beta_{yg231} + \pi_{yg2}\beta_{yg132}\beta_{yg232} + (1-\pi_{yg1}-\pi_{yg2})(1-\beta_{yg123}-\beta_{yg113})(1-\beta_{yg213}-\beta_{yg223})),$$

où la taille de l'échantillon pour l'année $y$ et la sous-population $g$ est représentée par $n_{yg..}$.

Le modèle compte 14 paramètres (six taux d'erreur pour l'enquête initiale, $r = 1$, six taux d'erreur pour la réinterview, $r = 2$, et deux taux de prévalence uniques) pour chaque sous-population et année. Cependant, le

tableau $3 \times 3$ pour une année donnée et une sous-population donnée ne compte que 8 fréquences indépendantes, ou degrés de liberté. Le modèle compte donc trop de paramètres et il faut réduire leur nombre; nous utilisons pour ce faire le paradigme de Hui et Walter.

## 3. APPLICATION DU MODÈLE ET DU PROGRAMME DE RÉINTERVIEW DE LA CPS

Le programme de réinterview prévu dans le cadre de la Current Population Survey (CPS) du U.S. Bureau of the Census (U.S. Bureau of the Census 1963) est effectué deux semaines environ après l'enquête initiale, pour mesurer l'erreur de réponse et évaluer la performance de l'intervieweur. Le plan d'échantillonnage pour la réinterview consiste en un échantillon aléatoire auto-pondéré des ménages (Levy et Lemeshow 1980), choisi parmi les ménages affectés à l'intervieweur évalué. La taille de l'échantillon correspond à environ 1/18 de l'échantillon mensuel intégral de la CPS, lequel compte entre 50 000 et 60 000 ménages. Deux méthodes de réinterview sont utilisées. Dans un premier temps, une proportion allant des trois quarts aux quatre cinquièmes de l'échantillon participe à une étude sur le biais dans les réponses, pour laquelle on procède à une première réinterview; une fois cette réinterview terminée, l'intervieweur concilie les divergences entre les réponses obtenues à l'enquête initiale et celles de la première réinterview. Durant cette étude sur le biais dans les réponses, on peut donc obtenir jusqu'à deux réponses de réinterview pour un même sujet, la première étant la réponse non conciliée et la deuxième, la réponse de réinterview conciliée. Dans un deuxième temps, le reste des ménages (soit entre un cinquième et un quart de l'échantillon intégral) passent une réinterview sans qu'il y ait conciliation des résultats.

Pour l'étude sur le biais dans les réponses, l'intervieweur ne doit regarder les réponses de l'enquête initiale qu'après que la réinterview est terminée. Selon Forsman et Schreiner (1991) et Schreiner (1980), l'intervieweur modifierait la réponse obtenue lors de la première réinterview pour qu'elle concorde avec la réponse de l'enquête initiale, car ils ont constaté que l'écart entre les réponses de l'enquête initiale et celles de la première réinterview était plus grand dans l'échantillon non concilié. Sinclair (1994) et Sinclair et Gastwirth (1996) ont ensuite démontré que ces différences étaient statistiquement significatives. Le processus de conciliation crée donc une corrélation entre les réponses initiales et les réponses non conciliées de la réinterview, dans l'échantillon concilié. Aussi, avons-nous décidé de limiter notre analyse aux données de l'enquête initiale et aux données non conciliées de la réinterview. Pour notre étude, nous présumons que, dans l'échantillon non concilié, les erreurs qui proviennent de l'enquête initiale et de la réinterview non conciliée, et qui sont liées à la situation d'activité réelle du répondant, sont indépendantes.

Pour appliquer la méthode de Hui et Walter, il nous faut deux sous-populations dans lesquelles les taux de prévalence diffèrent. Comme il est un fait connu que les taux d'activité des hommes et des femmes diffèrent, nous avons choisi d'utiliser ces deux groupes. Nous devons également présumer que le taux d'erreurs de classification est égal dans les deux sous-populations, hommes et femmes, c'est-à-dire que $\beta_{y1rij} = \beta_{y2rij}$. À ce stade-ci, nous présumons également que les taux d'erreurs de classification lors de l'enquête initiale et de la réinterview non conciliée peuvent être différents et qu'ils peuvent aussi varier d'une année à l'autre. Compte tenu de cette réduction, nous obtenons pour les deux sous-populations, pour une année donnée, 12 paramètres du taux d'erreur et 4 taux de prévalence, pour un total de 16 paramètres. Comme deux tableaux $3 \times 3$ contiennent au total 16 degrés de liberté, il est possible de faire une estimation. Pour l'étude présentée ici, nous avons analysé les données de l'échantillon de réinterview non concilié de la CPS, pour la période allant de 1981 à 1990. Les données annuelles complètes pour 1987 et d'autres données plus récentes n'ont pu être obtenues du U.S. Bureau of the Census.

Le Bureau of Labor Statistics (BLS) publie régulièrement des estimations du taux de chômage, à partir des données de la CPS (voir Bureau of Labor Statistics 1992). Comme la réinterview est un sous-échantillon de l'échantillon intégral utilisé pour la CPS, les estimations du taux de chômage établies à partir de l'échantillon de réinterview différeront des données publiées par le BLS. En outre, des techniques de traitement des données sont appliquées à l'échantillon intégral de la CPS, mais non aux données de la réinterview. À titre d'exemple, l'échantillon complet de la CPS est pondéré en fonction de la probabilité de sélection, et des facteurs de correction pour la non-réponse sont appliqués aux données. En raison de ces différences, les prévalences estimées par notre modèle, à partir uniquement des données de la réinterview, ne peuvent être comparées directement aux données publiées par le BLS. Nous avons utilisé les données de réinterview principalement pour estimer les taux d'erreur dans l'enquête initiale. De plus, tout au long du présent document, nous considérons que les données non conciliées de la réinterview correspondent à un échantillon aléatoire simple de la population aux fins des analyses et des tests d'hypothèses. À partir de ces estimations du taux d'erreur, nous estimons les taux de chômage corrigés du Bureau of Labor Statistics (BLS), le terme «corrigé» signifiant ici que les chiffres publiés ont été modifiés pour tenir compte des erreurs de classification au moment de l'enquête. La formule qui est utilisée pour estimer le taux de chômage vrai, en fonction des prévalences publiées par le BLS d'après l'échantillon complet de la CPS et des taux d'erreurs de classification estimés à partir des données non conciliées de la réinterview, est indiquée en annexe.

## 4. ANALYSE DES DONNÉES ET RÉSULTATS

La première étape dans la préparation de nos estimations finales a été d'obtenir les estimations de paramètres pour chacune des neuf tables de données annuelles, en utilisant la méthode de régression non linéaire SAS avec la méthode des moindres carrés généralisés de Gauss-Newton. Comme les méthodes de réinterview sont demeurées inchangées durant cette période, nous avons décidé de vérifier l'hypothèse selon laquelle les taux d'erreur étaient égaux durant toutes les années à l'étude, c'est-à-dire que $\beta_{ygrij} = \beta_{y'grij}$ pour toutes les années $y \neq y'$. Si on combine maintenant cette hypothèse à l'hypothèse de base voulant que les taux d'erreur pour les hommes et les femmes soient égaux, c'est-à-dire que $\beta_{y1rij} = \beta_{y2rij}$, alors cela suppose que $\beta_{ygrij} = \beta_{y'g'rij}$ pour toutes les valeurs de $y \neq y'$ et $g \neq g'$.

À partir des deux séries de résultats, nous avons effectué un test du rapport des vraisemblances, en présumant que l'échantillon de la réinterview est un échantillon aléatoire simple de la population, utilisé pour vérifier l'hypothèse voulant que chaque taux d'erreur est égal pour toutes les années. La valeur du ratio de vraisemblance, $-2 \log \lambda$ avec 96 degrés de liberté (144 paramètres dans le modèle intégral, moins 48 paramètres dans le modèle réduit), est de 84,06 pour une valeur de $p$ de 0,8027. Les données sont donc cohérentes avec le modèle réduit, ce qui nous permet d'utiliser les estimations du modèle réduit et de simplifier la notation. Aussi, à partir de maintenant, nous utiliserons $\beta_{rij}$ pour représenter $\beta_{ygrij}$, pour toutes les valeurs de $g$ et $y$.

Les estimations du taux d'erreur pour l'enquête initiale et la réinterview non conciliée sont présentées respectivement aux tableaux 1 et 2, avec leurs erreurs-types estimées. Nous remarquons que les taux d'erreur estimés pour la réinterview (tableau 2) sont similaires aux taux d'erreur correspondants de l'enquête initiale. Cette similitude indique que la réinterview non conciliée, qui est menée par le U.S. Bureau of the Census, est une répétition efficace. Les estimations du taux d'erreur montrent en outre que les techniques d'enquête utilisées pour la CPS permettent de classer de façon assez exacte les personnes occupées et inactives. En revanche, ces techniques ne donnent pas d'aussi bons résultats pour ce qui est du classement des personnes en chômage, car la proportion des chômeurs qui sont en fait classés parmi les chômeurs, $(1 - \beta_{121} - \beta_{131})$, n'est que de 0,8397.

Pour fins de comparaison, nous avons fait une analyse des données conciliées de la réinterview – lesquelles correspondent à 75 % de l'échantillon intégral – et ce pour la même période (1981 à 1990), en présumant que les réponses conciliées étaient exemptes d'erreur. Nous avons là aussi créé un tableau $3 \times 3$, en y indiquant le nombre de personnes classées dans chacune des trois catégories d'activité au moment de l'enquête initiale, en fonction du nombre de personnes classées dans chaque catégorie, selon la réinterview conciliée. Ces données sont indiquées au tableau 3.

Le tableau présente les données agrégées par année et par sexe, de sorte que les taux d'erreur calculés à partir de ce tableau peuvent être comparés à notre modèle. En utilisant les données figurant à la colonne «situation véritable», nous calculons une estimation du taux d'erreur. À titre d'exemple, l'estimation de $\beta_{121}$, qui représente la probabilité qu'une personne en chômage soit classée parmi les personnes actives lors de l'enquête initiale, est de 332/17 681 = 0,0188. Ces taux d'erreur sont présentés au tableau 1, pour illustrer dans quelle mesure les taux d'erreur estimés par notre méthode basée sur les données non conciliées diffèrent de ceux obtenus en présumant que la réinterview conciliée donnent des résultats parfaits.

Le tableau 1 présente également les estimations des taux d'erreur pour l'enquête initiale, calculés par Poterba et Summers (1986) à partir des données de la réinterview (combinées pour les deux sexes), pour la première moitié de 1981. Poterba et Summers (1986) s'appuient sur des données conciliées et non conciliées pour estimer les taux d'erreur. Ces auteurs présument que, dans l'échantillon concilié, les intervieweurs utilisent les données de l'enquête initiale pour influencer la réponse au moment de la première réinterview. Ils supposent donc qu'il y a production d'une valeur conciliée uniquement pour la proportion des répondants pour qui il devrait y avoir divergence entre les résultats de l'enquête initiale et ceux de la première réinterview. Par ailleurs, lorsqu'ils obtiennent une valeur conciliée, Poterba et Summer (1986) présument que celle-ci est exempte d'erreur. À partir de ces hypothèses ils utilisent l'échantillon non concilié pour estimer la fréquence de l'erreur et les données conciliées pour fournir de l'information sur la véritable situation d'activité. En résumé, la méthode de Poterba et Summers (1986) et la réinterview conciliée estiment toutes deux que les données conciliées de la réinterview sont parfaites.

Le tableau 4 présente les taux de chômage annuels publiés par le BLS, hommes et femmes combinés, et les compare aux taux de chômage corrigés d'après: (1) nos estimations du taux d'erreur, (2) les taux d'erreur calculés par Poterba et Summers (1986) et (3) les taux d'erreur obtenus en présumant que les données conciliées de la réinterview sont parfaites. Lorsqu'on répartit les résultats du tableau 4 en fonction de la valeur du taux de chômage publié par le BLS, il s'en dégage une tendance évidente, quant au biais dans les estimations initiales de la CPS. La figure 1 montre ainsi que les chiffres publiés ont tendance à surestimer le taux de chômage réel durant les années où le chômage est faible (1989, 1988 et 1990) et, inversement, à le sous-estimer durant les années où le chômage est élevé (1982-1983). On remarque en outre qu'un biais vers le haut est associé à notre méthode, comparativement aux deux autres. Enfin, les trois méthodes indiquent un effet cyclique, qui est le plus faible lorsqu'on présume que la réinterview conciliée est parfaite.

**Tableau 1**
Taux d'erreur estimés dans les données de la CPS initiale

| Paramètre du taux d'erreur | Description | | Valeur estimée $\beta_{1ij}$ | | | Erreur-type estimée |
|---|---|---|---|---|---|---|
| | Classé parmi | Situation véritable | Notre méthode | P et S (1986) | Réint. conc. parfaite | Notre méthode |
| $\beta_{121}$ | Occupés | Chômeurs | 0,0407 | 0,0378 | 0,0188 | 0,01892 |
| $\beta_{131}$ | Inactifs | Chômeurs | 0,1196 | 0,1146 | 0,0838 | 0,01463 |
| $\beta_{112}$ | Chômeurs | Occupés | 0,0049 | 0,0054 | 0,0017 | 0,00124 |
| $\beta_{132}$ | Inactifs | Occupés | 0,0100 | 0,0172 | 0,0098 | 0,00154 |
| $\beta_{113}$ | Chômeurs | Inactifs | 0,0110 | 0,0064 | 0,0034 | 0,00155 |
| $\beta_{123}$ | Occupés | Inactifs | 0,0205 | 0,0116 | 0,0053 | 0,00247 |

**Tableau 2**
Taux d'erreur estimés à partir de la réinterview non conciliée de la CPS

| Paramètre du taux d'erreur | Description | | Valeur estimée Notre méthode $\beta_{2ij}$ | Erreur-type estimée |
|---|---|---|---|---|
| | Classé parmi | Situation véritable | | |
| $\beta_{121}$ | Occupés | Chômeurs | 0,0333 | 0,01772 |
| $\beta_{131}$ | Inactifs | Chômeurs | 0,1128 | 0,01360 |
| $\beta_{112}$ | Chômeurs | Occupés | 0,0057 | 0,00135 |
| $\beta_{132}$ | Inactifs | Occupés | 0,0145 | 0,00160 |
| $\beta_{113}$ | Chômeurs | Inactifs | 0,0157 | 0,00171 |
| $\beta_{123}$ | Occupés | Inactifs | 0,0248 | 0,00238 |

**Tableau 3**
Totalisations croisées entre les réponses agrégées de l'enquête initiale et celles de la réinterview conciliée, 1981-1990
Données de réinterview conciliées correspondant à 75 % de l'échantillon de reinterview de la CPS

| Résultat de l'enquête | Réinterview conciliée | | | |
|---|---|---|---|---|
| CPS initiale | Chômeurs | Occupés | Inactifs | Total |
| Chômeurs | 15 868 | 372 | 480 | 16 720 |
| Occupés | 332 | 213 987 | 744 | 215 063 |
| Inactifs | 1 481 | 2 123 | 138 077 | 141 681 |
| Total | 17 681 | 215 482 | 139 301 | 373 464 |

**Tableau 4**
Incidence des estimations des taux d'erreur

| Année $y$ | Taux de chômage publié par le BLS $UE_y^{BLS}$ | Prob. en chômage, par rapport au nombre classés comme chômeurs | Estimation corrigée du taux de chômage publié par le BLS $AUE_y^{BLS}$ | | | Différence entre taux publié et corrigé | Erreur-type estimée de la différence |
|---|---|---|---|---|---|---|---|
| | | | Notre méthode | Poterba et Summers (1986) | Données conciliées (1981-1990) parfaites | Notre méthode | Notre méthode |
| 1990 | 5,44 % | 0,8135 | 5,27 % | 5,36 % | 5,63 % | 0,17 % | 0,27 % |
| 1989 | 5,20 % | 0,8052 | 4,99 % | 5,09 % | 5,37 % | 0,21 % | 0,26 % |
| 1988 | 5,43 % | 0,8113 | 5,25 % | 5,35 % | 5,62 % | 0,18 % | 0,27 % |
| 1986 | 6,89 % | 0,8503 | 6,97 % | 7,04 % | 7,22 % | -0,08 % | 0,33 % |
| 1985 | 7,09 % | 0,8531 | 7,20 % | 7,27 % | 7,44 % | -0,11 % | 0,34 % |
| 1984 | 7,41 % | 0,8581 | 7,56 % | 7,63 % | 7,79 % | -0,15 % | 0,36 % |
| 1983 | 9,47 % | 0,8894 | 9,99 % | 10,00 % | 10,04 % | -0,52 % | 0,48 % |
| 1982 | 9,54 % | 0,8902 | 10,08 % | 10,09 % | 10,12 % | -0,54 % | 0,49 % |
| 1981 | 7,50 % | 0,8581 | 7,66 % | 7,72 % | 7,88 % | -0,16 % | 0,36 % |

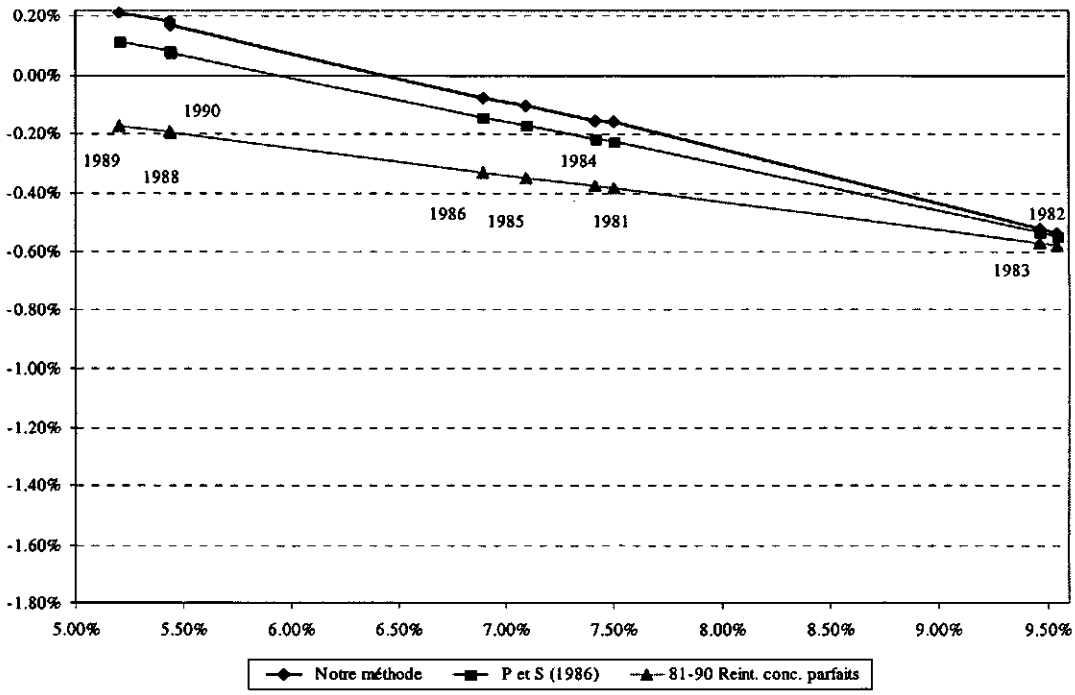



**Figure 1.** Comparaison entre le biais dans les taux de chômage publiés, selon les trois méthodes de calcul

Dans les études sur les tests de dépistage (Gastwirth 1987), la fraction des classifications positives qui sont exactes, appelée valeur prédictive d'un test positif, varie directement en fonction de la prévalence de la caractéristique. C'est ce qui explique que des tests de diagnostic assez précis peuvent donner des taux d'erreur inacceptables lorsqu'ils sont utilisés dans des populations où la prévalence de la maladie à l'étude est faible. L'analogie entre cette mesure et notre étude se situe dans la proportion des personnes qui sont classées parmi les chômeurs et qui sont véritablement en chômage. Cette proportion est indiquée à la troisième colonne, du tableau 4. Bien que l'intervalle des taux de chômage publiés soit assez restreint, on peut néanmoins voir une relation similaire avec le taux de chômage.

Par ailleurs, même si les résultats du test du rapport des vraisemblances indiquaient des taux d'erreur constants durant toute la période, les examinateurs ont proposé de pousser plus loin l'analyse de cette hypothèse. Nous avons donc réparti chacune des neuf années d'enquête en trois groupes, en fonction du taux de chômage publié pour chacune de ces années. Ainsi, les années d'enquête 1990, 1989 et 1988 ont été classées dans le groupe à faible taux de chômage, les taux déclarés pour ces années variant de 5,20 % à 5,44 %. À l'autre extrémité, les années 1982 et 1983 ont été classées dans la catégorie à taux élevé de chômage, avec des taux respectifs de 9,54 % et 9,47 %. Enfin, les autres années ont été classées dans le groupe à taux de chômage modéré, avec des taux variant de 6,89 % à 7,5 %. À partir de cette structure à trois groupes, nous avons conçu un autre modèle, selon lequel les taux d'erreur demeuraient constants à l'intérieur de chacun des groupes, mais pouvaient varier d'un groupe à l'autre. Les taux d'erreur estimés pour l'enquête initiale sont indiqués au tableau 5. On y présente également, pour fins de comparaison, les taux d'erreur indiqués au tableau 1, lesquels étaient basés sur le modèle à taux d'erreur égal.

Nous avons fait un test du rapport des vraisemblances, pour vérifier l'hypothèse selon laquelle le taux d'erreur était égal à l'intérieur de chacun de ces trois groupes, en établissant une comparaison avec le modèle initial sur neuf années. La valeur du ratio de vraisemblance, $-2 \log \lambda$ avec 72 degrés de liberté (144 paramètres dans le modèle complet moins 72 paramètres dans le modèle à trois groupes) a été de 69,25, pour une valeur de $p$ de 0,5697.

Dans l'ensemble, les estimations des taux d'erreur, pour les trois catégories du taux de chômage, semblent similaires. Cependant, comme les erreurs-types des taux d'erreur estimés sont assez élevées, un test formel d'homogénéité n'aurait pas une puissance suffisante pour déceler toute variation dans le taux d'erreur, sur les trois périodes.

Pour évaluer la sensibilité des estimations corrigées du taux de chômage, au tableau 4, nous les avons recalculées en utilisant les taux d'erreur obtenus par le modèle à trois groupes. Ces résultats sont présentés au tableau 6, lequel indique également l'erreur-type des estimations du taux de chômage, qui varie de près de 1,4 % à un sommet d'environ 2,6 %.

**Tableau 5**
Taux d'erreur estimés dans les données de la CPS initiale, pour trois catégories du taux de chômage

| Paramètre du taux d'erreur | Description | | Estimations des taux d'erreur | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Estimations basées sur le modèle à trois groupes | | | | | |
| | Classés parmi | Situation d'activité véritable | Modèle au tableau 1: taux d'erreur constant durant toutes les années | | Années à faible taux 1990,1989 et 1988 | | Années à taux modéré 1981, 1984-1986 | | Années à taux élevé 1982, 1983 | |
| | | | Est. | E-T | Est. | E-T | Est. | E-T | Est. | E-T |
| $\beta_{121}$ | Occupés | Chômeurs | 0,0407 | 0,0189 | 0,0635 | 0,1061 | 0,1113 | 0,1258 | 0,0974 | 0,0717 |
| $\beta_{131}$ | Inactifs | Chômeurs | 0,1196 | 0,0146 | 0,1680 | 0,0538 | 0,1000 | 0,0246 | 0,1084 | 0,0221 |
| $\beta_{112}$ | Chômeurs | Occupés | 0,0049 | 0,0012 | 0,0000 | 0,0047 | 0,0000 | 0,0098 | 0,0000 | 0,0069 |
| $\beta_{132}$ | Inactifs | Occupés | 0,0100 | 0,0015 | 0,0080 | 0,0038 | 0,0096 | 0,0025 | 0,0096 | 0,0031 |
| $\beta_{113}$ | Chômeurs | Inactifs | 0,0110 | 0,0015 | 0,0096 | 0,0040 | 0,0109 | 0,0024 | 0,0103 | 0,0029 |
| $\beta_{123}$ | Occupés | Inactifs | 0,0205 | 0,0025 | 0,0187 | 0,0065 | 0,0202 | 0,0034 | 0,0227 | 0,0044 |

**Tableau 6**
Répercussions des estimations des taux d'erreur, selon le modèle à trois groupes

| Année $y$ | Taux de chômage publié par le BLS | Prob en chômage d'après n$^{bre}$ classé en chômage – Modèle à trois groupes | Estimation corrigée du taux de chômage publié par le BLS | | Différence entre les taux publiés et corrigés | | |
|---|---|---|---|---|---|---|---|
| | | | Modèle initial à taux d'erreur égal | Modèle à trois groupes | Modèle initial à taux d'erreur égal | Modèle à trois groupes | Erreur-type estimée de la différence – méthode à trois groupes |
| 1990 | 5,44 % | 0,9124 | 5,27 % | 6,43 % | 0,17 % | -0,99 % | 1,40 % |
| 1989 | 5,20 % | 0,9088 | 4,99 % | 6,12 % | 0,21 % | -0,93 % | 1,35 % |
| 1988 | 5,43 % | 0,9105 | 5,25 % | 6,41 % | 0,18 % | -0,98 % | 1,41 % |
| 1986 | 6,89 % | 0,9170 | 6,97 % | 8,01 % | -0,08 % | -1,12 % | 2,35 % |
| 1985 | 7,09 % | 0,9178 | 7,20 % | 8,25 % | -0,11 % | -1,16 % | 2,42 % |
| 1984 | 7,41 % | 0,9199 | 7,56 % | 8,64 % | -0,15 % | -1,23 % | 2,53 % |
| 1983 | 9,47 % | 0,9400 | 9,99 % | 11,18 % | -0,52 % | -1,71 % | 2,05 % |
| 1982 | 9,54 % | 0,9404 | 10,08 % | 11,27 % | -0,54 % | -1,73 % | 2,08 % |
| 1981 | 7,50 % | 0,9191 | 7,66 % | 8,74 % | -0,16 % | -1,24 % | 2,56 % |

La figure 2 présente un graphique qui illustre le biais dans les taux de chômage calculés par le modèle à trois groupes et, pour fins de comparaison, par le modèle à taux d'erreur égal. Les résultats à la figure 2 sont très intéressants. Ils montrent ainsi que, même si l'effet cyclique est toujours présent, le biais estimé est ramené vers le bas et présente un biais négatif constant pendant toute la durée du cycle économique.

## 5. INCIDENCE DES ESTIMATIONS CORRIGÉES

Les résultats illustrés aux figures 1 et 2 montrent que toutes les méthodes utilisées pour corriger le taux de chômage en fonction des erreurs de classification indiquent que l'importance du biais dans le taux publié varie tout au long du cycle économique. Compte tenu des différences qui existent entre les estimations du biais obtenues par les deux méthodes, il est difficile de déterminer l'importance du biais. Malheureusement, les estimations sont sensibles à la spécification du modèle, en raison de la petite taille de l'échantillon pour la réinterview non conciliée; ceci se reflète dans les erreurs-types élevées qui sont associées aux estimations du taux d'erreur et, partant, aux estimations du biais.

Notre approche, qui présume que le taux d'erreur demeure constant durant toute la période, laisse croire que le biais dans les estimations de l'enquête est faible durant les années où le taux de chômage se situe entre 5,5 % et 7,5 %. Avec ce modèle, le taux de chômage publié semble sans biais lorsque le taux de chômage vrai se situe autour de 6,3 %, alors qu'il est sous-estimé lorsque le taux réel est supérieur à ce seuil et qu'il est surestimé lorsqu'il est inférieur. Par ailleurs, le biais dû à la sous-estimation devient assez apparent lorsque le taux de chômage atteint 9 %,

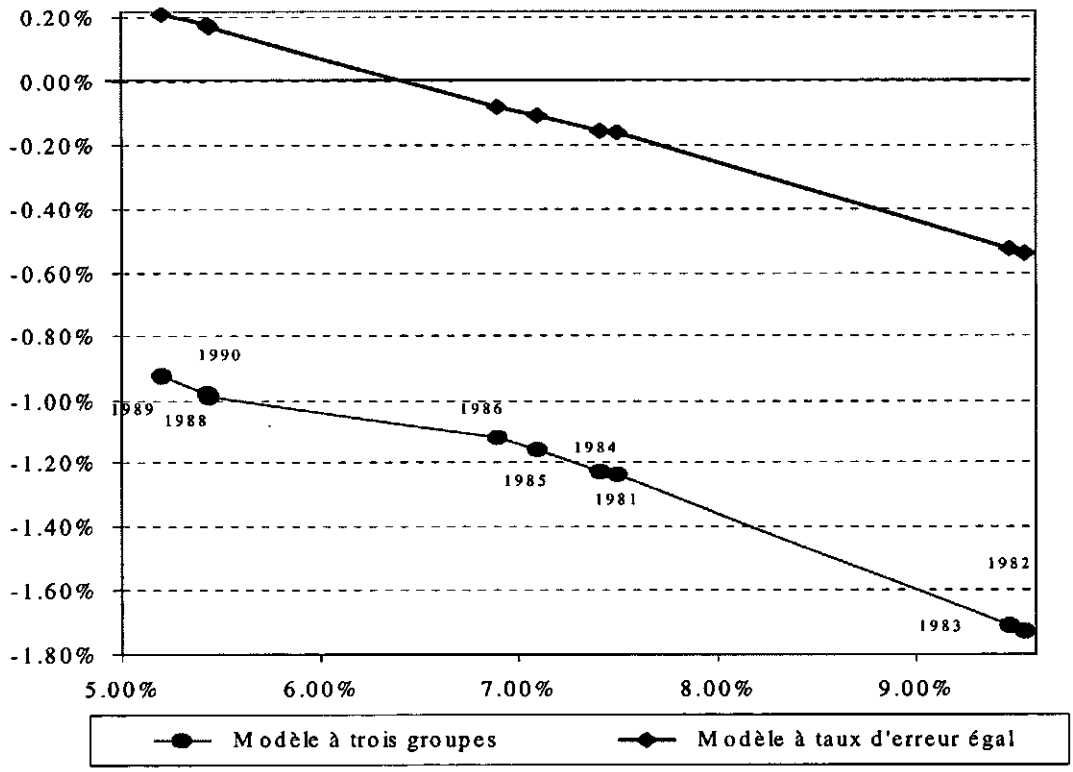



**Figure 2.** Comparaison entre le biais dans les taux de chômage publiés, calculé par le modèle à taux d'erreur égal et le modèle à trois groupes

tandis que le biais de surestimation peut être significatif lorsque le taux de chômage est inférieur à 5 %.

Les résultats du modèle à trois groupes laissent croire que les taux de chômage publiés sont sous-estimés. Si cette conclusion s'avère exacte, alors ces résultats indiquent que le biais durant les années de faible taux de chômage est malgré tout d'environ −0,7 % et qu'il peut atteindre jusqu'à −1,7 % durant les années de chômage élevé. Ces résultats tranchent avec ceux obtenus par le modèle à taux d'erreur égal.

Le fait que l'importance et la direction du biais dans le taux de chômage publié varient durant le cycle économique pourrait influer sur l'utilisation de ce taux dans les études sur le «taux naturel» de chômage et sur la relation entre inflation et chômage. En termes plus précis, nos résultats indiquent que l'intervalle à l'intérieur duquel se situe le taux réel de chômage durant le cycle économique est supérieur à l'intervalle des taux déclarés (voir tableau 4). Hughes et Perlman (1984) ont passé en revue la documentation sur le taux «naturel» de chômage et sur la relation entre inflation et chômage, ainsi que sur le rôle de la théorie de recherche d'emploi pour expliquer pourquoi le chômage n'est pas aussi faible que l'on croit, en période de «plein» emploi. McKenna (1985) présente une étude plus poussée sur la théorie de la recherche d'emploi et sur ses liens avec la durée du chômage et le caractère volontaire du chômage. Déterminer quel modèle est à la base des taux d'erreurs de classification dans la CPS a d'importantes répercussions économiques. En effet, si c'est le modèle à taux d'erreur égal qui est exact, alors le taux publié serait légèrement surestimé en périodes de faible chômage et il y aurait moins de chômage véritable à expliquer par les théories de recherche d'emploi et autres théories connexes. Par contre, si c'est le modèle à trois groupes qui est exact alors, même en périodes de chômage peu élevé, le nombre réel de chômeurs serait en fait plus élevé.

## 6. DISCUSSION

Nous avons présenté dans ce document une autre méthode pour estimer les taux d'erreur dans la CPS. Notre étude diffère des travaux antérieurs, en ce que nous suivons la méthode de Hui et Walter (1980) pour estimer les taux d'erreur, en présumant que le taux d'erreur est le même chez les hommes et les femmes et que les erreurs dans l'enquête initiale sont indépendantes de celles qui surviennent lors de la réinterview non conciliée. Bien qu'il pourrait y avoir une légère corrélation entre les erreurs, l'hypothèse d'indépendance est courante pour une analyse de données de ce type (voir Bailar 1968, Chua et Fuller 1987 et Singh et Rao 1995). On trouve une discussion du biais de la méthode de H et W avec erreurs dépendantes dans Vasek (1995). Quant à l'hypothèse d'un taux d'erreur égal entre les deux sous-populations, plusieurs auteurs cités ici (par ex., Poterba et Summers 1986) ont relevé des

différences de légères à modérées entre les taux d'erreur chez les hommes et les femmes, en s'appuyant sur l'hypothèse de données conciliées parfaites. Il convient cependant de préciser que cette dernière hypothèse est remise en question. Prenons par exemple l'estimation de $\beta_{121}$, c'est-à-dire la probabilité qu'une personne en chômage soit classée parmi les personnes actives lors de l'enquête initiale. À partir du tableau 3, nous estimons cette valeur selon l'hypothèse d'une réinterview conciliée sans biais, en divisant $n_{21}$ par $n_{.1}$ (332/17 681 = 0,0188), où $n_{ij}$ a été défini précédemment et où $j$ correspond maintenant à la situation d'activité selon la réinterview conciliée. En utilisant la valeur probable de ces deux fréquences obtenues à la section 2, nous pouvons représenter comme suit l'expression de la valeur probable de l'estimation pour de gros échantillons:

$$E(n_{21}/n_{.1})$$

$$= \frac{\pi_1\beta_{121}(1-\beta_{221}-\beta_{231})+\pi_2(1-\beta_{112}-\beta_{132})\beta_{212}+(1-\pi_1-\pi_2)\beta_{123}\beta_{213}}{\pi_1(1-\beta_{221}-\beta_{231})+\pi_2\beta_{212}+(1-\pi_1-\pi_2)\beta_{213}}$$

$$= \beta_{121} + \beta_{121}\left[\frac{\pi_1(1-\beta_{221}-\beta_{231})}{\pi_1(1-\beta_{221}-\beta_{231})+\pi_2\beta_{212}+(1-\pi_1-\pi_2)\beta_{213}} - 1\right]$$

$$+\left[\frac{\pi_2(1-\beta_{112}-\beta_{132})\beta_{212}+(1-\pi_1-\pi_2)\beta_{123}\beta_{213}}{\pi_1(1-\beta_{221}-\beta_{231})+\pi_2\beta_{212}+(1-\pi_1-\pi_2)\beta_{213}}\right]. \quad (1)$$

De (1) il s'ensuit que, si le taux d'erreur dans la réinterview conciliée, $\beta_{2ij}$, est égal à zéro, alors cet estimateur est sans biais. Si toutefois la réinterview conciliée ne donne pas des résultats parfaits, le biais dans l'estimateur dépend alors des taux de prévalence dans la population à l'étude. Par conséquent, si les taux d'erreur réels dans l'enquête initiale sont en fait égaux dans les deux sous-populations étudiées et que les classifications obtenues par la conciliation ne sont pas parfaites, alors les taux d'erreur estimés dans l'enquête initiale seront différents pour les deux populations. On ne peut donc pas utiliser les similitudes ou les différences entre les taux d'erreur estimés pour les hommes et les femmes, observées lors d'études antérieures, pour confirmer ou infirmer les hypothèses présentées ici.

Nous avons également fait une analyse de sensibilité de la méthode de Hui et Walter (1986) appliquée à des réponses dichotomiques (Sinclair 1994) et constaté que la méthode est sensible, dans certains cas, à un manquement à l'hypothèse sur les taux d'erreur égaux, mais qu'elle est très robuste dans d'autres circonstances. Il faudra donc poursuivre les recherches afin de mettre au point des techniques de réinterview et d'analyse qui permettent d'assouplir les hypothèses restrictives actuellement requises pour l'analyse des données de réinterview.

Il est à noter que Chua et Fuller (1987) ont eux aussi obtenu des estimations des erreurs de classification à trois résultats, à partir des données de réinterview correspondant à 25 % de l'échantillon de reinterview de la CPS, pour la période de 1977 à 1980. Leurs résultats, qui sont analogues aux nôtres, révèlent que les taux d'erreur les plus élevés sont associés à la classification des personnes véritablement en chômage. Poterba et Summers (1995), de même que Singh et Rao (1995), ont eux aussi constaté que ce groupe était le plus difficile à classer. Comme tous les modèles examinés indiquent que le taux global d'erreurs de classification des personnes en chômage est d'environ 20 %, les futures réinterviews devraient chercher à déterminer pourquoi ces taux sont aussi élevés et ainsi, nous l'espérons, contribuer à améliorer l'enquête.

Les estimations «corrigées», présentées au tableau 4, pourraient servir notamment à une analyse de sensibilité des études (par ex. Abowd et Zellner 1985; Poterba et Summers 1995) sur les mouvements bruts et sur la dynamique du marché du travail qui présumaient que l'interview conciliée était parfaite. Ceci équivaudrait pour eux à utiliser les estimations à l'avant-dernière colonne du tableau 3. De même, les estimations des erreurs de classification pourraient être introduites dans les méthodes utilisées pour estimer les modèles probit et logit avec les variables des réponses mal classées (Hausman et Scott-Morton 1994), ainsi que pour le développement de techniques statistiques formelles pour les données d'enquête (Rao et Thomas 1991). Il convient en terminant d'insister sur le fait que toutes les estimations corrigées en fonction des erreurs de classification n'en sont encore qu'au stade de la recherche et que les estimations des taux d'erreur ne sont pas encore suffisamment exactes pour corriger les données de l'enquête régulière, en particulier du fait qu'un nouveau questionnaire et que de nouvelles techniques d'interview ont été introduites en janvier 1994 (Bureau of Labor Statistics 1993).

## REMERCIEMENTS

Les auteurs aimeraient remercier Irv Schreiner du U.S. Bureau of the Census pour les données qui ont servi à la présente étude et pour ses précieux conseils. Nous voulons également remercier John Thompson, Henry Woltman et Jon Clark, également du U.S. Bureau of the Census, de même que les examinateurs et le rédacteur adjoint, pour leurs nombreux commentaires utiles. Cette recherche a été financée en partie par une subvention de la National Science Foundation.

## ANNEXE TECHNIQUE A

### Examen de la méthode de Hui et Walter

La méthode de Hui et Walter a été conçue pour évaluer les tests de diagnostic. L'avantage de cette technique est qu'elle permet au chercheur de mesurer le taux d'erreur dans une épreuve donnée, sans avoir à utiliser une épreuve

de comparaison exempte d'erreurs. La méthode prévoit l'utilisation de deux populations (ou sous-populations), auxquelles correspondent des taux de prévalence différents, pour estimer les paramètres. Les données d'une telle étude peuvent être résumées dans un tableau $2 \times 2$ comme celui illustré à la figure A qui suit. Ce tableau, lorsqu'il se rapporte à une sous-population précise, est indexé par la lettre $g$. Nous représentons par $n_{ij}$ la fréquence des cas qui, dans la sous-population $g$, sont classés au premier test dans la catégorie $i$ ($i = 1$ pour ceux qui possèdent la caractéristique à l'étude et $i = 2$ pour ceux qui ne l'ont pas) et, au deuxième test, dans la catégorie $j$ ($j = 1$ ou 2). Supposons que $\pi$ représente le taux de prévalence inconnu véritable de la caractéristique à l'étude et que $\alpha_r$ et $\beta_r$ représentent les faux taux positif et négatif inconnus. Ces taux d'erreur sont indexés par la lettre $r$, où $r = 1$ correspond au résultat du premier test et $r = 2$, au résultat du deuxième (dans notre cas, $r = 1$ correspond à l'enquête initiale et $r = 2$ à la réinterview). Le faux taux positif, $\alpha_r$ indique la probabilité que l'évaluation au $r$-ième test classe la personne comme étant positive alors qu'elle devrait en fait être classée parmi les négatifs. De même, le faux taux négatif, $\beta_r$ est la probabilité que l'évaluation au $r$-ième test classe le cas comme négatif alors que la personne présente la caractéristique à l'étude. «Un moins (1 -)» chacun de ces paramètres reflètent respectivement la spécificité et la sensibilité des méthodes de classification du test (ou de l'enquête).

| Résultat du test 1 (Enquête initiale) | Résultat du test 2 (Réinterview) | | |
|---|---|---|---|
| | Positif | Négatif | Total |
| Positif | Cellule 1 | Cellule 3 | $n_{1\cdot}$ |
| Négatif | Cellule 2 | Cellule 4 | $n_{2\cdot}$ |
| Total | $n_{\cdot 1}$ | $n_{\cdot 2}$ | $n_{\cdot\cdot}$ |

**Figure A.** Classification croisée des résultats des tests 1 et 2

Si l'on présume que les erreurs dans le premier et le deuxième tests sont indépendantes l'une de l'autre (compte tenu de la situation vraie), les probabilités prévues, représentées par $P_{ij}$, qui sont associées aux fréquences par cellule illustrées à la figure A pour une sous-population donnée $g$, se définissent comme suit:

Pour

$$\text{Cellule 1 } P_{g11} = \pi_g(1-\beta_{1,g})(1-\beta_{2,g}) + (1-\pi_g)(\alpha_{1,g}\alpha_{2,g})$$

$$\text{Cellule 2 } P_{g21} = \pi_g(\beta_{1,g})(1-\beta_{2,g}) + (1-\pi_g)(1-\alpha_{1,g})(\alpha_{2,g})$$

$$\text{Cellule 3 } P_{g12} = \pi_g(1-\beta_{1,g})\,\beta_{2,g} + (1-\pi_g)(\alpha_{1,g})(1-\alpha_{2,g})$$

$$\text{Cellule 4 } P_{g22} = \pi_g(\beta_{1,g}\,\beta_{2,g}) + (1-\pi_g)(1-\alpha_{1,g})(1-\alpha_{2,g}). \quad \text{(A.1)}$$

À partir de (A.1), nous remarquons que nous avons au total cinq paramètres mais seulement trois entrées indépendantes (ou degrés de liberté) pour faire les estimations. Le nombre de paramètres doit donc être réduit.

Pour réduire les paramètres, Hui et Walter posent d'abord comme hypothèse que la proportion de cas qui présentent la caractéristique diffère d'une sous-population à une autre et donc que $\pi_1 \neq \pi_2$. Deuxièmement, il faut que les taux d'erreur pour chaque test soient identiques dans les deux sous-populations, mais les taux d'erreur associés aux deux tests peuvent différer. Pour les deux sous-populations, ceci suppose que, dans (A.1), $\beta_r = \beta_{r,1} = \beta_{r,2}$ et $\alpha_r = \alpha_{r,1} = \alpha_{r,2}$ et que $\beta_1 \neq \beta_2$, et $\alpha_1 \neq \alpha_2$. Dans de telles conditions, le nombre de paramètres est réduit à six (deux taux de prévalence, un pour chaque sous-population, et deux taux d'erreur pour chaque test, 1 et 2). Comme les deux tableaux $2 \times 2$ contiennent six degrés de liberté, il est possible de faire des estimations. À noter que, si $\pi_1 = \pi_2$, et que les taux d'erreur sont les mêmes dans les deux sous-populations, alors les probabilités dans (A.1) seraient les mêmes pour les deux sous-populations, de sorte que nous n'aurions en fait qu'un tableau et donc qu'il serait impossible de faire des estimations. Selon le modèle de Hui et Walter (1986), les estimations par les moindres carrés non linéaires généralisés peuvent être calculées en utilisant l'algorithme de Gauss Newton tiré de la méthode de régression non linéaire SAS. Avec cette méthode, il est possible d'exprimer les fréquences observées, $n_{ij}$, en fonction de la taille totale de l'échantillon, $n_{\cdot\cdot}$ et de les multiplier par les probabilités dans l'expression (A.1). Hui et Walter (1986) présentent également les estimateurs en forme analytique fermée dans (A.2) en fonction des probabilités observées par cellule, lesquelles sont représentées par $p_{gij}$.

$$\hat{\alpha}_r = \frac{(p_{r1\cdot}p_{\bar{r}\cdot 1} - p_{r\cdot 1}p_{\bar{r}1\cdot} + p_{211} - p_{111} + D)}{2E_r}$$

$$\hat{\beta}_r = \frac{(p_{r\cdot 2}p_{\bar{r}2\cdot} - p_{r2\cdot}p_{\bar{r}\cdot 2} + p_{122} - p_{222} + D)}{2E_r} \quad \text{(A.2)}$$

où,

$\bar{r} = 2$ if $r = 1$, $\bar{r} = 1$ if $r = 2$

$$p_{g\cdot j} = \sum_{i=1}^{2} p_{gij},\ p_{gi\cdot} = \sum_{j=1}^{2} p_{gij};$$

$$\hat{\pi}_g = \frac{1}{2} + \frac{[p_{g1\cdot}(p_{1\cdot 1} - p_{2\cdot 1}) + p_{g\cdot 1}(p_{1\cdot 1} - p_{21\cdot}) + p_{211} - p_{111}]}{2D}$$

où,

$$D = \pm\,[(p_{11\cdot}p_{21\cdot} - p_{1\cdot 1}p_{111} + p_{111} - p_{211})^2 - 4(p_{11\cdot} - p_{21\cdot})(p_{111}p_{2\cdot 1} - p_{211}p_{1\cdot 1})]^{\frac{1}{2}}$$

avec,

$$E_1 = p_{2\cdot 1} - p_{1\cdot 1},\ E_2 = p_{21\cdot} - p_{11\cdot}.$$

À noter qu'il existe deux points distincts dans l'ensemble de solutions, pour une valeur positive ou négative de $D$; cependant, une seule de ces valeurs donnera des estimations raisonnables. Les variances des estimateurs, calculées à partir de la matrice d'information asymptotique estimée, sont définies dans le document de Hui et Walter (1980).

## ANNEXE TECHNIQUE B

### Correction des taux de chômage publiés

Pour évaluer les répercussions des taux d'erreur estimés, nous avons besoin d'une expression pour estimer les taux de prévalence réels (les quatre paramètres $\pi$), en fonction des taux d'erreur estimés et des taux de prévalence observés (ou fréquences dans l'échantillon), pour une enquête donnée. Nous présentons ici la formule utilisée pour ces calculs, laquelle peut également servir à calculer les taux corrigés du BLS en utilisant, comme valeurs observées, les taux de chômage et les taux d'emploi publiés par le BLS. Cette expression est définie en (B.1).

À noter que, dans l'expression (B.1), nous avons supprimé l'indice $g$-ième des paramètres $\pi$, afin que l'expression représente les taux de prévalence parmi la population en général, hommes et femmes réunis. Pour cette étude, nous présumons que les taux d'erreur pour les hommes et les femmes sont égaux.

$$\begin{bmatrix}\hat{\pi}_{y1}\\ \hat{\pi}_{y2}\end{bmatrix} = \begin{bmatrix}1-\hat{\beta}_{121}-\hat{\beta}_{131}-\hat{\beta}_{113} & \hat{\beta}_{112}-\hat{\beta}_{113}\\ \hat{\beta}_{121}-\hat{\beta}_{123} & 1-\hat{\beta}_{112}-\hat{\beta}_{132}-\hat{\beta}_{123}\end{bmatrix}^{-1} \begin{bmatrix}\frac{n_{y1\cdot}}{n_{y\cdot\cdot}} - \hat{\beta}_{113}\\ \frac{n_{y2\cdot}}{n_{y\cdot\cdot}} - \hat{\beta}_{123}\end{bmatrix}. \quad \text{(B.1)}$$

Dans l'analyse présentée ici, nous avons trois ensembles de valeurs observées, soit: deux taux de prévalence observés – établis à partir de l'échantillon de réinterview (lequel est un sous-échantillon de l'échantillon intégral de la CPS) et incluant les données non conciliées de la réinterview et les données conciliées de l'étude sur le biais dans les réponses – et les taux de prévalence publiés par le BLS à partir de l'enquête initiale intégrale (CPS). Nous nous concentrerons ici sur les premier et dernier ensembles de statistiques, à savoir les données non conciliées de la réinterview et les estimations publiées du BLS. Afin de maintenir une distinction entre ces deux ensembles, nous définissons,

$$U_y^R = \frac{n_{y1\cdot}}{n_{y\cdot\cdot}}$$

$$E_y^R = \frac{n_{y2\cdot}}{n_{y\cdot\cdot}} \quad \text{(B.2)}$$

comme étant les taux de prévalence observés de chômeurs et de personnes occupées, d'après les données non conciliées de la réinterview. Les taux de prévalence correspondants, publiés par le BLS et basés sur les données pondérées de l'enquête initiale intégrale (CPS), sont définis par $U_y^{BLS}$ et $E_y^{BLS}$.

De même, le taux de chômage observé dans la population active, d'après l'échantillon de réinterview non conciliée, est représenté par $UE_y^R$, qui est égal à $U_y^R$ divisé par $(U_y^R + E_y^R)$, alors que le taux de chômage publié par le BLS est représenté par $UE_y^{BLS}$.

Si on simplifie l'expression (B.1) en fonction des taux de prévalence observés lors de la réinterview, $U_y^R$ et $E_y^R$, nous obtenons:

$$\begin{aligned}\hat{\pi}_{y1} = &\{U_y^R - \hat{\beta}_{113} - \hat{\beta}_{112}U_y^R + \hat{\beta}_{113}\hat{\beta}_{112} - \hat{\beta}_{113}\hat{\beta}_{132}\\ &- \hat{\beta}_{123}U_y^R - \hat{\beta}_{112}E_y^R + \hat{\beta}_{123}\hat{\beta}_{112} + \hat{\beta}_{113}E_y^R\}\\ &\{1 - \hat{\beta}_{112} - \hat{\beta}_{132} - \hat{\beta}_{123} - \hat{\beta}_{121}(1 + \hat{\beta}_{132} + \hat{\beta}_{123} + \hat{\beta}_{113})\\ &- \hat{\beta}_{131}(\hat{\beta}_{112} + \hat{\beta}_{132} - 1) - \hat{\beta}_{113}(\hat{\beta}_{112} + \hat{\beta}_{132} - 1) + \hat{\beta}_{123}\hat{\beta}_{112}\}\end{aligned}$$

$$\begin{aligned}\hat{\pi}_{y2} = &\{-\hat{\beta}_{121}U_y^R + \hat{\beta}_{121}\hat{\beta}_{113} + \hat{\beta}_{123}U_y^R + E_y^R - \hat{\beta}_{123} - \hat{\beta}_{121}E_y^R\\ &+ \hat{\beta}_{122}\hat{\beta}_{123} - \hat{\beta}_{131}E_y^R + \hat{\beta}_{131}\hat{\beta}_{123} - \hat{\beta}_{123}E_y^R\}\\ &\{1 - \hat{\beta}_{112} - \hat{\beta}_{132} - \hat{\beta}_{123} - \hat{\beta}_{121}(1 + \hat{\beta}_{132} + \hat{\beta}_{123} + \hat{\beta}_{113})\\ &- \hat{\beta}_{131}(\hat{\beta}_{112} + \hat{\beta}_{132} - 1) - \hat{\beta}_{113}(\hat{\beta}_{112} + \hat{\beta}_{132} - 1) + \hat{\beta}_{123}\hat{\beta}_{112}\}. \quad \text{(B.3)}\end{aligned}$$

À partir de l'expression (B.3), nous pouvons estimer le taux de chômage corrigé à partir de la réinterview, représenté par $AUE_{y_R}^R$ égal à $\hat{\pi}_{yg1}$ divisé par $(\hat{\pi}_{yg1} + \hat{\pi}_{yg2})$. À noter que $AUE_y^R$ peut être représenté par:

$$\begin{aligned}A\hat{U}E_y^R = &\{-U_y^R + E_y^R + \hat{\beta}_{112}(U_y^R - \hat{\beta}_{113} + E_y^R)\\ &+ \hat{\beta}_{132}(U_y^R - \hat{\beta}_{113}) + \hat{\beta}_{123}(U_y^R - \hat{\beta}_{112}) - \hat{\beta}_{113}E_y^R\}\\ &\{U_y^R + \hat{\beta}_{113}(1 + \hat{\beta}_{112} - \hat{\beta}_{121} - \hat{\beta}_{123}) + \hat{\beta}_{112}(U_y^R + E_y^R - \hat{\beta}_{113})\\ &+ \hat{\beta}_{121}(U_y^R + E_y^R - \hat{\beta}_{123}) - E_y^R + \hat{\beta}_{123} + \hat{\beta}_{131}(U_y^R - \hat{\beta}_{123})\} \quad \text{(B.4)}\end{aligned}$$

Enfin, pour obtenir l'estimation corrigée du taux de chômage publié par le BLS, laquelle est représentée par $AUE_y^{BLS}$, nous remplaçons $U_y^R$ par $U_y^{BLS}$ et $E_y^R$ par $E_y^{BLS}$ dans l'expression (B.4). À noter que les erreurs-types estimées des estimations pour $AUE_y^{BLS}$ qui ont été présentées à la section quatre, ont été calculées par une méthode d'approximation par série de Taylor (Wolter 1985). Nous avons d'abord présumé, pour ce faire, que la variance dans les estimations publiées de $U_y^{BLS}$ et $E_y^{BLS}$

était négligeable. Même si ce n'est pas le cas, cette hypothèse simplifie grandement le calcul de la variance et saisit la majeure partie de la variation totale. Par ailleurs, cette hypothèse est corroborée par le fait que, compte tenu de la grande taille de l'échantillon annuel intégral qui est utilisé pour la CPS, la variance dans ces échantillons est négligeable lorsqu'on la compare à l'erreur d'échantillonnage associée aux estimations des taux d'erreur, lesquelles sont basées sur le petit échantillon de la réinterview non conciliée. En résumé, lorsque la substitution de $U_y^R$ par $U_y^{BLS}$ et de $E_y^R$ par $E_y^{BLS}$ dans l'expression (B.4) est complétée, nous présumons que $U_y^{BLS}$ et $E_y^{BLS}$ sont des valeurs fixes connues dans cette équation. Enfin, la variance d'échantillonnage, qui est associée à la différence entre la valeur corrigée et la valeur publiée et qui définit le biais dans l'estimation initiale, est calculée en faisant la somme des variances. Par conséquent, si nous présumons que la valeur publiée est exempte de variance d'échantillonnage, la variabilité d'échantillonnage associée à la différence ou au biais est tout simplement égale à la variabilité d'échantillonnage de la valeur corrigée.

## ANNEXE TECHNIQUE C

### Estimation des erreurs-types des taux de chômage corrigés

Pour une fonction complexe de plusieurs paramètres estimés, les estimations de la variance associée à cette fonction peuvent être calculées par approximation par série de Taylor, comme l'a proposé Wolter (1985). Supposons que le paramètre à étudier est $Y = G(\Theta)$, où $\Theta$ représente un $n$ vecteur dimensionnel des paramètres de la population, $\Theta = \{\theta_1, ..., \theta_n\}$. Si $G$ possède des dérivées secondes continues qui se situent dans un intervalle acceptable pour $\Theta$ et $\Theta$-accent, alors Wolter présente la relation:

$$\hat{Y} - Y = A + R(\hat{\Theta},\Theta)$$

où,

$$A = \sum_{k=1}^{n} \frac{\partial G(\ominus)}{\partial \theta_k}(\hat{\theta}_k - \theta_k)$$

$$R(\hat{\ominus},\ominus) = \sum_{k=1}^{n}\sum_{i=1}^{n} (1/2!)\frac{\partial^2 G(\Lambda)}{\partial\theta_k \partial\theta_i}(\hat{\theta}_k - \theta_k)(\hat{\theta}_i - \theta_i)$$

$$\hat{\ominus} \le \Lambda \le \ominus. \qquad \text{(C.1)}$$

Le terme qui reste est souvent considéré comme ayant peu de conséquence et est éliminé de la relation. Avec une approximation du premier ordre, Wolter présente,

$$\begin{aligned} \mathrm{MSE}(\hat{Y}) &= E[G(\hat{\ominus}) - G(\ominus)]^2 \\ &= \mathrm{Var}(A) \\ &= \sum_{k=1}^{n}\sum_{i=1}^{n} \frac{\partial G(\ominus)}{\partial\theta_k}\frac{\partial G(\ominus)}{\partial\theta_i}\,\mathrm{Cov}(\hat{\theta}_k,\hat{\theta}_i) \\ &= d\,\Sigma_{\hat{\theta}}\,d^T \end{aligned} \qquad \text{(C.2)}$$

où $d$ est un vecteur ligne de dimension $n$, formé des éléments

$$d_k = \left[\frac{\partial G(\ominus)}{\partial\theta_k}\right]. \qquad \text{(C.3)}$$

Wolter désigne cet estimateur l'approximation du premier ordre de l'erreur quadratique moyenne (laquelle est égale à la somme de la variance d'échantillonnage et du biais de l'estimateur au carré). On peut obtenir des approximations d'ordre supérieur, en conservant des termes additionnels dans le développement. Pour l'estimation de la variance, nous remplaçons $\sum_\theta$ par la matrice de covariance estimée et nous évaluons $d$ d'après les valeurs estimées de $\theta$. Dans le cas particulier qui nous intéresse, nous voulons estimer la variance associée à la fonction des estimations dans l'expression (C.4) indiquée ci-après.

$$\begin{aligned} G(\ominus) = G(&\hat{\beta}_{121},\hat{\beta}_{131},\hat{\beta}_{112},\hat{\beta}_{132},\hat{\beta}_{113},\hat{\beta}_{123},U_y^{BLS},E_y^{BLS}) = \\ &\Big\{-U_y^{BLS}+E_y^{BLS}+\hat{\beta}_{112}(U_y^{BLS}-\hat{\beta}_{113}+E_y^{BLS}) \\ &\quad +\hat{\beta}_{132}(U_y^{BLS}-\hat{\beta}_{113})+\hat{\beta}_{123}(U_y^{BLS}-\hat{\beta}_{112})-\hat{\beta}_{113}E_y^{BLS}\Big\} \\ &\Big\{U_y^{BLS}+\hat{\beta}_{113}(1+\hat{\beta}_{112}-\hat{\beta}_{121}-\hat{\beta}_{123})+\hat{\beta}_{112}(U_y^{BLS}+E_y^{BLS}-\hat{\beta}_{113}) \\ &\quad +\hat{\beta}_{121}(U_y^{BLS}+E_y^{BLS}-\hat{\beta}_{123})-E_y^{BLS}+\hat{\beta}_{123}+\hat{\beta}_{131}(U_y^{BLS}-\hat{\beta}_{123})\Big\}. \end{aligned} \qquad \text{(C.4)}$$

Pour obtenir les estimations, nous présumons que les valeurs de $U_y^{BLS}$ et $E_y^{BLS}$ sont fixes (c'est-à-dire que la variance d'échantillonnage est négligeable). Si nous prenons les dérivées partielles de l'équation (C.4) pour les six taux d'erreur et que nous évaluons ces expressions en fonction des taux d'erreur estimés, nous obtenons un vecteur $d$ qui dépend des estimations des taux d'erreur ainsi que des proportions de chômeurs et de personnes occupées publiées par le BLS pour chaque année à l'étude. Dans notre modèle initial qui présume que les taux d'erreur demeurent fixes d'une année à l'autre, ce vecteur $d$ pour la période à l'étude ne varie, d'une année à l'autre, que pour les valeurs publiées. À titre d'exemple, pour l'année 1989, le vecteur $d$ estimé à partir des taux de chômage et d'emploi publiés par le BLS (soit 0,0347 et 0,6329), est égal à:

$$\hat{d} = \begin{bmatrix} \hat{\beta}_{121} & 0{,}07851 \\ \hat{\beta}_{131} & 0{,}07558 \\ \hat{\beta}_{112} & -1{,}2918 \\ \hat{\beta}_{132} & -0{,}04813 \\ \hat{\beta}_{113} & -0{,}64214 \\ \hat{\beta}_{123} & 0{,}03884 \end{bmatrix}.$$

La matrice de covariance estimée, par notre analyse de régression non linéaire SAS, laquelle est basée sur le modèle initial qui présume que le taux d'erreur est fixe d'année en année et donc est le même pour toutes les années à l'étude, est définie ci-après.

| $\Sigma$ | β121 | β131 | β112 | β132 | β113 | β123 |
|---|---|---|---|---|---|---|
| $\beta_{121}$ | 0,000358 | -4,7E-05 | -3,5E-07 | -2,6E-08 | -3,9E-07 | 2,9E-07 |
| $\beta_{131}$ | -4,7E-05 | 0,000214 | -1,7E-07 | -5,2E-07 | -1,4E-06 | -2,8E-07 |
| $\beta_{112}$ | -3,5E-07 | -1,7E-07 | 1,54E-06 | 2,14E-07 | -2,3E-08 | 9,9E-10 |
| $\beta_{132}$ | -2,6E-08 | -5,2E-07 | 2,14E-07 | 2,37E-06 | -1,5E-08 | -6,1E-08 |
| $\beta_{113}$ | -3,9E-07 | -1,4E-06 | -2,3E-08 | -1,5E-08 | 2,4E-06 | -8E-08 |
| $\beta_{123}$ | 2,9E-07 | -2,8E-07 | 9,9E-10 | -6,1E-08 | -8,0E-08 | 6,1E-06 |

La multiplication a priori et a posteriori du vecteur $d$ par la matrice de covariance estimée donne une variance estimée pour $AUE^{BLS}$ qui est de 6,72 E-6 pour 1989 et une erreur-type de 0,0026 (0,26 %), comme l'indique le tableau 4.

## BIBLIOGRAPHIE


ABOWD, J., et ZELLNER, A. (1985). Estimated gross labor force flows. *Journal of Economic and Business Statistics*, 3, 253-283.

BAILAR, B.A. (1968). Recent research in reinterview procedures. *Journal of the American Statistical Association*, 63, 41-63.

BIEMER, P.P., et FORSMAN, G. (1992). On the quality of reinterview data with application to the current population survey. *Journal of the American Statistical Association*, 87, 915-923.

BUREAU OF LABOR STATISTICS (1993). Overhauling the population survey. *Monthly Labor Review*, 116, 9. Washington DC: U.S. Government Printing Office.

BUREAU OF LABOR STATISTICS (1992). *Employment and Earnings*. 38, 8. Washington DC: U.S. Government Printing Office.

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.

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

GASTWIRTH, J.L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDs antibodies test data. *Statistical Science*, 2, 213-238.

HAUSMAN, J.A., et MORTON, S. (1994). Misclassification of a Dependent Variables in a Discrete Response Setting. Document de travail, Department of Economics, MIT, Cambridge.

HUI, S.L., et WALTER, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167-171.

HUGHES, J.J., et PERLMAN, R. (1984). *The Economics of Unemployment*. New York: Cambridge University Press.

LEVY, P., et LEMESHOW, S. (1980). *Sampling for Health Professionals*. California: Lifetime Learning Publications.

McKENNA, C.J. (1985). *Uncertainty and the Labor Market: Recent Developments in Job Search Theory*. New York: St. Martins Press.

POTERBA, J.M., et SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.

POTERBA, J.M., et SUMMERS, L.H. (1995). Unemployment benefits and labor market transitions: a multinominal logit model with errors in classification. *The Review of Economics and Statistics*, 77, 207-216.

RAO, J.N.K., et THOMAS, D.R. (1991). Chi-squared tests with complex survey data subject to misclassification error. Dans *Measurement Errors in Surveys*, (Éds. Paul Biemer et coll.). New York: Wiley.

SCHREINER, I., (1980). Reinterview Results From the CPS Independent Reconciliation Experiment (second quarter 1978 through third quarter 1979). Note de service non publiée, U.S. Bureau of the Census, Mai 7, 1980.

SINCLAIR, M.D. (1994). Evaluating Reinterview Survey Methods for Measuring Response Errors. Thèse de doctorat, George Washington University, Septembre.

SINCLAIR M.D., et GASTWIRTH, J.L. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Association*, 91, 961-969.

SINGH, A.C., et RAO, J.N.K (1995). On the adjustment of gross flow estimate for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90, 478-488.

U.S. BUREAU OF THE CENSUS (1963). *The Current Population Survey Reinterview Program: Some Notes and Discussion*. Document Technique no. 6. Washington, D.C.: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1963). *Evaluating Censuses of Population and Housing*. Statistical Training Document #ISP-TR-5. Washington, D.C.: U.S. Government Printing Office, 1985.

VACEK, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41, 959-968.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 185-199
Statistique Canada



# Utilisation de méthodes d'appariement statistique dans l'estimation de calage

ROBBERT H. RENSSEN[1]

RÉSUMÉ

Le présent article porte sur un essai de mise en tableau croisé de deux variables nominales qui ont été recueillies de manière distincte à partir de deux échantillons indépendants de grande taille, et recueillies conjointement à partir d'un seul échantillon de petite taille. Dans le cadre de cet essai, on a présumé que les échantillons de grande taille présentent un grand ensemble de variables communes. La méthode d'estimation qui est proposée peut être considérée comme un mélange entre les méthodes de calage et l'appariement statistique. Grâce aux méthodes de calage, il est possible d'intégrer les plans d'échantillonnage complexes à la méthode d'estimation, afin de répondre à certaines exigences en matière de cohérence entre des estimations provenant de sources différentes, ainsi que pour obtenir des estimations plutôt non biaisées dans le cas du tableau à double entrée. Grâce aux méthodes d'appariement statistique, il est possible d'intégrer un ensemble relativement important de variables communes à l'estimation de calage, à l'aide de laquelle on peut améliorer la justesse du tableau à double entrée qui est estimé. La méthode d'estimation nous permet de mieux comprendre le biais qui accompagne généralement l'estimation du tableau à double entrée lorsqu'on utilise uniquement les échantillons de grande taille. Nous montrons l'utilité de la méthode d'estimation dans l'imputation des valeurs provenant d'un des grands échantillons (source donneuse) à l'autre grand échantillon (source-hôte). Bien que la méthode soit élaborée principalement pour les valeurs nominales $Y$ et $Z$, une modification mineure permet de l'appliquer également à des valeurs $Y$ et $Z$ continues.

MOTS CLÉS: Cohérence entre les estimations; estimateur de régression générale; imputation; informations auxiliaires multidimensionnelles; tableau à double entrée.

## 1. INTRODUCTION

La plupart des enquêtes statistiques ont pour but d'obtenir de simples paramètres descriptifs relatifs à une population finie. Les estimations sont souvent présentées sous forme de tableaux comprenant des cellules qui contiennent les estimations de totaux de population ou de sous-groupes. Les données sont souvent recueillies d'après un vaste ensemble de variables, ce qui produit de nombreux résultats liés aux variables en question ainsi qu'aux rapports existant entre celles-ci. Afin d'économiser des ressources et de réduire le fardeau des répondants, les organismes qui s'occupent de statistique souhaitent réduire la taille des échantillons et la longueur des questionnaires. Ces organismes ont recours à des sources de données administratives et aux données d'enquêtes sur échantillon existantes qui sont menées à grande échelle, ou appliquent des plans d'échantillonnage à questionnaire fractionné (voir Raghunathan et Grizzle 1995). En raison de ces choix, les méthodes utilisées pour combiner des sources de données sont devenues des outils très prisés dans la production de données statistiques. La combinaison de sources de données peut être effectuée de nombreuses façons différentes; deux méthodes bien connues dans l'échantillonnage d'enquête sont l'appariement statistique et l'estimation de calage.

Singh, Mantel, Kinack et Rowe (1993) décrivent l'appariement statistique comme un cas particulier d'imputation qui comporte deux sources distinctes de micro-données contenant différentes informations relatives à des unités différentes. Une des sources de données sert de fichier hôte ou de fichier receveur dans lequel sont imputées de nouvelles informations pour chaque enregistrement à partir des données provenant de l'autre source, qui est le fichier donneur. Plus précisément, ces auteurs utilisent un fichier hôte A contenant des informations sur les variables $(X, Y)$ et un fichier donneur B contenant des informations sur les variables $(X, Z)$. La variable commune $X$ peut être utilisée pour identifier des unités similaires dans les deux fichiers. En général, l'appariement statistique a trait à la difficulté que comporte l'achèvement des enregistrements dans le fichier A en imputant des valeurs pour $Z$ à l'aide d'informations relatives au rapport existant entre $X$ et $Z$ qui proviennent du fichier B. Ces valeurs de $Z$ imputées comportent un inconvénient important parce que le rapport réel existant entre $Y$ et $Z$ peut être perdu en entier dans le fichier hôte qui est enrichi. Cet inconvénient correspond à ce qu'on appelle l'hypothèse d'indépendance conditionnelle entre $Y$ et $Z$ compte tenu de $X$. Afin de se débarrasser de cette hypothèse d'indépendance conditionnelle, Singh et coll. (1993) considèrent un troisième ensemble de données (fichier C) contenant des informations auxiliaires sur l'ensemble complet $(X, Y, Z)$. Cet ensemble de données pourrait provenir, par exemple, d'une enquête spéciale menée à une petite échelle. Ces auteurs présentent plusieurs méthodes d'imputation permettant de compléter le fichier A par l'ajout de $Z$, à partir du fichier B, à l'aide de renseignements provenant de A, B et C et portant sur les

[1] Robbert H. Renssen, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, Netherlands.

rapports existant entre $X$, $Y$ et $Z$. Singh et coll. (1993) donnent de nombreuses références bibliographiques pertinentes concernant les méthodes d'appariement statistique. Nous mentionnons ici uniquement Rodgers (1984), Rubin (1986) et Paass (1986).

Dans l'ouvrage de Deville et Särndal (1992), l'estimation de calage est utilisée comme une méthode générale servant à pondérer des enquêtes sur échantillon, en tenant compte du plan complexe d'échantillonnage et d'informations auxiliaires obtenues de sources extérieures (voir aussi Deville, Särndal et Sautory 1993). L'utilisation d'informations auxiliaires, c'est-à-dire de variables de contrôle, vise principalement trois objectifs: réduire la variance d'échantillonnage, réduire le biais dû à la non-réponse et assurer la cohérence entre les estimations établies à partir de diverses sources par rapport aux variables de contrôle qui sont utilisées. Il existe une volumineuse documentation sur les méthodes de pondération utilisées dans le cadre d'enquêtes sur échantillon. Nous citons à cet égard Bethlehem et Keller (1987), Alexander (1987), Lemaître et Dufour (1987), et Zieschang (1990).

Le présent article porte sur la façon d'estimer le produit croisé de $Y$ et $Z$ (par ex., le tableau à double entrée de $Y$ et $Z$ lorsque ces variables sont nominales, ou la covariance entre $Y$ et $Z$ lorsque ces variables sont continues) à l'aide de méthodes d'appariement statistique et de l'estimation de calage. Nous présupposons que deux fichiers de données A et B représentent deux vastes enquêtes sur échantillon, qui peuvent toutes deux être fondées sur un plan d'échantillonnage complexe. Afin de pondérer l'échantillon de l'enquête à petite échelle effectuée spécialement pour la combinaison de sources de données (fichier C), des informations auxiliaires sont tirées des deux grands échantillons. Il est peut-être difficile de déterminer si les grands échantillons devraient être considérés comme des sources d'informations auxiliaires pour l'échantillon de petite taille et vice-versa. L'appariement statistique permet d'intégrer un grand ensemble de variables $X$ à la méthode d'estimation et de répondre à certaines exigences en matière de cohérence. La majeure partie de l'article est consacrée aux variables nominales $Y$ et $Z$ en raison des caractéristiques de ces variables. Ainsi, nous montrons que les dénombrements marginaux du tableau $YZ$ estimé correspondent toujours aux estimations des totaux de population de $Y$ et de $Z$ lorsque l'estimateur de calage courant est appliqué en utilisant les variables $X$ comme variables de contrôle dans le cas des premier et deuxième échantillons de grande taille respectivement. Cependant, la méthode qui est proposée est également applicable à des variables $Y$ et $Z$ continues. Tout au long du présent article, nous présupposons que $X$ consiste en plusieurs variables, qui peuvent être nominales ou continues. Nous soutenons que lorsque les variables $X$ présentent une forte corrélation avec $Y$ ou $Z$, notre méthode d'estimation donne des estimations relativement justes du produit croisé de $Y$ et $Z$, par exemple dans le cas du tableau complet $YZ$ lorsque $Y$ et $Z$ sont des variables nominales.

La méthode d'estimation qui est proposée ressemble étroitement à une méthode présentée par Singh et coll. (1993, section 2) et qui sert à estimer un coefficient de corrélation entre $Y$ et $Z$. Dans le présent article, nous présupposons que ces variables sont unidimensionnelles. Cependant, notre méthode diffère de celle précitée par le fait qu'elle intègre à la méthode d'estimation le plan d'échantillonnage complexe de toutes les sources de données et qu'elle utilise les sources de données de grande taille d'une manière plus efficace dans l'estimation de paramètres de population provenant de la source de données de petite taille. Lorsque $Y$ et $Z$ sont des variables nominales et qu'il n'y a pas de corrélation linéaire entre $X$ et $Y$, ni entre $X$ et $Z$, notre méthode correspond à une poststratification incomplète (Deville et Särndal 1992, Bethlehem et Keller 1987). D'autre part, s'il y a une corrélation parfaite entre $Y$ et $X$, notre méthode donne un tableau estimé à double entrée de $Y$ et $Z$ qui correspond à un tableau estimé à double entrée qu'on obtiendrait à partir du fichier B si on imputait d'abord les valeurs de $Y$. On obtient un résultat similaire lorsqu'il y a une corrélation parfaite entre $Z$ et $X$.

Bien que la combinaison de sources de données distinctes par l'entremise de variables communes puisse être fructueuse d'un point de vue théorique, dans la pratique, on peut rencontrer des difficultés parce qu'on ne trouve pas facilement des variables communes au sens strict, principalement en raison de différences concernant les définitions, les méthodes d'observation et la période de référence. Ces difficultés peuvent être réduites si l'on harmonise les processus d'enquête à un stade initial. Une application prometteuse de l'utilisation de variables communes est le recours à des plans d'échantillonnage intégrés comme celui de la Dutch Household Survey on Living Conditions; voir van Tuinen (1995), Bakker et Winkels (1998), Winkels et Everaers (1998), et Hofmans (1998). Le questionnaire utilisé pour cette enquête comporte une structure à trois menus. Le premier menu contient des questions qui portent sur des aspects démographiques et socio-économiques, ainsi que sur le niveau d'instruction. Le deuxième menu contient des questions fondamentales auxquelles il est facile de répondre et qui portent sur tous les aspects pertinents des conditions de vie. Les questions contenues dans le troisième menu portent également sur les conditions de vie, mais elles sont plus exhaustives que les questions du deuxième menu. Afin de réduire le temps nécessaire pour répondre au questionnaire, le troisième menu est fractionné. Chaque répondant doit répondre à toutes les questions des premier et deuxième menus et remplir un sous-questionnaire du troisième menu. En raison du troisième menu, l'échantillon est divisé en sous-échantillons qui correspondent aux divers sous-questionnaires. Le plan d'échantillonnage de chaque sous-échantillon peut être décrit comme un échantillonnage à deux degrés pour l'estimateur de régression générale.

Le présent article est structuré de la manière décrite ci-après. Le cadre théorique est présenté dans la section 2.

Aux fins de cette présentation, il est pratique de décrire un estimateur de calage utilisé pour l'échantillon de petite taille, des informations auxiliaires étant obtenues de deux enregistrement distincts au lieu de deux grands échantillons distincts. Un des enregistrements contient des valeurs relatives à $X$ et à $Y$, tandis que l'autre contient des valeurs relatives à $X$ et à $Z$. Les sections 2.1 à 2.4 sont consacrées à des variables nominales $Y$ et $Z$. Dans la section 2.1, les enregistrements sont utilisés pour obtenir une première estimation synthétique du tableau $YZ$ par des méthodes d'imputation à régression. Nous montrons que ce tableau synthétique à double entrée possède des caractéristiques intéressantes. Dans la section 2.2, nous proposons un ensemble d'équations de calage qui servent à pondérer l'échantillon de petite taille d'après les caractéristiques en question. Nous décrivons brièvement le rapport qui existe entre cet ensemble et la poststratification complète, et entre cet ensemble et la poststratification incomplète. Nous donnons un exemple numérique dans la section 2.3. Le lien avec les méthodes d'appariement statistique, tel qu'il est présenté dans Singh et coll. (1993) est décrit dans la section 2.4. Le traitement des variables nominales $Y$ et $Z$ est excessivement restrictif. Dans la section 2.5, nous montrons que la méthode de pondération qui est proposée peut être appliquée également dans le cas de variables $Y$ et $Z$ continues ou dans le cas d'une variable $Y$ continue et d'une variable $Z$ nominale. Dans la section 3, nous modifions la méthode et utilisons des informations auxiliaires tirées de deux grands échantillons distincts plutôt que de deux enregistrements. À l'aide d'une étude de simulation, la méthode de pondération modifiée est comparée à la stratification double incomplète classique. Enfin, la section 4 est consacrée à quelques remarques conclusives.

## 2. COMBINAISON D'ENREGISTREMENTS PAR L'ENTREMISE DE VARIABLES COMMUNES

Prenons une population finie $\Omega = \{1, ..., N\}$ composée d'un nombre $N$ de personnes et supposons qu'il existe deux enregistrements concernant ces personnes. Le premier enregistrement contient, pour chaque personne $k$, un dossier contenant les scores $y_k$ et $x_k$ relatifs aux variables $Y$ et $X$ respectivement, tandis que le deuxième enregistrement contient, pour chaque personne $k$ un dossier contenant les scores $z_k$ et $x_k$ relatifs aux variables $Z$ et $X$ respectivement, $k = 1, ..., N$. Comme on peut voir, la variable $X$ est présente dans les deux enregistrements. Nous faisons remarquer que les dossiers des deux enregistrements correspondent à la même population finie. Le processus de fusion de ces deux enregistrements serait un appariement exact si l'on utilisait $X$ pour comparer les dossiers contenus dans l'un des enregistrements avec ceux contenus dans l'autre, dans le but de déterminer quelles paires de dossiers ont trait à la même unité de population (voir Fellegi et Sunter 1969). Dans le présent article, nous procédons d'une manière différente.

### 2.1 Établissement des totaux de population synthétiques

Soient $Y$ le niveau d'instruction avec un nombre de catégories $p$ et $Z$ l'emploi avec un nombre de catégories $q$. Nous avons alors $y_k$ comme vecteur d'ordre $p$, qui représente des variables factices $p$. Chaque variable factice correspond à une catégorie précise et est égale à 1 si la personne $k$ appartient à la catégorie en question, sinon elle est égale à 0. De façon analogue, $z_k$ est un vecteur d'ordre $q$. En outre, $X$ peut être le résultat d'un croisement (stratification) d'un certain nombre de caractéristiques (par ex.: sexe, âge, région, état civil, *etc.*). Les scores $x_k$ sont un vecteur d'ordre $r$. Lorsque $X$ consiste en une stratification complète, $x_k$ représente un nombre de variables factices $r$. Dans le reste du présent article, $r$ devrait être considéré comme un nombre important comparativement à $p \times q$. Les totaux de population dans le cas de $Y$ et $Z$ représentent les distributions statistiques marginales relatives au niveau d'instruction et à l'emploi. À l'aide de la variable commune $X$, les prédictions pour $Y$ et $Z$ peuvent être définies à l'aide d'un modèle de régression linéaire multiple:

$$\hat{y}_k = B' x_k, \quad k = 1, ..., N,$$

et

$$\hat{z}_k = A' x_k, \quad k = 1, ..., N,$$

où $B$ et $A$ représentent les coefficients ordinaires de régression des moindres carrés qui satisfont aux équations

$$\left( \sum_{k=1}^{N} x_k x_k' \right) B = \sum_{k=1}^{N} x_k y_k' \qquad (1)$$

et

$$\left( \sum_{k=1}^{N} x_k x_k' \right) A = \sum_{k=1}^{N} x_k z_k'. \qquad (2)$$

L'exposant «$t$» représente la transposition. Ce modèle est appelé modèle de probabilité linéaire (voir Maddala 1983, chap. 2). Il existe des modèles plus élégants, comme des modèles probit et logit, pour prédire des variables binaires. Cependant, nous ne nous intéressons pas aux prédictions elles-mêmes, mais plutôt aux totaux de population synthétiques de ces prédictions. Ces totaux semblent avoir des propriétés intéressantes lorsqu'on utilise le modèle de prédiction linéaire; c'est pourquoi il est possible de justifier l'utilisation de ce modèle. Il est à noter que $B$ est calculé à partir du premier enregistrement et $A$ à partir du deuxième. À l'aide de la variable commune $X$ et des coefficients de régression $B$ et $A$, nous créons un enregistrement synthétique qui contient un dossier pour chaque personne $k$, avec

les scores $x_k$, $B'x_k$, et $A'x_k$. En fait, on peut ajouter soit $y_k$, soit $z_k$ à l'enregistrement, mais dans le cas qui nous intéresse, cet ajout paraît superflu (voir prochain paragraphe). Lorsqu'il existe un vecteur $a$ d'ordre $r$ de nombres fixes qui fait que $a'x_k = 1$ dans le cas de toutes les personnes $k$, alors les totaux de population relatifs aux nouvelles variables $B'x_k$ et $A'x_k$ équivalent aux totaux de population relatifs aux variables initiales correspondantes (voir, par ex., Bethlehem et Keller 1987). Cela peut être démontré facilement en prémultipliant d'abord les équations normales (1) et (2) par $a'$ et en remplaçant par la suite $a'$ par $a'x_k = 1$ dans les équations qui sont obtenues.

À partir de l'enregistrement synthétique, on peut définir un tableau synthétique à double entrée par $\sum_{k=1}^{N}(B'x_k)(A'x_k)'$. Ce tableau synthétique à double entrée peut être considéré comme une approximation de la distribution statistique (simultanée) $\sum_{k=1}^{N} y_k z_k'$. À l'aide des équations normales (1) et (2), ont peut dériver les identités suivantes:

$$\sum_{k=1}^{N}(B'x_k)(A'x_k)' = \sum_{k=1}^{N} y_k(A'x_k)' = \sum_{k=1}^{N}(B'x_k)z_k'.$$

De toute évidence, les croisements entre $B'x_k$ $A'x_k$, $y_k$ et $A'x_k$, ou entre $B'x_k$ et $z_k$ donnent tous des tableaux synthétiques à double entrée identiques. Par conséquent, il suffit de considérer uniquement $\sum_{k=1}^{N}(B'x_k)(A'x_k)'$ et de supprimer $y_k$ ou $z_k$ dans l'enregistrement synthétique. La différence entre, d'une part, la distribution statistique réelle entre $Y$ et $Z$ et, d'autre part, son «approximation» synthétique peut être obtenue à l'aide de la décomposition suivante:

$$\sum_{k=1}^{N} y_k z_k' = \sum_{k=1}^{N}(B'x_k)(A'x_k)' + \sum_{k=1}^{N}(y_k - B'x_k)(z_k - A'x_k)'. \qquad (3)$$

On remarquera la forte ressemblance avec la décomposition ordinaire de la variance effectuée dans l'analyse de régression (voir, par ex., Searle 1971). Lorsque $B'x_k = y_k$ ou $A'x_k = z_k$ dans le cas de toutes les personnes $k$, le tableau à double entrée dérivé de l'enregistrement synthétique équivaut à la distribution statistique simultanée réelle existant entre $Y$ et $Z$.

Soit $l$ un vecteur d'un ordre approprié consistant en des un, et $l'y_k = 1$ et $l'z_k = 1$ dans le cas de toutes les personnes $k$. S'il existe une constante $a$ qui fait que $a'x_k = 1$ dans le cas de toutes les personnes $k$, nous avons alors également

$$l'\hat{y}_k = l'B'x_k = l'\left(\sum_{k=1}^{N} y_k x_k'\right)\left(\sum_{k=1}^{N} x_k x_k'\right)^{-1} x_k = a'\left(\sum_{k=1}^{N} x_k x_k'\right)\left(\sum_{k=1}^{N} x_k x_k'\right)^{-1} x_k = a'x_k = 1$$

dans le cas de toutes les personnes $k$, et, de façon analogue, $l'\hat{z}_k = l'A'x_k = 1$ dans le cas de toutes les personnes $k$. Il s'ensuit que

$$l'\sum_{k=1}^{N}(B'x_k)(A'x_k)' = \sum_{k=1}^{N}(A'x_k)' = \sum_{k=1}^{N} z_k' \qquad (4)$$

et

$$\sum_{k=1}^{N}(B'x_k)(A'x_k)'l = \sum_{k=1}^{N}(B'x_k) = \sum_{k=1}^{N} y_k. \qquad (5)$$

Ainsi les totaux de ligne et de colonne du tableau synthétique à double entrée équivalent aux dénombrements marginaux de population correspondants en ce qui a trait à $Y$ et $Z$.

Ce qu'il reste à considérer est la situation lorsque $a'x_k = 1$ dans le cas de toutes les valeurs $k$ et d'une constante quelconque $a$. Cette condition est satisfaite lorsque $X$ représente une variable nominale. De façon plus générale, cette condition est toujours satisfaite si le vecteur $X$ peut être divisé en deux sous-vecteurs, dont l'un représente une variable nominale.

### 2.2 Formulation des contraintes dans l'estimation de calage

Supposons qu'un échantillon aléatoire $s$ de taille $n$ est tiré de la population finie $\Omega = \{1, ..., N\}$ d'après un plan d'échantillonnage $p(s)$, de sorte que les probabilités d'inclusion de premier et de deuxième ordre $\Pr(k \in s) = \pi_k$ et $\Pr(k, l \in s) = \pi_{kl}$ sont strictement positives. Dans le cas de chaque $k \in s$, on observe le vecteur de scores $(x_k, y_k, z_k)$. On dispose de deux enregistrements distincts pour obtenir des informations auxiliaires. Le premier enregistrement contient, dans le cas de chaque $k \in \Omega$, des dossiers avec des scores relatifs à $x_k$ et à $y_k$, tandis que le deuxième enregistrement contient, dans le cas de chaque $k \in \Omega$, des scores relatifs à $x_k$ et à $z_k$. Le but est d'estimer le tableau $YZ$ à partir de l'échantillon $s$, en utilisant des informations auxiliaires tirées des deux enregistrements. Il existe un large éventail d'estimateurs à pondération qui peuvent être utilisés en présence d'informations auxiliaires multidimensionnelles. Dans Särndal, Swensson et Wretman (1992), l'estimateur de régression générale est décrit de manière approfondie. Il définit implicitement des poids d'échantillon qui reproduisent les totaux de population connus relatifs

aux variables auxiliaires qui sont utilisées comme variables de contrôle dans l'estimateur. Une telle propriété de cohérence est intéressante si les informations auxiliaires sont utilisées tant pour la diffusion que pour la pondération. L'estimateur de calage est créé sous forme de généralisation de l'estimateur de régression générale (Deville et Särndal 1992, et Deville et coll. 1993).

Afin de donner un exemple précis, soit $G$ une fonction fixe telle que définie dans Deville et coll. (1993) et considérons l'estimateur à pondération suivant dans le cas de notre tableau $YZ$:

$$\hat{T} = \sum_{k=1}^{n} w_k (y_k z_k'), \tag{6}$$

où $w_k$ est un scalaire qui représente un poids attribué à la personne $k \in s$. Soit $d_k = \pi_k^{-1}$. Un estimateur de calage pour le tableau $YZ$ utilise des poids qui sont obtenus en minimisant $\sum_{k=1}^{n} d_k G(w_k/d_k)$ par rapport à $w_k$, soumis à un ensemble de contraintes imposées à $w_k$, dans le cas de tout échantillon $s$. Nous considérons d'abord l'ensemble de contraintes suivant:

$$\sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{N} y_k \text{ et } \sum_{k=1}^{n} w_k z_k = \sum_{k=1}^{N} z_k. \tag{I}$$

Ce (premier) ensemble de contraintes utilise uniquement les dénombrements (marginaux) relatifs à $Y$ et $Z$. La variable commune $X$ n'est pas utilisée. Une des $p + q$ équations est redondante, de sorte que pour résoudre le problème de la minimisation, une des équation peut être supprimée. Dans le cas de $G(w_k/d_k) = (w_k/d_k - 1)^2$, l'estimateur de calage qui est obtenu correspond à une stratification double incomplète, telle que définie dans Bethlehem et Keller (1987). En utilisant $G(w_k/d_k) = 1 + w_k/d_k(\log(w_k/d_k) - 1)$, on obtient l'estimateur classique itératif du quotient (voir, par ex., Oh et Scheuren 1987). Copeland, Peitzmeier et Hoy (1987) ont comparé ces méthodes, d'après des données de la Current Population Survey. Ces auteurs en arrivent à la conclusion que les estimations obtenues à l'aide des deux méthodes sont très semblables. Dans Deville et coll. (1993), sont présentées deux autres fonctions de distance, qui sont particulièrement intéressantes compte tenu du problème lié aux poids extrêmes. L'estimation de tableaux à double entrée avec des contraintes sur les dénombrements marginaux est souvent effectuée dans le cas d'enquête sur échantillon. Souvent, les contraintes imposées aux dénombrements marginaux sont nécessaires pour deux raisons: premièrement, afin de réduire l'erreur et le biais d'échantillonnage, et deuxièmement, pour répondre à des exigences de cohérence relatives à des dénombrements de populations qui sont publiés.

Supposons que $x_k$ est un score nominal qui comporte un nombre de catégories $r$. Étant donné qu'on dispose d'informations sur la population relatives aux croisements entre $Y$ et $X$, et aux croisements entre $Z$ et $X$, nous pourrions aussi considérer l'ensemble de contraintes suivant:

$$\sum_{k=1}^{n} w_k (y_k x_k') = \sum_{k=1}^{N} y_k x_k' \text{ et}$$

$$\sum_{k=1}^{n} w_k (z_k x_k') = \sum_{k=1}^{N} z_k x_k'.$$

Le nombre de contraintes non redondantes dans cet ensemble est égal à $r(p + q - 1)$. Dans le cas d'une valeur $r$ plus grande, cet ensemble peut ne pas être applicable parce qu'il contient trop de contraintes par rapport à la taille de l'échantillon. L'ensemble pourrait présenter un intérêt pratique uniquement lorsque $r$ est petit. Dans le reste du présent article, nous ne tiendrons pas compte de cet ensemble de contraintes.

En vue d'intégrer un ensemble important de variables communes à la méthode de pondération, nous considérons un ensemble de contraintes qui fait appel aux informations bidimensionnelles sur la population que nous avons dans le tableau synthétique:

$$\sum_{k=1}^{n} w_k (B'x_k)(A'x_k)' = \sum_{k=1}^{N} (B'x_k)(A'x_k)'. \tag{II}$$

Ce (deuxième) ensemble de contraintes est une application directe de la théorie des estimateurs de calage. Les totaux de population relatifs aux croisements entre $B'x_k$ et $A'x_k$ sont connus; ces combinaisons sont donc utilisées comme variables auxiliaires pour formuler l'ensemble de contraintes. Évidemment, dans le cas d'une valeur $r$ importante, le nombre de contraintes non redondantes demeure limité par $p \times q$. Un inconvénient important des poids de calage qui sont obtenus est que ceux-ci ne reproduisent pas nécessairement les dénombrements (marginaux) de population en ce qui a trait à $Y$ et $Z$ lorsqu'on applique ces poids à $y_k$ et à $z_k$ respectivement. Autrement dit, les poids de calage obtenus ne satisfont pas nécessairement au premier ensemble de contraintes. Il s'agit là d'un inconvénient important, surtout si cet ensemble est formulé en vue de respecter des exigences de cohérence.

Par conséquent, comme solution de remplacement, nous considérons un troisième ensemble de contraintes:

$$\sum_{k=1}^{n} w_k \left( y_k z_k' - (y_k - B'x_k)(z_k - A'x_k)' \right) =$$

$$\sum_{k=1}^{N} (B'x_k)(A'x_k)' \tag{III}$$

En présupposant qu'il existe une constante $a$ qui fait que $a'x_k = 1$ dans le cas de toutes les personnes $k$, cet ensemble de contraintes permet d'atteindre l'objectif de cohérence. Soit $l$ un vecteur de un d'un ordre approprié, et rappelons-nous que $l'y_k = l'B'x_k = l'z_k = l'A'x_k = 1$ dans le cas de toutes les personnes $k$, que $B'\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} y_k$, et que

$A' \sum_{k=1}^{N} x_k = \sum_{k=1}^{N} z_k$. En prémultipliant le troisième ensemble d'équations, des deux côtés, par $l'$, nous obtenons le premier ensemble de contraintes en ce qui a trait à $Z$, tandis qu'en postmultipliant le troisième ensemble des deux côtés par $l$, on obtient le premier ensemble de contraintes relativement à $Y$. L'estimateur de calage qui en résulte peut être exprimé comme suit:

$$\hat{T} = \sum_{k=1}^{n} w_k (y_k z_k') = \sum_{k=1}^{N} (B' x_k)(A' x_k)' + \sum_{k=1}^{N} w_k (y_k - B' x_k)(z_k - A' x_k)'.$$

De toute évidence, cet estimateur suit la décomposition donnée par (3). Il est égal au tableau à deux entrées défini synthétiquement, un terme de correction en plus. Ce terme de correction est une estimation de calage de la différence relative à la distribution statistique réelle existant entre, d'une part, $Y$ et $Z$, et d'autre part le tableau à deux entrées défini synthétiquement. De façon analogue au deuxième ensemble de contraintes, dans le troisième ensemble, le nombre de contraintes non redondantes est borné par $p \times q$.

Un cas spécial important est celui où $G(w_k/d_k) = (w_k/d_k - 1)^2$. Dans ce cas, chaque cellule estimée est une estimation de régression générale avec $(y_k z_k)$, $\text{vec}(B' x_k x_k' A)$, et $\text{vec}(y_k z_k' - (y_k - B' x_k)(z_k - A' x_k)')$ comme variables de contrôle, dans le cas des premier, deuxième et troisième ensembles de contraintes respectivement. Des formules analytiques pour la variance de plan d'échantillonnage de l'estimateur de régression générale sont données, par exemple, dans Särndal et coll. (1992, chap. 6). En fait, ces formules sont des approximations pour des tailles d'échantillon importantes. Dans Deville et Särndal (1992) sont décrites les conditions suffisantes dans lesquelles ces approximations sont valides, dans le cas d'estimateurs de calage en général.

Dans Deville et coll. (1993), la poststratification complète est décrite comme une méthode de calage pour laquelle on utilise tous les dénombrements de population relatifs à la classification croisée dans l'ensemble de contraintes. L'élaboration d'une poststratification complète donne l'estimateur de poststratification complète ordinaire, quelle que soit la fonction de distance $G$. Solution de remplacement, la poststratification incomplète est décrite comme une méthode de calage dans laquelle on utilise une connaissance moins détaillée de tous les dénombrements de cellule dans l'ensemble de contraintes. L'estimateur de calage défini d'après le premier ensemble de contraintes est un exemple couramment utilisé de poststratification incomplète. Plusieurs cas dans lesquels la poststratification incomplète est préférable à la poststratification complète ont été décrits. Deux de ces cas sont le manque d'informations relatives à la population et certains dénombrements de cellule zéro ou extrêmement petits (voir aussi Oh et Scheuren 1987). L'estimateur de calage défini d'après les deuxième et troisième ensembles de contraintes correspond à la poststratification, dans le sens que tous les croisements sont utilisés comme informations auxiliaires. À l'exception du cas où il existe un rapport linéaire parfait, soit entre $Y$ et $X$, soit entre $Z$ et $X$, la méthode diffère de la poststratification complète par l'utilisation de totaux de population synthétiques plutôt que des dénombrements de population réels. La poststratification complète donne des résultats instables si certaines cellules d'échantillon contiennent seulement quelques observations. Dans de tels cas, la poststratification incomplète présente un intérêt sur le plan pratique. De façon analogue, l'estimateur de calage défini d'après les deuxième et troisième ensembles de contraintes peut être instable. Comme dans le cas de la poststratification incomplète, on peut considérer plutôt l'utilisation d'un croisement incomplet dans les contraintes.

### 2.3 Un exemple numérique

Nous illustrons l'utilisation de l'estimateur de calage dans le cas de trois ensembles de contraintes différents, à l'aide d'un exemple hypothétique. L'exemple est fondé sur des données réelles tirées d'un échantillon pour la Dutch National Travel Survey (1994). Le plan d'échantillonnage est grosso modo un échantillon en grappes d'adresses autopondéré. Toutes les personnes vivant à une adresse choisie sont incluses dans l'échantillon. La taille nette de l'échantillon est d'environ 80 000 personnes réparties dans 34 000 adresses. À partir de cet échantillon sont établis deux enregistrements hypothétiques d'environ $N = 80\ 000$ personnes. Dans l'un des enregistrements, on inscrit l'âge (dans six catégories) et dans l'autre, la possession d'une automobile (dans deux catégories). La variable commune entre les enregistrements est un nombre clé pour les adresses, ce qui donne $r = 34\ 000$ catégories dans le cas de la variable $X$. Dans le cas de cet exemple en particulier, le tableau synthétique à double entrée est simplifié comme suit:

$$\sum_{k=1}^{N} (B' x_k)(A' x_k)' = \sum_{j=1}^{r} N_j \bar{y}_j \bar{z}_j',$$

où $N_j$ représente la taille de l'adresse $j$, $\bar{y}_j$ la moyenne des six catégories d'âge de l'adresse $j$, et $\bar{z}_j$ la moyenne des deux catégories de possession d'une automobile de l'adresse $j$.

Afin de calculer le tableau synthétique à double entrée, les deux enregistrements sont combinés comme suit: premièrement, les enregistrements font l'objet d'un tri d'après le numéro clé pour les adresses; deuxièmement, on calcule les dénombrements des six catégories d'âge et des deux catégories de possession d'une automobile troisièmement, chaque dénombrement d'adresses relatif à l'âge est relié à son dénombrement d'adresses correspondant relatif

à la possession d'une automobile. À l'aide de cet enregistrement synthétique, de $r$ = 34 000, on peut calculer le tableau synthétique à double entrée. Le résultat est présenté dans le tableau 1. Ce tableau peut être considéré comme une première approximation de la distribution statistique réelle entre l'âge et la possession d'une voiture. Une condition suffisante pour une approximation étroite est l'homogénéité en ce qui a trait à l'âge ou à la possession d'une automobile à l'intérieur de toutes les adresses, c'est-à-dire que toutes les personnes vivant à une même adresse devraient être soit dans la même catégorie d'âge, soit dans la même catégorie de possession d'une voiture. Dans le cas de la plupart des adresses à plus d'une personne, cette proposition semble peu probable. Il s'ensuit des équations (4) et (5) que les totaux de ligne et de colonne du tableau 1 correspondent aux dénombrements (marginaux) de population réels relatifs à l'âge et la possession d'une automobile respectivement.

À l'aide d'un échantillon aléatoire simple de $n$ = 1 000 personnes, les dénombrements de population sont estimés au moyen d'un estimateur de régression générale. Trois ensembles de variables auxiliaires sont utilisés, conformément aux trois ensembles de contraintes mentionnés dans la section précédente. Les tableaux estimés sont indiqués ci-après (pour plus de commodité, nous avons utilisé la mesure de distance quadratique: $G(w_k/d_k) = (w_k/d_k - 1)^2$). Les écarts-types estimés correspondants sont indiqués entre parenthèses. Ces écarts-types estimés sont fondés sur les formules de variance habituelles de l'estimateur de régression générale (voir Särndal et coll. 1992, chap. 6).

**Tableau 1**
Totaux de population synthétiques dans le cas de croisements de variables relatives à l'âge et à la possession d'une automobile

| | 1 | 2 | 3 | 4 | 5 | 6 | total |
|---|---|---|---|---|---|---|---|
| oui | 3 461 | 1 659 | 5 739 | 10 770 | 6 536 | 3 334 | 31 499 |
| non | 9 827 | 4 692 | 7 902 | 17 102 | 6 424 | 5 389 | 51 336 |
| total | 13 288 | 6 351 | 13 641 | 27 872 | 12 960 | 8 723 | 82 835 |

Dans le tableau 2, les dénombrements de population sont estimés d'après la stratification incomplète double ordinaire (Bethlehem et Keller 1987). Il n'y a pas de jeunes (catégories d'âge 1 et 2) qui possèdent une automobile dans l'échantillon, ce qui est une constatation susceptible d'être représentative de la population; ces cellules sont donc estimées à zéro. En raison des exigences de conversence, c'est-à-dire le premier ensemble de contraintes, les dénombrements estimés de jeunes ne possédant pas une automobile équivalent aux dénombrements de cellule marginaux correspondants. Une façon d'essayer d'améliorer le tableau 2 consiste à utiliser l'adresse à variable commune dans la méthode de pondération. Dans le tableau 3, les estimations relatives aux cellules sont indiquées d'après le deuxième ensemble de contraintes. Comme on l'a déjà mentionné dans la section précédente, les totaux estimés de ligne et de colonne peuvent différer des dénombrements de population réels. Une comparaison entre le tableau 2 et le tableau 3 montre que ces différences peuvent être importantes. En outre, presque tous les dénombrements de cellule estimés du tableau 2 présentent des écarts-types estimés plus faibles que les dénombrements de cellule estimés correspondants du tableau 3. Ainsi, le deuxième ensemble de contraintes donne des résultats assez insatisfaisants. Cela implique deux choses: 1) une cohérence des dénombrements de cellule marginaux estimés par rapport aux dénombrements de population connus correspondants; 2) des variances asymptotiques plus faibles de tous les dénombrements de cellule estimés. Les résultats sont indiqués dans le tableau 4. Les dénombrements marginaux estimés de cellule sont effectivement cohérents, et les écarts-types estimés sont, au plus, deux fois moins importants que les écarts-types estimés correspondants indiqués dans le tableau 2.

**Tableau 2**
Totaux de population estimés dans le cas de croisements entre des variables relatives à l'âge et à la possession d'une automobile, d'après le premier ensemble de contraintes

| | 1 | 2 | 3 | 4 | 5 | 6 | total |
|---|---|---|---|---|---|---|---|
| oui | 0 (0) | 0 (0) | 4 968 (423) | 15 414 (543) | 7 518 (458) | 3 599 (375) | 31 499 |
| non | 13 288 (0) | 6 351 (0) | 8 673 (423) | 12 458 (543) | 5 422 (458) | 5 124 (375) | 51 336 |
| total | 13 288 | 6 351 | 13 641 | 27 872 | 12 960 | 8 723 | 82 835 |

**Tableau 3**
Totaux de population estimés dans le cas de croisements entre des variables relatives à l'âge et à la possession d'une automobile, d'après le deuxième ensemble de contraintes

| | 1 | 2 | 3 | 4 | 5 | 6 | total |
|---|---|---|---|---|---|---|---|
| oui | 0 (0) | 0 (0) | 4 791 (435) | 13 826 (811) | 6 887 (494) | 3 421 (321) | 28 923 (1005) |
| non | 14 385 (782) | 7 012 (595) | 8 118 (563) | 12 893 (796) | 5 853 (464) | 5 654 (306) | 53 912 (1005) |
| total | 14 385 (782) | 7 012 (595) | 12 908 (603) | 26 718 (958) | 12 739 (419) | 9 074 (177) | 82 835 |

**Tableau 4**
Totaux de population estimés dans le cas de croisements entre des variables relatives à l'âge et à la possession d'une automobile, d'après le troisième ensemble de contraintes

| | 1 | 2 | 3 | 4 | 5 | 6 | total |
|---|---|---|---|---|---|---|---|
| oui | 0 (0) | 0 (0) | 5 501 (226) | 15 647 (227) | 6 898 (177) | 3 453 (78) | 31 499 |
| non | 13 288 (0) | 6 351 (0) | 8 139 (226) | 12 224 (227) | 6 062 (177) | 5 270 (78) | 51 336 |
| total | 13 288 | 6 351 | 13 641 | 27 872 | 12 960 | 8 723 | 82 835 |

## 2.4 Imputation de valeurs tirées d'un des enregistrements dans l'autre enregistrement

Jusqu'ici, nous avons élaboré une méthode de pondération pour estimer un tableau à double entrée réunissant deux variables qui sont inscrites dans deux enregistrements distincts. Souvent, on voudrait disposer non seulement de

tableaux estimés à double entrée ou de rapports linéaires estimés plus généraux, mais également d'enregistrements complets dans lesquels sont inscrites simultanément les deux variables. Les utilisateurs de statistiques trouvent ce genre de bases de données complètes faciles à analyser. La création de tels enregistrements enrichis peut être considérée comme un cas spécial d'imputation. Un des enregistrements sert de source hôte ou de receveur, tandis que l'autre enregistrement sert de source donneuse. En admettant que le deuxième enregistrement est la source donneuse, la difficulté consiste à imputer dans le premier enregistrement les valeurs de $Z$ tirées du deuxième enregistrement, en utilisant les informations contenues dans le tableau estimé à double entrée que nous avons décrit dans la section 2.2 comme informations auxiliaires. Des difficultés d'appariement statistique dans l'utilisation de données tirées d'une troisième source ont déjà été décrites par Rubin (1986) et Paass (1986). Singh et coll. (1993) présentent un examen de leurs méthodes. En outre, ils proposent certaines modifications des méthodes de Rubin (1986) et Paass (1986). Notre propre méthode d'imputation est fondée sur la méthode de régression proposée par Rubin (1986) et par Singh et coll. (1993).

Après avoir défini des variables prédictives pour les variables $Z$ à l'aide du modèle de régression

$$\hat{z}_k = A^t x_k,\ k = 1, ..., N,$$

où $A$ est donné par (2), nous définissons de nouvelles prédictions de ces variables au moyen du modèle de régression élargi

$$\tilde{z}_k = \alpha_1^t x_k + \alpha_2^t y_k,\ k = 1, ..., N,$$

avec

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \left[ \sum_{k=1}^{N} \begin{pmatrix} x_k x_k^t & x_k y_k^t \\ y_k x_k^t & y_k y_k^t \end{pmatrix}^{-1} \right] \left[ \sum_{k=1}^{N} \begin{pmatrix} x_k z_k^t \\ y_k z_x^t \end{pmatrix} \right].$$

À l'aide de résultats bien connus concernant les coefficients de régression partielle du modèle linéaire général (voir, par ex., Seber 1977), $\alpha_1$ et $\alpha_2$ peuvent être exprimés sous la forme

$$\alpha_1 = A - B\alpha_2$$

et sous la forme

$$\alpha_2 = \left[ \sum_{k=1}^{N} (y_k - B^t x_k)(y_k - B^t x_k)^t \right]^{-1} \times \left[ \sum_{k=1}^{N} (y_k - B^t x_k)(z_k - A^t x_k)^t \right],$$

où $B$ et $A$ sont donnés par (1) et (2) respectivement. Ils peuvent être calculés à partir des premier et deuxième enregistrements. Les coefficients de régression partielle devraient être estimés à partir de la troisième source. Nous proposons

$$\hat{\alpha}_1 = A - B\hat{\alpha}_2$$

et

$$\hat{\alpha}_2 = \left[ \sum_{k=1}^{N} (y_k - B^t x_k)(y_k - B^t x_k)^t \right]^{-1} \times \left[ \sum_{k=1}^{n} w_k (y_k - B^t x_k)(z_k - A^t x_k)^t \right],$$

où $w_k$ représente des poids de calage qui sont décrits dans la section 2.2. D'après ces estimations, nous définissons de nouvelles prédictions pour les valeurs $Z$:

$$\hat{z}_k = \hat{\alpha}_1^t x_k + \hat{\alpha}_2^t y_k = A^t x_k + \hat{\alpha}_2^t (y_k - B^t x_k),\ k = 1, ..., N. \quad (7)$$

Ces nouvelles prédictions sont égales aux anciennes (voir section 2.1), un terme de correction en plus. Ce terme de correction dépend de la différence entre la valeur de $Y$ et son (ancienne) prédiction. Cependant, on peut considérer cela comme une tentative d'amélioration de la prédiction relative à $Z$ et, plus important encore, comme un moyen de reconstruire l'estimateur à pondération d'après le troisième ensemble de contraintes (section 2.2). En effet, l'égalité suivante est vérifiée:

$$\sum_{k=1}^{N} y_k \hat{z}_k^t = \sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t + \sum_{k=1}^{n} w_k (y_k - B^t x_k)(z_k - A^t x_k)^t.$$

Il s'agit là uniquement de l'estimateur à pondération établi d'après le troisième ensemble de contraintes lorsque les poids de calage correspondants sont utilisés pour estimer $\alpha_2$. Il est facile de montrer que

$$\sum_{k=1}^{N} x_k \hat{z}_k^t = \sum_{k=1}^{N} x_k \hat{z}_k^t = \sum_{k=1}^{N} x_k z_k^t.$$

Ainsi, le tableau $XZ$ peut également être reconstruit. Au début de la présente section, nous avons présupposé que le deuxième enregistrement est la source donneuse. Ce choix est arbitraire. Si nous avions imputé les valeurs de $Y$ au lieu des valeurs de $Z$, nous aurions obtenu une estimation identique pour le tableau $YZ$. En outre, le tableau $XY$ aurait pu être reconstruit.

Les nouvelles prédictions relatives aux valeurs de $Z$ peuvent être utilisées pour l'imputation. Singh et coll. (1993) donnent des algorithmes pour l'imputation en utilisant des modèles de régression. Ces valeurs de $Z$ peuvent être imputées dans le premier enregistrement en

deux étapes. Lors de la première étape, les prédictions données par (7) sont calculées pour chaque $(x_k, y_k)$ dans le premier enregistrement. Nous avons montré que les croisements entre les valeurs de $Y$ et ces valeurs prédites de $Z$ peuvent être considérés comme des estimateurs à pondération. Cependant, les prédictions calculées n'ont en général pas de valeur réaliste; c'est pourquoi la première étape est suivie d'une deuxième. Dans cette deuxième étape, chaque valeur prédite de $Z$ qui se trouve dans le premier enregistrement est remplacée par une valeur réelle de $Z$, tirée du deuxième enregistrement, qui se trouve le plus près dans $(X, Z)$ d'après une certaine distance euclidienne

### 2.5 Estimation de produits croisés dans le cas de variables $Y$ et $Z$ continues

La caractéristique de cohérence du troisième ensemble de contraintes (section 2.2) est vérifiée également par rapport à des variables $Y$ et $Z$, pourvu qu'il y ait des constantes $a_y$ et $a_z$ d'un ordre approprié, telles que $a_y' y_k = 1$ et $a_z' z_k = 1$ dans le cas de toutes les personnes $k$. Afin de voir cela, nous élargissons légèrement les résultats de la section 2.1. Notons d'abord que

$$a_y' B' x_k = a_y' \sum_{k=1}^{N} y_k x_k' \left( \sum_{k=1}^{N} x_k x_k' \right)^{-1} x_k =$$

$$a' \sum_{k=1}^{N} x_k x_k' \left( \sum_{k=1}^{N} x_k x_k' \right)^{-1} x_k = a' x_k = 1$$

(on présume toujours qu'il existe une constante $a$ qui fait que $a' x_k = 1$ dans le cas de toutes les personnes $k$). De manière analogue, $a_z' A' x_k = 1$. Les équations équivalentes de (4) et (5) pour les variables continues sont obtenues facilement. Par conséquent, en prémultipliant les deux côtés de (III) avec $a_y'$, on obtient $\sum_{k=1}^{n} w_k z_k' = \sum_{k=1}^{N} z_k'$, et en postmultipliant les deux côtés de (III) avec $a_z$, on obtient $\sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{N} y_k$. Ainsi, le troisième ensemble de contraintes permet d'atteindre l'objectif de cohérence, c'est-à-dire l'équation de calage du premier ensemble de contraintes, dans le cas de variables $Y$ et $Z$ assez générales. Voici deux exemples.

Dans le premier exemple, nous prenons $y_k = (1, y_{2k})'$ et $z_k = (1, z_{2k})'$, où l'on présuppose que $y_{2k}$ et $z_{2k}$ sont des variables continues. En prenant $a_y = a_z = (1, 0)'$, nous voyons que $a_y' y = a_z' z = 1$ dans le cas de toutes les personnes $k$. Le produit croisé de $Y$ et $Z$ est égal à

$$\sum_{k=1}^{N} y_k z_k' = \begin{pmatrix} N & \sum_{k=1}^{N} z_{2k} \\ \sum_{k=1}^{N} y_{2k} & \sum_{k=1}^{N} y_{2k} z_{2k} \end{pmatrix},$$

d'où l'on peut dériver facilement la covariance existant entre $y_{2k}$ et $z_{2k}$. Ce produit croisé peut être estimé à l'aide du troisième ensemble de contraintes. Une élaboration de cet ensemble donne les quatre contraintes suivantes dans le cas de l'exemple qui nous intéresse:

$$\sum_{k=1}^{n} w_k = N, \sum_{k=1}^{n} w_k y_{2k} = \sum_{k=1}^{N} y_{2k}, \sum_{k=1}^{n} w_k z_{2k} = \sum_{k=1}^{N} z_{2k},$$

et

$$\sum_{k=1}^{n} w_k (y_{2k} z_{2k} - (y_{2k} - B_2' x_k)(z_{2k} - A_2' x_k)) = \sum_{k=1}^{N} (B_2' x_k)(A_2' x_k),$$

où les coefficients de régression sont donnés par

$$B_2 = \left( \sum_{k=1}^{N} x_k x_k' \right)^{-1} \sum_{k=1}^{N} x_k y_{2k}$$

et par

$$A_2 = \left( \sum_{k=1}^{N} x_k x_k' \right)^{-1} \sum_{k=1}^{N} x_k z_{2k}.$$

Si l'on s'intéresse en particulier au coefficient de corrélation entre $y_{2k}$ et $z_{2k}$, les contraintes suivantes pourraient également être prises en considération:

$$\sum_{k=1}^{n} w_k y_{2k}^2 = \sum_{k=1}^{N} y_{2k}^2 \text{ et } \sum_{k=1}^{n} w_k z_{2k}^2 = \sum_{k=1}^{N} z_{2k}^2.$$

Dans le deuxième exemple, nous supposons que $y_k = (1, y_{2k})'$, où $y_{2k}$ peut être continue et où $z_k$ est nominale avec un nombre de catégories $q$. En prenant $a_y = (1, 0)'$ et $a_z = l$, où $l$ est un vecteur de un d'un ordre approprié, nous constatons que $a_y' y_k = a_z' z_k = 1$ dans le cas de toutes les personnes $k$. Le produit croisé de $Y$ et $Z$ est

$$\sum_{k=1}^{N} y_k z_k' = \begin{pmatrix} N_1 & N_2 & \cdots & N_q \\ \sum_{k \in C_1} y_{2k} & \sum_{k \in C_2} y_{2k} & \cdots & \sum_{k \in C_q} y_{2k} \end{pmatrix},$$

où $C_h$ représente l'ensemble d'éléments de population appartenant à la catégorie $h$ de $Z$, et $N_h$ la taille de $C_h$. Nous nous assurons que les poids de calage relatifs au troisième ensemble de contraintes satisfont aux équations «marginales» de calage. $\sum_{k=1}^{n} w_k z_k = \sum_{k=1}^{N} z_k = (N_1 \ldots N_q)'$ et $\sum_{k=1}^{n} w_k y_{2k} = \sum_{k=1}^{N} y_{2k}$, qui peuvent être toutes deux intéressantes en ce qui a trait aux exigences de cohérence.

## 3. COMBINAISON D'ÉCHANTILLONS INDÉPENDANTS PAR L'ENTREMISE DE VARIABLES COMMUNES

Dans la section précédente, nous avons présenté une méthode permettant de combiner deux enregistrements par

l'entremise de variables communes et en utilisant des informations auxiliaires tirées d'un échantillon de petite taille. Dans la présente section, nous adaptons cette méthode en combinant deux échantillons indépendants. Nous considérons un enregistrement complet de personnes, deux enquêtes sur échantillon à grande échelle et une enquête sur échantillon à petite échelle. L'enregistrement contient un ensemble limité de variables comme le sexe, l'âge, la région et l'état civil. Ces variables sont désignées par $X$. Dans l'un des grands échantillons, on observe les variables $Y$, $U$ et $X$, tandis que dans l'autre grand échantillon, on observe les variables $Z$, $U$ et $X$. Dans l'échantillon de petite taille, on observe toutes les variables, c'est-à-dire $Y$, $Z$, $U$ et $X$. L'échantillon de petite taille peut provenir d'une enquête à petite échelle effectuée spécialement aux fins de la combinaison, ou du chevauchement des deux grandes enquêtes. Dans la figure 1, les sources de données sont indiquées schématiquement. Pour plus de commodité, on présuppose que tous les échantillons correspondent à des unités différentes, c'est-à-dire que l'on présume qu'il n'y a pas de chevauchement des échantillons.

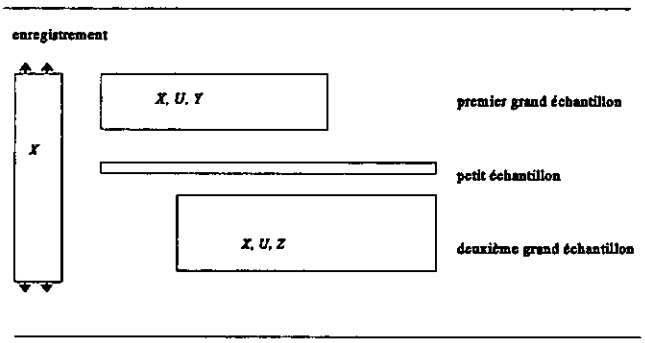



**Figure 1.** Aperçu des diverses sources de données

Les variables $X$ et $U$ sont réparties dans $C = (X\ U)$, où $X$ désigne l'ensemble de variables communes comprenant des totaux de population connus et $U$ l'ensemble de variables communes comprenant des totaux de population inconnus. Tous les échantillons peuvent être recueillis selon un plan d'échantillonnage complexe. On présume que $Y$ et $Z$ sont toutes deux des variables nominales; cependant, comme dans la section 2.5, les méthodes de pondération qui sont proposées peuvent aussi être appliquées à des variables $Y$ et $Z$ continues. Le but est d'estimer le tableau à double entrée composé par $Y$ et $Z$. Nous considérons deux estimateurs. L'un est fondé sur une stratification incomplète double (analogue au premier ensemble de contraintes décrit dans la section 2.2), et l'autre est fondé sur un mélange entre l'appariement statistique et le calage (analogue au troisième ensemble de contraintes mentionné dans la section 2.2).

### 3.1 Stratification double incomplète

Tout d'abord, on estime les totaux de population relatifs à $Y$ et $Z$ au moyen des premier et deuxième (grands) échantillons respectivement. Ces totaux de population sont estimés en deux phases. Dans la première phase, les deux grands échantillons sont pondérés en utilisant $X$ comme un ensemble de variables de contrôle. Cela implique que les deux grands échantillons soient pondérés de manière à ce qu'ils reproduisent les totaux de population connus de $X$, qui sont représentés par $t_x$. D'après ces poids, une estimation regroupée des totaux de population de $U$ donne

$$\hat{t}_u = \lambda \sum_{k \in n_1} w_{1k} u_k + (1 - \lambda) \sum_{k \in n_2} w_{2k} u_k,$$

où $w_{1k}$ et $w_{2k}$ représentent les poids de calage (de première phase) des premier et deuxième échantillons, et $\lambda \in [0, 1]$. Dans la deuxième phase, les deux échantillons sont pondérés à nouveau en utilisant simultanément $X$ et $U$ comme variables de contrôle. Soient $v_{1k}$ et $v_{2k}$ les poids de calage de cette deuxième phase. Les estimateurs pour les totaux de population de $Y$ et $Z$ qui sont obtenus peuvent être considérés comme des estimateurs de calage en deux phases (voir Renssen et Nieuwenbroek 1997, section 6). Ces estimateurs sont représentés respectivement par $\hat{t}_y$ et $\hat{t}_z$:

$$\hat{t}_y = \sum_{k \in n_1} v_{1k} y_k \text{ et } \hat{t}_z = \sum_{k \in n_2} v_{2k} z_k.$$

Nous remarquons que les deux estimateurs sont fondés sur un ensemble similaire de variables de contrôle. Si l'ensemble commun de variables est volumineux, on peut considérer l'utilisation d'un sous-ensemble plus petit pour pondérer les deux échantillons. En général, le sous-ensemble servant à pondérer le premier échantillon peut différer du sous-ensemble utilisé pour pondérer le deuxième échantillon. Cependant, nous présupposons dorénavant que les deux (grands) échantillons sont pondérés d'après le même ensemble de variables de contrôle.

Le tableau à double entrée composé de $Y$ et $Z$ peut être estimé par la pondération du troisième échantillon (petit échantillon), en utilisant simultanément $Y$ et $Z$ comme variables de contrôle, c'est-à-dire

$$\hat{T} = \sum_{k \in n_3} w_{3k} (y_k z_k^t),$$

où les poids de calage $w_{3k}$ satisfont aux contraintes

$$\sum_{k \in n_3} w_{3k} y_k = \hat{t}_y \text{ et } \sum_{k \in n_3} w_{3k} z_k = \hat{t}_z.$$

Il s'agit d'une stratification incomplète double, où les totaux de population inconnus de $Y$ et $Z$ sont remplacés par leurs estimations. Ces ensembles de contraintes permettent d'obtenir des estimations justes des dénombrements marginaux du tableau $YZ$, si les variables communes $C$ présentent une forte corrélation avec $Y$ et $Z$.

### 3.2 Stratification double synthétique

Dans la présente section, nous considérons un autre estimateur pour le tableau $YZ$ qui a fait appel également aux (grands) échantillons comme sources d'informations auxiliaires. Cependant, plutôt que d'utiliser des dénombrements estimés marginaux comme informations auxiliaires, on utilise des dénombrements synthétiques estimés. Soit $B$ le coefficient de régression de population relatif à $Y$ et $C$, qui est estimé d'après le premier (grand) échantillon:

$$\hat{B} = \left( \sum_{k \in n_1} v_{1k} c_k c_k^t \right)^{-1} \left( \sum_{k \in n_1} v_{1k} c_k y_k^t \right).$$

De façon analogue, soit $A$ le coefficient de régression de population relatif à $Z$ et $C$, qui est estimé d'après le deuxième (grand) échantillon:

$$\hat{A} = \left( \sum_{k \in n_2} v_{2k} c_k c_k^t \right)^{-1} \left( \sum_{k \in n_2} v_{2k} c_k z_k^t \right).$$

Il est à noter que ces coefficients de régression estimés sont fondés sur les poids de calage de deuxième phase, plutôt que sur les poids d'inclusion. S'il y a une constante $a$ qui fait que $a^t c_k = 1$ dans le cas de toutes les personnes $k$, nous avons alors toujours $l^t \hat{B}^t c_k = l^t \hat{A}^t c_k = 1$ dans le cas de toutes les personnes $k$. Maintenant, inspirés par la décomposition donnée par (3), c'est-à-dire,

$$\sum_{k=1}^{N} y_k z_k^t = B^t \sum_{k=1}^{N} (c_k c_k^t) A + \sum_{k=1}^{N} (y_k - B^t c_k)(z_k - A^t c_k)^t,$$

nous proposons l'estimation du tableau à double entrée en deux étapes. Dans la première étape, le premier terme du côté droit est estimé en remplaçant les coefficients de régression de population $B$ et $A$ par leurs estimations $\hat{B}$ et $\hat{A}$. En outre, nous proposons d'estimer $\sum_c = \sum_{k=1}^{N} c_k c_k^t$ d'après l'estimation combinée

$$\hat{\sum}_c = \gamma \sum_{k \in n_1} v_{1k} (c_k c_k^t) + (1 - \gamma) \sum_{k \in n_2} v_{2k} (c_k c_k^t),$$

où $v_{1k}$ et $v_{2k}$ représentent les poids (de deuxième phase) des premier et deuxième échantillons, et $\gamma \in [0, 1]$. Finalement, le premier terme est estimé d'après $\hat{B}^t \hat{\sum}_c \hat{A}$. Jusqu'à maintenant, nous n'avons pas utilisé le petit échantillon (troisième). Si on veut, les estimations de $B$, $A$ et $\sum_c$ peuvent être légèrement améliorées en utilisant également l'échantillon de petite taille.

Dans la deuxième étape, on estime le tableau complet à double entrée composé de $Y$ et $Z$ en pondérant le petit échantillon (troisième) d'après l'estimateur de calage assujetti au troisième ensemble de contraintes (voir section 2.2), où $B$, $A$ et $\sum_c$ sont remplacés par leurs estimations $\hat{B}$, $\hat{A}$ et $\hat{\sum}_c$. L'estimateur qui est obtenu correspond à

$$\sum_{k=1}^{n_3} w_3 (y_k z_k^t) = \hat{B}^t \hat{\sum}_c \hat{A} + \sum_{k=1}^{n_3} w_{3k} (y_k - \hat{B}^t c_k)(z_k - \hat{A}^t c_k)^t. \quad (8)$$

Le premier terme du côté droit est une estimation pour le tableau synthétique à double entrée. Cette estimation est presque non biaisée dans le cas du tableau $YZ$ si l'hypothèse d'indépendance conditionnelle est confirmée. Nous signalons que ce type d'estimateur est obtenu essentiellement en appliquant la méthode d'appariement statistique avec contraintes (voir, par ex., Barr et Turner 1980, Rodgers 1984 ou Rubin 1986). Le deuxième terme est un terme de correction utilisé pour obtenir une estimation presque non biaisée pour le tableau $YZ$ sans l'hypothèse précitée. S'il existe une constante $a$ qui fait que $a^t c_k = 1$ dans le cas de tous les éléments échantillonnés, nous obtenons alors, en prémultipliant les deux côtés de (8) par $l^t$, l'estimateur suivant pour le total de population de $Z$:

$$\sum_{k \in n_3} w_{3k} z_k^t = \left( \gamma \sum_{k \in n_1} v_{1k} c_k^t + (1 - \gamma) \sum_{k \in n_2} v_{2k} c_k^t \right) \hat{A} = \left( \sum_{k \in n_2} v_{2k} c_k^t \right) \hat{A} = a^t \left( \sum_{k \in n_2} v_{2k} c_k c_k^t \right) \hat{A} = \hat{t}_z^t.$$

De la même façon, en postmultipliant les deux côtés par $l$, nous obtenons un estimateur pour le total de population de $Y$:

$$\sum_{k \in n_3} w_{3k} y_k = \hat{B}^t \left( \gamma \sum_{k \in n_1} v_{1k} c_k + (1 - \gamma) \sum_{k \in n_2} v_{2k} c_k \right) = \hat{B}^t \left( \sum_{k \in n_1} v_{1k} c_k \right) = \hat{B}^t \left( \sum_{k \in n_1} v_{1k} c_k c_k^t \right) a = \hat{t}_y.$$

Il s'ensuit que les dénombrements marginaux de cellule du tableau estimé à double entrée sont les estimateurs à deux phases pour les totaux de population de $Y$ et $Z$ tels que nous les avons définis dans la section 3.1.

### 3.3 Une étude de simulation: intégration d'enquêtes-ménages

Dans la présente sous-section, nous allons comparer les méthodes de pondération que sont la stratification double incomplète (présentée en 3.1) et la stratification double synthétique (présentée en 3.2), à l'aide d'une étude de simulation. Nous utilisons un ensemble de données qui provient d'une étude pilote de la Dutch Household Survey on Living Conditions; voir van Tuinen (1995).

L'ensemble de données comprend 1 085 dossiers dont on a observé les variables suivantes: âge (six catégories: 15-24, 25-34, 35-44, 45-54, 55-64, 65 et plus), sexe (deux catégories: homme, femme), possession d'une maison (deux catégories: oui, non), occupation (cinq catégories: emploi, entretien ménager, études, travail bénévole, autre) et santé (deux catégories: oui, non). Aux fins de l'étude de simulation, cet ensemble de données est considéré comme une population finie. On présuppose que les totaux de population relatifs à l'âge et au sexe sont connus.

Afin de simuler les méthodes de pondération, nous avons exécuté un algorithme de Monte Carlo. Nous avons notamment produit 500 échantillons, de manière indépendante les uns des autres, d'après un plan d'échantillonnage à deux phases. Dans la première phase, on recueille un échantillon aléatoire simple d'une taille de 20 500 personnes, avec remplacement. Dans cet échantillon, on observe l'âge, le sexe et la possession d'une maison. Dans la deuxième phase, l'échantillon recueilli lors de la première phase est divisé de façon aléatoire en deux grands sous-échantillons d'une taille de 10 000 personnes chacun et un petit sous-échantillon de 500 personnes; dans l'un des grands échantillon, on observe l'occupation (représentée par $Y$), dans l'autre, on observe la santé (représentée par $Z$), tandis que dans le petit échantillon, on observe à la fois l'occupation et la santé. Lors de chaque observation, nous avons estimé le tableau à double entrée composé de $Y$ et $Z$ d'après quatre méthodes de pondération que nous allons présenter ci-après.

L'échantillon de la première phase est pondéré au moyen d'un croisement du sexe et de l'âge qui est utilisé comme variable de contrôle. Il s'agit là uniquement d'une poststratification avec douze poststrates. D'après ces poids, les totaux de population peuvent être estimés dans le cas de toutes les variables observées dans l'échantillon de la première phase et des croisements effectués entre ces variables. En particulier, nous pouvons reproduire les totaux de population dans le cas de combinaisons de l'âge et du sexe, et obtenir des totaux de population estimés dans le cas de croisements de variables relatives à l'âge, au sexe et à la possession d'une maison. Maintenant, nous distinguons deux ensembles de variables communes pour pondérer les grands sous-échantillons, ainsi que pour obtenir une estimation pour le tableau synthétique à double entrée composé de $Y$ et $Z$. Le premier ensemble est une combinaison de l'âge et du sexe (12 catégories) et le deuxième une combinaison de l'âge, du sexe et de la possession d'une maison (24 catégories). Dans le cas de chaque simulation, cela donne deux estimations différentes pour les dénombrements marginaux, c'est-à-dire deux estimations différentes pour les totaux de population de $Y$ et $Z$ (il est à noter que les deux estimations sont fondées sur une poststratification) et deux estimations différentes pour le tableau synthétique à double entrée. Afin de pondérer le petit sous-échantillon, nous faisons la distinction entre la méthode de pondération fondée sur la stratification double incomplète et la méthode de pondération fondée sur la stratification synthétique double. Étant donné qu'on utilise deux ensemble différents de variables communes pour pondérer les grands sous-échantillons ainsi que pour effectuer l'appariement statistique, nous obtenons quatre ensemble de poids de calage pour chaque cycle de simulation par rapport au sous-échantillon de petite taille, qui donne à son tour, dans le cas de chaque cycle de simulation, quatre tableaux estimés à double entrée différents composés de $Y$ et $Z$. Afin de faciliter les calculs, nous avons utilisé la mesure de distance quadratique dans l'estimation de calage, ce qui implique que chaque cellule estimée correspond à une estimation de régression générale. Enfin, nous avons pris les moyennes et les variances de ces tableaux à double entrée au cours des 500 simulations. Les résultats sont indiqués dans les tableaux 5 à 8.

**Tableau 5**
Stratification double incomplète combinée au premier ensemble de variables communes

| | 1 | 2 | 3 | 4 | 5 | total |
|---|---|---|---|---|---|---|
| oui | $447_{(96)}$ | $232_{(97)}$ | $89_{(28)}$ | $25_{(21)}$ | $59_{(49)}$ | $852_{(17)}$ |
| non | $61_{(79)}$ | $104_{(90)}$ | $11_{(21)}$ | $11_{(19)}$ | $46_{(46)}$ | $233_{(17)}$ |
| total | $508_{(23)}$ | $336_{(19)}$ | $100_{(8)}$ | $36_{(3)}$ | $105_{(10)}$ | 1 085 |

**Tableau 6**
Stratification double incomplète combinée au deuxième ensemble de variables communes

| | 1 | 2 | 3 | 4 | 5 | total |
|---|---|---|---|---|---|---|
| oui | $447_{(96)}$ | $232_{(97)}$ | $89_{(28)}$ | $25_{(21)}$ | $59_{(49)}$ | $852_{(17)}$ |
| non | $61_{(79)}$ | $104_{(90)}$ | $11_{(21)}$ | $11_{(19)}$ | $46_{(46)}$ | $233_{(17)}$ |
| total | $508_{(23)}$ | $336_{(19)}$ | $100_{(8)}$ | $36_{(3)}$ | $105_{(9)}$ | 1 085 |

**Tableau 7**
Stratification double synthétique combinée au premier ensemble de variables communes

| | 1 | 2 | 3 | 4 | 5 | total |
|---|---|---|---|---|---|---|
| oui | $447_{(75)}$ | $231_{(74)}$ | $89_{(17)}$ | $25_{(20)}$ | $59_{(42)}$ | $851_{(17)}$ |
| non | $61_{(58)}$ | $105_{(65)}$ | $11_{(12)}$ | $11_{(19)}$ | $46_{(38)}$ | $234_{(17)}$ |
| total | $508_{(23)}$ | $336_{(19)}$ | $100_{(8)}$ | $36_{(3)}$ | $105_{(10)}$ | 1 085 |

**Tableau 8**
Stratification synthétique double combinée au deuxième ensemble de variables communes

| | 1 | 2 | 3 | 4 | 5 | total |
|---|---|---|---|---|---|---|
| oui | $447_{(70)}$ | $231_{(70)}$ | $89_{(16)}$ | $25_{(18)}$ | $59_{(40)}$ | $851_{(17)}$ |
| non | $61_{(52)}$ | $105_{(60)}$ | $11_{(11)}$ | $11_{(16)}$ | $46_{(37)}$ | $234_{(17)}$ |
| total | $508_{(23)}$ | $336_{(19)}$ | $100_{(8)}$ | $36_{(3)}$ | $105_{(9)}$ | 1 085 |

Les moyennes observées au cours des 500 simulations sont presque identiques dans le cas des quatre types d'estimateur, comme le montrent ces tableaux. Il est à noter que les dénombrements de cellule indiqués sont arrondis. Nous avons également calculé le tableau *YZ* réel à partir de la population finie. Les dénombrement réels correspondent exactement aux moyennes qui sont indiquées dans le tableau 5 (ou 6). Dans le cas de cette simulation, nous en arrivons à la conclusion que tous les estimateurs comportent un biais très faible.

Les variances observées au cours des 500 simulations sont indiquées entre parenthèses. Les variances relatives aux dénombrements marginaux estimés des tableaux 5 et 7 coïncident parce que ces estimations sont fondées sur le même estimateur. Pour la même raison, les variances relatives aux dénombrements marginaux estimés des tableaux 6 et 8 coïncident également. Il est à noter que les variances relatives aux dénombrements marginaux estimés des tableaux 6 et 8 sont légèrement plus faibles que les variances relatives aux dénombrements marginaux estimés des tableaux 5 et 7, en raison du plus grand ensemble de variables communes. Cependant, dans le cas de la plupart des dénombrements marginaux estimés, cette réduction de la variance peut être considérée comme négligeable.

Les tableaux 5 et 6 donnent des variances identiques en ce qui a trait à tous les dénombrements estimés. Les variances dans le cas de la plupart des dénombrements estimés du tableau 7 sont clairement plus faibles que celles indiquées dans les tableaux 5 et 6. Dans le tableau 8, cette réduction de la variance est encore plus importante. Dans le cas de cet exemple, nous en arrivons à la conclusion que l'utilisation de l'ensemble plus grand de variables communes avec la première méthode de pondération réduit légèrement les taux de variance relatifs aux dénombrements marginaux estimés, mais n'a pas d'incidence sur les taux de variance des dénombrements de cellule estimés. Évidemment, l'utilisation de l'ensemble plus grand de variables communes avec la deuxième méthode de pondération réduit également légèrement les taux de variance des dénombrements de cellule marginaux. Enfin, compte tenu de l'ensemble de variables communes, la méthode de pondération fondée sur un appariement synthétique donne des variances plus faibles dans le cas des dénombrements de cellule estimés que la méthode fondée sur la stratification double incomplète.

### 3.4 Imputation de valeurs tirées d'un des grands échantillons dans l'autre grand échantillon

Au moyen de deux échantillons de grande taille et du petit échantillon, on peut créer un échantillon synthétique dans lequel les valeurs réelles de $Y$ et les valeurs prédites de $Z$, ou les valeurs prédites de $Y$ et les valeurs réelles de $Z$ sont enregistrées simultanément. Nous définissons des prédictions pour les valeurs de $Y$ et $Z$ de manière analogue à (7), à savoir

$$\hat{y}_k = \hat{B}'c_k + \tilde{\beta}_2' \ (z_k - \hat{A}'c_k), k = 1, ..., n_2, \qquad (9)$$

et

$$\hat{z}_k = \hat{A}'c_k + \tilde{\alpha}_2' \ (y_k - \hat{B}'c_k), k = 1, ..., n_1, \qquad (10)$$

où

$$\tilde{\beta}_2 = \left[\sum_{k=1}^{n_2} v_{2k}(z_k - \hat{A}'c_k)(z_k - \hat{A}'c_k)'\right]^{-1} \times \left[\sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)'\right],$$

et

$$\tilde{\alpha}_2 = \left[\sum_{k=1}^{n_1} v_{1k}(y_k - \hat{B}'c_k)(y_k - \hat{B}'c_k)'\right]^{-1} \times \left[\sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)'\right].$$

Dans le cas de chaque $(c_k, y_k)$, les valeurs de $Z$ peuvent être imputées dans le premier grand échantillon au moyen de (10), $k = 1, ..., n_1$, et de façon analogue, dans le cas de chaque $(c_k, z_k)$, les valeurs de $Y$ peuvent être imputées dans le deuxième grand échantillon au moyen de (9), $k = 1, ..., n_2$. D'après ces valeurs imputées, nous pouvons définir les estimations suivantes pour le tableau à double entrée composé de $Y$ et Z:

$$\sum_{k=1}^{n_1} v_{1k}\, y_k \hat{z}_k' = \hat{B}' \sum_{k=1}^{n_1} v_{1k}\, c_k\, c_k'\, \hat{A} + \sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)' \qquad (11)$$

et

$$\sum_{k=1}^{n_2} v_{2k}\, \hat{y}_k z_k' = \hat{B}' \sum_{k=1}^{n_2} v_{2k}\, c_k\, c_k'\, \hat{A} + \sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)'. \qquad (12)$$

Une des estimations est fondée sur le premier échantillon synthétique, l'autre sur le deuxième échantillon synthétique. En réunissant les échantillons synthétiques, on obtient un échantillon synthétique combiné de taille $n_1 + n_2$, à partir duquel on peut établir une estimation combinée pour le tableau à double entrée. Cette estimation combinée montre une ressemblance plus étroite avec (8). Il est à noter que

si $C$ et $Z$ présentent entre elles une corrélation parfaite, le côté gauche de (11) se réduit à $\sum_{k=1}^{n_1} v_{1k} y_k z_k^t$, c'est-à-dire que notre tableau estimé à double entrée correspond à un tableau estimé à double entrée pondéré fondé sur le premier échantillon, comme si les valeurs réelles de $Z$ étaient imputées dans cet échantillon. De manière analogue, si $C$ et $Y$ présentent entre elles une corrélation parfaire, (12) se réduit à $\sum_{k=1}^{n_2} v_{2k} y_k z_k^t$.

Un cas particulier important qui doit être pris en considération est celui où $c$ est une variable nominale. Dans ce cas, les égalités suivantes sont vérifiées:

$$\sum_{k \in n_1} v_{1k}(c_k c_k^t) = \sum_{k \in n_2} v_{2k}(c_k c_k^t) = \text{diag}\begin{pmatrix} t_x \\ \hat{t}_u \end{pmatrix},$$

ainsi (11) et (12) coïncident. En outre, nous avons, dans le cas d'une variable $c$ nominale:

$$\sum_{k \in n_1} v_{1k} c_k \hat{z}_k^t = \sum_{k \in n_2} v_{2k} c_k z_k^t$$

et

$$\sum_{k \in n_2} v_{2k} \hat{y}_k c_k^t = \sum_{k \in n_1} v_{1k} y_k c_k^t.$$

Évidemment, si $c$ est nominale, il suffit de créer un échantillon synthétique qui est fondé soit sur le premier échantillon synthétique, soit sur le deuxième. Dans un cas comme dans l'autre, les estimations à pondération pour les tableaux $CZ$, $CY$ et $YZ$ peuvent être reconstituées. Enfin, nous signalons que les valeurs imputées dans tous les échantillons synthétiques peuvent être irréalistes. Comme nous l'avons mentionné dans la section 2.4, les prédictions calculées peuvent être remplacées par des valeurs réelles d'après un quelconque algorithme.

## 4. RÉCAPITULATION

Dans le présent article nous avons décrit une méthode de pondération qui permet de combiner des informations provenant de différentes enquêtes sur échantillon. La charnière entre ces enquêtes est un ensemble de variables communes (voir figure 1). Nous prétendons que ces échantillons devraient être pondérés selon une structure séquentielle. Dans un premier temps, nous avons pondéré les deux échantillons de grande taille en utilisant les valeurs de $X$ comme variables de contrôle. D'après ces échantillons pondérés, nous avons pu obtenir une estimation combinée du total de population de $U$. Ensuite, les deux échantillons de grande taille ont été pondérés à nouveau en utilisant simultanément $X$ et $U$ comme variables de contrôle. Cela a donné une estimation du total de population de $Y$ et Z.

À l'aide de méthodes d'appariement statistique, et en utilisant $X$ et $U$ comme variables communes, nous pouvons également obtenir une estimation dans le cas d'un tableau synthétique à double entrée composé de $Y$ et $Z$. Finalement, l'échantillon de petite taille a été pondéré d'après deux ensembles de variables de contrôle différents. Le premier ensemble de variables de contrôle correspondait aux totaux de population de $Y$ et $Z$, tandis que le deuxième ensemble de variables de contrôle correspondait au tableau synthétique à double entrée. L'utilisation du premier ensemble de variables de contrôle s'apparente fortement à la stratification double incomplète. Le cadre théorique qui est nécessaire pour créer la deuxième méthode de pondération a été présenté tout au long du présent article. À l'aide des deux méthodes de pondération, on peut estimer le tableau $YZ$ (on présuppose tacitement que $Y$ et $Z$ sont des variables nominales). Les dénombrements marginaux du tableau $YZ$ relatifs à la première méthode de pondération correspondent, par définition des équations de calage, aux totaux estimés de population de $Y$ (qui est fondée sur le premier grand échantillon) et de $Z$ (qui est fondée sur le deuxième grand échantillon). Nous avons montré que cette propriété de cohérence est confirmée dans le cas de la deuxième méthode de pondération. Nous avons également effectué une étude numérique pour évaluer le rendement des méthodes de pondération en ce qui a trait aux dénombrements de cellule. Nous avons constaté à cet égard que les deux méthodes de pondération donnent des tableaux à double entrée qui sont presque exempts de biais de plan d'échantillonnage. Les variances simulées (de plan d'échantillonnage) relatives à la deuxième méthode de pondération semblaient être plus faibles que les variances (de plan d'échantillonnage) correspondantes relatives à la première méthode de pondération, en ce qui a trait à tous les dénombrements de cellule estimés. En principe, on a présupposé que les variables $Y$ et $Z$ étaient des variables nominales; cependant, nous avons précisé que les idées qui sont présentée peuvent être appliquées également dans le cas de variables $Y$ et $Z$ continues, ou dans le cas d'une variable $Y$ continue et d'une variable $Z$ nominale.

## REMERCIEMENTS

L'auteur remercie Peter Kooiman, Nico Nieuwenbroek, et Ger Slootbeek pour leur lecture attentive du présent article et leurs judicieuses observations. L'auteur remercie également deux examinateurs anonymes et un rédacteur associé pour leur précieux conseils visant à améliorer l'article. Les opinions exprimées dans celui-ci sont celles de l'auteur et ne reflètent pas nécessairement la politique de Statistics Netherlands.

## BIBLIOGRAPHIE

ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.

BAKKER, B.F.M., et WINKELS, J.W. (1998). Why integration of household surveys? - Why POLS?. *Netherlands Official Statistics*, 13, 5-7.

BARR, R.S., et TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Survey. Technical report, U.S. Department of the Treasury, Office of Tax Analysis.

BETHLEHEM, J.G., et KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

COPELAND, K.R., PEITZMEIER, F.K., et HOY, C.E. (1987). Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population. *Techniques d'enquête*, 13, 183-191.

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

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

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

HOFMANS, M.G. (1998). Innovative weighting in POLS. Making use of core questions. *Netherlands Official Statistics*, 13, 12-15.

LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.

MADDALA, G.S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

OH, H.L., et SCHEUREN, F. (1987). Variante de la méthode itérative du quotient. *Techniques d'enquête*, 13, 221-232.

PAASS, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. Dans *Microanalytic Simulation Models to Support Social and Financial Policy*. Amsterdam: Elsevier Science.

RAGHUNATHAN, T.E., et GRIZZLE, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.

RENSSEN, R.H., et NIEUWENBROEK, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.

RODGERS, W.L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, 2, 91-102.

RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4, 87-94.

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

SEARLE, S.R. (1971). *Linear Models*. New York: John Wiley & Sons.

SEBER, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons.

SINGH, A.C., MANTEL, H.J., KINACK, M.D, et ROWE, R. (1993). Appariement statistique: l'utilisation d'information supplémentaire comme solution de remplacement à l'hypothèse d'indépendance conditionnelle. *Techniques d'enquête*, 19, 67-89.

TUINEN VAN, H.K. (1995). Social indicators, social surveys and integration of social statistics. *Statistical Journal of the United Nations ECE*, 12, 379-394.

WINKELS, J.W., et EVERAERS, P.C.J. (1998). Design of an integrated survey in the Netherlands. The case POLS. *Netherlands Official Statistics*, 13, 6-11.

ZIESCHANG, K.D. (1990). A generalized least squares weighting system for the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 201-208
Statistique Canada



# Échantillonnage en deux cycles: Estimation de la population totale

RAGHUNATH ARNAB[1]

RÉSUMÉ

Deux stratégies d'échantillonnage sont proposées pour estimer les chiffres d'une population finie au cycle le plus récent, à partir des échantillons prélevés sur deux cycles, au moyen de plans d'échantillonnage à probabilité variable. Nous utilisons les données recueillies au premier cycle sur une variable à l'étude comme mesure de la taille et comme variable de stratification pour la sélection de l'échantillon apparié, au deuxième cycle. L'efficacité relative des stratégies proposées est comparée ici à d'autres méthodes appropriées.

MOTS CLÉS: Estimateur composite; échantillon apparié; plan d'échantillonnage; stratégie d'échantillonnage; plan d'échantillonnage à probabilité variable.

## 1. INTRODUCTION

Il arrive très souvent qu'une même population fasse l'objet d'une enquête menée à intervalles réguliers, dans le but d'estimer les mêmes caractéristiques de la population en regard des changements qui surviennent dans le temps. Ainsi, un grand nombre de pays recueillent des données sur une base annuelle ou trimestrielle pour estimer, par exemple, le nombre total de personnes en chômage, de personnes infectées par le VIH, d'immigrants, *etc.* Dans le présent article, nous examinons une population finie $U = (U_1, ..., U_i, ..., U_N)$ composée de $N$ unités identifiables, qui est censée être échantillonnée à deux reprises pour estimer la population totale d'une variable à l'étude au deuxième cycle (actuellement). Dans le cas d'un échantillonnage successif, les données recueillies au cycle précédent (le premier) sont utilisées pour élaborer une stratégie efficace qui tienne compte des coûts et qui produise un estimateur efficace de l'ensemble de la population au cycle présent. De nombreux auteurs ont traité de ces méthodes. Singh (1967), de même que Avadhani et Sukhatme (1970), ont utilisé l'information recueillie au premier cycle comme mesure de la taille, pour sélectionner l'échantillon apparié au deuxième cycle; de son côté, Arnab (1991) a utilisé cette information à la fois comme variable de stratification et mesure de la taille pour la sélection de l'échantillon au deuxième cycle. Plus récemment, Prasad et Graham (1994) ont modifié les stratégies d'échantillonnage de Raj (1965) et de Chotai (1974), en utilisant l'information obtenue au premier cycle comme mesure de la taille pour la sélection de l'échantillon apparié au deuxième cycle; ils ont constaté, de façon empirique, qu'une de leurs stratégies donnait de meilleurs résultats que celle de Chotai (1974). Dans cet article, deux stratégies différentes sont proposées. L'une d'elle utilise l'information recueillie au premier cycle comme mesure de la taille, alors que l'autre utilise cette information à la fois comme mesure de la taille et comme variable de stratification pour la sélection de l'échantillon apparié au deuxième cycle. Nous démontrons ici qu'une des stratégies proposées est meilleure que celle de Prasad et Graham (1994); il nous est toutefois impossible de tirer de conclusion théorique définie en ce qui concerne la deuxième stratégie. Des données empiriques montrent cependant que cette dernière stratégie est plus efficace que celle décrite par Prasad et Graham (1994) et aussi que la première stratégie proposée, du fait qu'elle utilise les valeurs du premier cycle pour tous les stades possibles, c'est-à-dire stratification, estimation et sélection de l'échantillon apparié au deuxième cycle.

Les méthodes générales de sélection et d'estimation aux deux cycles sont décrites ci-après.

### 1.1 Plans d'échantillonnage

Lors du premier cycle d'enquête, un échantillon $s_1$, de taille $n$, est choisi au moyen d'un plan d'échantillonnage approprié, disons $P_1$, et les données $y_{1i}$ $i\varepsilon s_1$ sont obtenues, où $y_{1i}$ ($y_{2i}$) est la valeur de la variable aléatoire à l'étude, $y$, pour la $i$-ième unité au premier (deuxième) cycle. Au deuxième cycle, un échantillon apparié (sous-échantillon) $s_m$ de taille $m$ ($= n\lambda$, est censé être un nombre entier, $0 \leq \lambda \leq 1$) est choisi à partir de $s_1$, par un plan d'échantillonnage approprié $P_m$, et est complété par un échantillon non apparié $s_u$ de taille $u$ ($= n\mu = n - m, \mu = 1 - \lambda$), lequel est choisi, soit parmi l'ensemble de la population $U$, soit parmi $U/s_1$, la série d'unités qui n'ont pas été sélectionnées au premier cycle par le plan d'échantillonnage $P_u$; l'information $y_{2i}$ ($i\varepsilon s_m$, $i\varepsilon s_u$) au deuxième cycle est ainsi obtenue. Bien sûr, on s'attend à ce que le coût de l'enquête pour les unités de l'échantillon apparié soit beaucoup moins élevé que celui pour les unités non appariées; cependant, pour simplifier l'étude, nous présumons ici que le coût d'enquête est le même pour toutes les unités, au deuxième cycle.

[1] Raghunath Arnab, Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban - 4000, South Africa.

### 1.2 Méthode d'estimation

À partir des données $y_{1i}$, $i\varepsilon s_1$, et $y_{2i}$, $i\varepsilon s_m$, obtenues de l'échantillon initial $s_1$, et de l'échantillon apparié $s_m$, un estimateur sans biais $\hat{Y}_{2m}$ pour $Y_2$ – la population totale au deuxième cycle – est obtenu en utilisant $y_{1i}$, $i\varepsilon s_1$ comme étant de l'information auxiliaire. Par conséquent, $\hat{Y}_{2m}$ désigne habituellement un estimateur de la différence, un estimateur par quotient ou un estimateur de régression. À partir de l'échantillon non apparié $s_u$, un estimateur sans biais $\hat{Y}_{2u}$ est également construit pour $Y_2$. Enfin, un estimateur composite – une combinaison de $\hat{Y}_{2m}$ et $\hat{Y}_{2u}$ – est obtenu en utilisant un poids approprié de $\varphi$ $(0 \le \varphi \le 1)$

$$\hat{Y}_2 = \varphi\, \hat{Y}_{2m} + (1 - \varphi)\, \hat{Y}_{2u}. \qquad (1)$$

On obtient la valeur optimale de $\varphi = \varphi(\lambda)$ en réduisant au minimum la valeur de $V(\hat{Y}_2)$, qui désigne la variance de $\hat{Y}_2$ par rapport à $\varphi$ pour une valeur donnée de $m$ (*i.e.*, $\lambda$). Les expressions pour $\varphi(\lambda)$ et $V(\hat{Y}_2 I\lambda)$, la variance de $\hat{Y}_2$ lorsque $\varphi = \varphi(\lambda)$, sont calculées comme suit, lorsque $\hat{Y}_{2m}$ et $\hat{Y}_{2u}$ sont indépendantes:

$$\varphi(\lambda) = (1/V_m)\,[1/V_m + 1/V_u]^{-1},$$

$$V(\hat{Y}_2 I\lambda) = [1/V_m + 1/V_u]^{-1},$$

où $V_m$ et $V_u$ sont les variances de $\hat{Y}_{2m}$ et $\hat{Y}_{2u}$ respectivement. La proportion optimale de l'échantillon apparié $\lambda = \lambda_0$ est déterminée en réduisant au minimum la valeur de $V(\hat{Y}_2 I\lambda)$ par rapport à $\lambda$. Enfin, lorsqu'on introduit $\lambda = \lambda_0$ dans l'expression $V(\hat{Y}_2 I\lambda)$, on obtient la variance minimale de $\hat{Y}_2$, représentée par $V_{\min}(\hat{Y}_2) = V(\hat{Y}_2 I\lambda_0)$. Notre objectif est de trouver une stratégie appropriée, qui soit une combinaison de $P = (P_1, P_m, P_u)$ et $\hat{Y}_2$ et qui permette de maintenir la valeur de $V_{\min}(\hat{Y}_2)$ à un minimum.

### 1.3 Quelques stratégies d'échantillonnage

#### 1.3.1 Avadhani et Sukhatme (1970)

Au premier cycle, l'échantillon initial $s_1$ de taille $n$ a été sélectionné par échantillonnage aléatoire simple sans remise (ÉASSR), en présumant qu'aucune information auxiliaire n'était disponible avant l'enquête. Au deuxième cycle, l'échantillon apparié $s_m$ de taille $m$ a été prélevé de $s_1$, par la méthode de Rao, Hartley et Cochran (ou méthode de RHC 1962) en utilisant $y_{1i}$ comme mesure de la taille de la $i$-ième unité $i\varepsilon s_1$ en supposant que $y_{1i}$ est positif. Selon le plan d'échantillonnage de RHC, les $n$ unités sélectionnées de $s_1$ sont réparties de façon aléatoire en $m$ groupes, chacun de taille $n/m$ que l'on présume être un nombre entier. Dans chaque groupe sélectionné, une unité est choisie de façon indépendante, avec une probabilité proportionnelle à la mesure de taille. Par conséquent, si la $i$-ième unité, $U_i$, appartient au $j$-ième groupe $G_j$ $(j = 1, ..., m)$, alors la probabilité de sélectionner $U_i$ est égale à $q_i^*(i\varepsilon s_1) = y_{1i}/\sum_{i\varepsilon s_1} y_{1j}$. L'échantillon non apparié $s_u$ a été prélevé de $U/s_1$ par ÉASSR.

#### 1.3.2 Chotai (1974)

Au premier cycle, l'échantillon initial $s_1$ de taille $n$ a été sélectionné par le plan d'échantillonnage de RHC décrit précédemment (en présumant que $N/n$ est un nombre entier), avec une probabilité proportionnelle à $z_i$, la mesure de la taille de la $i$-ième unité que l'on présume être positive et connue pour chaque $i\varepsilon U$. Supposons que $\Delta_j = \sum_{k\varepsilon G_j} p_k$, la somme des valeurs de $p_k$ $(= z_k/Z, Z = \sum_{i\varepsilon U} z_i)$ qui appartiennent au groupe aléatoire $G_j$ $(j = 1, ..., n)$ formé en sélectionnant l'échantillon $s_1$ par la méthode de RHC. L'échantillon apparié $s_m$ a été prélevé de $s_1$ par le plan de RHC, en utilisant la mesure de taille normée $\Delta_i$ pour la $i$-ième unité $i\varepsilon s_1$ $(\sum_{i\varepsilon s_1} \Delta_i = 1)$, en supposant que $n/m$ est un nombre entier. L'échantillon non apparié, $s_u$ a été sélectionné selon le plan d'échantillonnage de RHC, avec la mesure de taille normée $p_i$ pour la $i$-ième unité, en présumant là encore que $N/u$ est un entier. Supposons que $p_i^+$ $(P_i')$ = total des valeurs de $\Delta_i$ $(p_i)$ associées aux unités qui appartiennent au groupe aléatoire duquel la $i$-ième unité a été sélectionnée dans $s_m$ $(s_u)$ par la méthode d'échantillonnage de RHC avec $\sum_{i\varepsilon s_m} P_i^+ = 1$ $(\sum_{i\varepsilon s_u} P_i' = 1)$.

L'estimateur composite de $Y_2$ est représenté par

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

où

$$\hat{Y}_{2m} = \sum_{i\varepsilon s_m} (y_{2i}/p_i)\, P_i^+ -$$

$$\gamma\left[\sum_{i\varepsilon s_m} (y_{1i}/p_i)\, P_i^+ - \sum_{i\varepsilon s_1} (y_{1i}/p_i)\, \Delta_i\right];$$

$$\hat{Y}_{2u} = \sum_{i\varepsilon s_u} (y_{2i}/p_i) P_i' \qquad (2)$$

où $\gamma$ est une constante choisie avec soin de manière à réduire au minimum la variance de $\hat{Y}_{2m}$. Chotai (1974) a défini l'expression du minimum de la variance de $\hat{Y}_2$ comme étant

$$V_{\min}(\hat{Y}_2) = k\,[1 - f + \sqrt{(1 - \delta^{*2})}]\, \sigma_2^2/2 = V_c \text{ (disons)} \qquad (3)$$

où

$$k = N/\{n(N-1)\},\ f = n/N,$$

$$\sigma_t^2 = \sum_{i\varepsilon U} p_i (y_{ti}/p_i - Y_t)^2,\ t = 1, 2$$

$$Y_t = \sum_{i\varepsilon U} y_{ti},\ t = 1, 2$$

$$\delta^* = \sum_{i\varepsilon U} p_i (y_{2i}/p_i - Y_2)\ (y_{1i}/p_i - Y_1)/(\sigma_1\sigma_2). \qquad (4)$$

#### 1.3.3 Arnab (1991)

Arnab (1991) a proposé plusieurs stratégies selon lesquelles l'échantillon initial $s_1$ est construit par la

méthode de sélection avec probabilité proportionnelle à la taille avec remise (PPTAR), en utilisant la mesure de taille normée $p_i = z_i / Z$ pour la $i$-ième unité. À partir des valeurs pré-établies $y_{1i}$ $(i \varepsilon s_1)$ obtenues sur la base de certains critères, les $n$ unités de l'échantillon sont réparties en un nombre approprié de $L$ strates. Supposons que $s_{1h}$ représente l'échantillon de taille $n_h$, dans la $h$-ième strate ($s_1 = U_h s_{1h}$ et $\sum_h n_h = n$). Nous supposons ici que $n$ est suffisamment grand pour s'assurer que $n_h$ est positif pour chaque $h$ dans la pratique. Au deuxième cycle, les sous-échantillons $s_{mh}$ de taille $m_h$ ($= v_h n_h$, $v_h$ est une fraction prédéterminée et $m_h$ est censé être un nombre entier) sont sélectionnés de façon indépendante de $s_{1h}$, selon des plans d'échantillonnage appropriés utilisant $y_{1i}$, $i \varepsilon s_1$ pour la construction des échantillons appariés $s_{mh}$. L'échantillon non apparié $s_u$ est sélectionné par PPTAR, à partir de l'ensemble de la population $U$, en utilisant $z_i'$ comme mesure de taille.

### 1.3.4 Prasad et Graham (1994)

Ici, l'échantillon initial $s_1$ est sélectionné par un plan d'échantillonnage de RHC similaire à la méthode de Chotai (1974), en utilisant une mesure de taille normée $p_i = z_i / Z$ pour la $i$-ième unité. L'échantillon apparié $s_m$ est prélevé de $s_1$ par la méthode de RHC en utilisant $p_i^* = (y_{1i}\Delta_i / p_i) / \sum_{i \varepsilon s_1} (y_{1i}\Delta_i / p_i)$ pour la $i$-ième unité, $i \varepsilon s_1$ ; où $\Delta_i$ est la somme des valeurs de $p_j$ pour le groupe qui renferme la $i$-ième unité obtenue en sélectionnant $s_1$ par le plan d'échantillonnage de RHC. L'échantillon non apparié, $s_u$ a été prélevé de l'ensemble de la population $U$ par un plan d'échantillonnage de RHC similaire à celui proposé par Chotai (1974). Ici encore, on présume que $N/n$, $n/m$ et $N/u$ sont des nombres entiers. Prasad et Graham (1994) ont proposé l'estimateur composite suivant pour $Y_2$ :

$$\hat{Y}_2 = \varphi \hat{Y}_{2m} + (1 - \varphi) \hat{Y}_{2u}$$

où $\hat{Y}_{2m} = \sum_{i \varepsilon s_m} (y_{2i}^* / p_i^*) \tilde{P}_i$ ; $\hat{Y}_{2u} = \sum_{i \varepsilon s_u} (y_{2i} / p_i) P_i'$ ; $y_{2i}^* = y_{2i} \Delta_i / p_i$ ; $\tilde{P}_i (P_i')$ = total des valeurs de $p_i^*$ $(p_i)$ associées aux unités qui appartiennent au groupe aléatoire duquel la $i$-ième unité a été sélectionnée dans $s_m$ $(s_u)$. L'expression pour la variance minimum de $\hat{Y}_2$ est représentée par:

$$V_{\min}(\hat{Y}_2) = k(1 - f + \sqrt{\zeta})\, \sigma_2^2 / 2 = V_{PG} \text{ (disons)} \qquad (5)$$

où

$$\zeta = \sigma_3^2 / \sigma_2^2,\ \sigma_3^2 = \sum_{i \varepsilon U} q_i (y_{2i}/q_i - Y_2)^2,\ q_i = y_{1i} / Y_1; \qquad (6)$$

$k$, $f$, $\sigma_2^2$ et $Y_1$ sont définis en (4).

Dans l'expression de Prasad et Graham (1994) de $V_{\min}(\hat{Y}_2)$, le diviseur 2 a été omis et il s'agit de toute évidence d'une erreur typographique.

**Remarque 1.1**

Si l'on se rapporte aux stratégies décrites à la section 1.3, nous constatons que le plan de Avadhani et Sukhatme (1970) ne requiert pas d'information sur les mesures de taille pour l'ensemble de la base de sondage et qu'il est donc moins exigeant que les autres. Chotai (1974) a utilisé les mesures de taille initiales $p_i$ pour la sélection, mais les valeurs obtenues lors de la première enquête $y_{1i}$, $i \varepsilon s_1$ ont été utilisées pour l'estimation seulement. L'utilisation de cette information supplémentaire $p_i$ pour la sélection de l'échantillon initial $s_1$ rend la stratégie de Chotai (1974) plus efficace que celle de Avadhani et Suhkatme (1970). Cependant, pour utiliser l'estimateur optimal $\hat{Y}_2$ avec la stratégie de Avadhani et Sukhatme (1970), il faut estimer $\varphi$, le seul paramètre inconnu. Par contre, dans la stratégie de Chotai (1974), les deux paramètres $\varphi$ et $\gamma$ doivent être estimés pour pouvoir utiliser l'estimateur optimal $\hat{Y}_2$. Prasad et Graham (1994) ont utilisé ces deux variables pour la sélection de l'échantillon apparié (donc automatiquement aussi pour l'estimation) et ils ont démontré, de façon empirique, que leur stratégie donne de meilleurs résultats que celle de Chotai (1974). Pour une plus grande efficacité, on peut également utiliser la stratégie de Prasad et Graham (1994) dans la pratique, car $\hat{Y}_2$ ne suppose qu'un seul paramètre inconnu, $\varphi$. Il est à noter que Arnab (1991) a été le premier à introduire le principe de la stratification en utilisant $y_{1i}$, $i \varepsilon s_1$ comme variable de stratification. Cette technique doit toujours être utilisée en pratique, chaque fois que l'information nécessaire est disponible, en particulier pour la sélection de larges unités qui présentent de grandes différences au niveau de la taille, comme celles examinées dans les exemples numériques indiqués à la section 3. On s'attend à ce que la stratégie de Arnab (1991) soit plus efficace que les stratégies précédentes, car elle utilise les valeurs du premier cycle à la fois pour la stratification et l'estimation. L'estimateur optimal $\hat{Y}_2$ contient toutefois plusieurs paramètres inconnus (pour plus de détails, voir Arnab 1991) qui peuvent nuire à l'application de la stratégie, en particulier lorsque la taille de l'échantillon est insuffisante.

## 2. STRATÉGIES PROPOSÉES

Nous examinons ici deux stratégies d'échantillonnage qui sont des modifications des stratégies proposées respectivement par Prasad et Graham (1994) et par Arnab (1991).

### 2.1 Stratégie 1

Le plan d'échantillonnage pour cette stratégie est le même que celui examiné par Prasad et Graham (1994), et qui est décrit à la section 1.3.4. Ici, seul l'estimateur basé sur l'échantillon apparié $s_m$ a été modifié en introduisant la

mesure initiale de taille dans l'estimation. L'estimateur modifié proposé $\hat{Y}_{2m}^{*}$ et l'estimateur composite de $Y_2$ se définissent comme suit:

$$\hat{Y}_{2m}^{*} = \sum_{i\epsilon s_m} (y_{2i}^{*}/p_i^{*})\,\tilde{P}_i - \beta\left[\sum_{i\epsilon s_m} (z_i^{*}/p_i^{*})\,\tilde{P}_i - Z\right] =$$

$$\sum_{i\epsilon s_m} (r_i^{*}/p_i^{*})\,\tilde{P}_i + \beta Z$$

où $z_i^{*} = z_i\Delta_i/p_i, y_{2i}^{*} = y_{2i}\Delta_i/p_i, r_i^{*} = r_i\Delta_i/p_i, r_i = y_{2i} - \beta z_i$ et $\beta$ est une constante choisie avec soin de manière à réduire au minimum la variance de $\hat{Y}_{2m}^{*}$; $p_i^{*}, \tilde{P}_i$ et $\Delta_i$ correspondent aux descriptions indiquées à la section 1.3.4;

$$\hat{Y}_2 = \varphi\,\hat{Y}_{2m}^{*} + (1 - \varphi)\,\hat{Y}_{2u}$$

où $\hat{Y}_{2u}$ est obtenu en (2).

Si $E_1(V_1)$ désigne l'espérance mathématique non conditionnelle (variance) pour la sélection de l'échantillon $s_1$, et $E_2(V_2)$ désigne l'espérance mathématique conditionnelle (variance) par rapport à $s_m$ lorsque la valeur de $s$ est fixe, la variance de $\hat{Y}_{2m}^{*}$ pour une valeur donnée de $\beta$ correspond à

$$V(\hat{Y}_{2m}^{*} I\beta) = E_1 V_2(\hat{Y}_{2m}^{*} I\beta) + V_1 E_2(\hat{Y}_{2m}^{*} I\beta).$$

Selon Prasad et Graham (1994), nous obtenons

$$E_1 V_2(\hat{Y}_{2m}^{*} I\beta) = k_1 \sigma_3^{*2}(\beta)$$

et

$$V_1 E_2(\hat{Y}_{2m}^{*}) = k(1 - f)\,\sigma_2^2$$

où

$$k_1 = N(n - m)/\{nm(N - 1)\};$$

$$\sigma_3^{*2}(\beta) = \sum_{i\epsilon U} q_i (r_i/q_i - R)^2$$

$$= \sigma_3^2 + \beta^2\sigma_0^2 - 2\beta\sigma_0\sigma_3\delta;$$

$$R = \sum_{i\epsilon U} R_i = Y_2 - \beta Z, \delta = \sigma_{03}/(\sigma_0\sigma_3),$$

$$\sigma_0^2 = \sum_{i\epsilon U} q_i (z_i/q_i - Z)^2,$$

$$\sigma_{03} = \sum_{i\epsilon U} q_i (y_{2i}/q_i - Y_2)(z_i/q_i - Z) \qquad (7)$$

$\sigma_2^2, k$ et $\sigma_3^2, q_i$ sont obtenus respectivement de (4) et (6). La valeur optimale de $\beta$ qui réduit au minimum la valeur de $V(\hat{Y}_{2m}^{*} I\beta)$ est définie par la formule opt $\beta = \beta_0 = \delta\sigma_3/\sigma_0$.

Si l'on introduit la valeur optimale de $\beta = \beta_0$ dans l'expression de $V(\hat{Y}_{2m}^{*} I\beta)$, on obtient la valeur optimale de

$$V(\hat{Y}_{2m}^{*} I\beta) = V(\hat{Y}_{2m}^{*} I\beta_0) = k[(1 - f) + (1 - \lambda)\,\zeta*/\lambda]\,\sigma_2^2$$

où $\zeta^{*} = (1 - \delta^2)\,\zeta$; $k, f$ et $\zeta$ sont définis respectivement dans (4) et (6).

On obtient la variance optimale de $\hat{Y}_2$ pour une valeur donnée de $\lambda$ en réduisant au minimum la variance de $\hat{Y}_2$ par rapport à $\varphi$ lorsque $\beta = \beta_0$, et la variance optimale est obtenue par la formule

$$V_{\text{opt}}(\hat{Y}_2 I\lambda) = [1/V(\hat{Y}_{2m}^{*} I\beta_0) + 1/(\hat{Y}_{2u})]^{-1}$$

$$= [1/\{k(1 - f) + (1 - \lambda)\,\zeta^{*}/\lambda\} + \mu/\{k(1 - f\mu)\}]^{-1}\,\sigma_2^2.$$

Enfin, en réduisant au minimum la valeur de $V_{\text{opt}}(\hat{Y}_2 I\lambda)$ par rapport à $\lambda$, la proportion optimale de l'échantillon apparié et la variance minimum de $\hat{Y}_2$ se définissent respectivement comme suit:

$$\text{opt } \lambda = \lambda_0 = \sqrt{\zeta^{*}}/(1 + \sqrt{\zeta^{*}})$$

et

$$V_{\min}(\hat{Y}_2) = k(1 - f + \sqrt{\zeta^{*}})\,\sigma_2^2/2 = M_1 \text{ (disons)} \qquad (8)$$

**Remarque 2.1**

L'estimateur $\hat{Y}_{2m}^{*}$, décrit en (1) peut être utilisé en pratique lorsque la valeur optimale de $\beta = \beta_0$ est connue ou que l'on possède une bonne indication de la valeur de $\beta_0$ à partir d'enquêtes précédentes. Si, au lieu de l'estimateur de régression $\hat{Y}_{2m}^{*}$ décrit précédemment, on utilise l'estimateur de différence $\hat{Y}_{2m}^{**} = \sum_{i\epsilon s_m} (y_{2i}^{*}/p_i^{*})\tilde{P}_i - [\sum_{i\epsilon s_m} (z_i^{*}/p_i^{*})\tilde{P}_i - Z]$ basé sur l'échantillon apparié, l'expression de la variance minimum de $\hat{Y}_2$ devient alors:

$$V_{\min}(\hat{Y}_2) = k(1 - f + \sqrt{\tilde{\zeta}})\,\sigma_2^2/2 = \bar{M}_1 \text{ (disons)}$$

avec

$$\tilde{\zeta} = (1 + \tau^2 - 2\tau\delta)\,\zeta,\ \tau = \sigma_0/\sigma_3.$$

### 2.1.1 Estimation de la variance

Pour obtenir des estimateurs sans biais approximatifs de $V_{\text{opt}}(\hat{Y}_2)$, nous énonçons d'abord les théorèmes non démontrés qui suivent:

**Théorème 1**

$$\hat{V}(\hat{Y}_{2m}^{*}) = \{k/(1-k)\}\left[\left\{\sum_{i\epsilon s_m}(y_{2i}^2\Delta_i/p_i^2)\tilde{P}_i/p_i^{*} - \hat{Y}_{2m}^{*2}\right\} + \{k_2/k\}\sum_{i\epsilon s_m}\tilde{P}_i\left(r_i^{*}/p_i^{*} - \sum_{i\epsilon s_m}\tilde{P}_i r_i^{*}/p_i^{*}\right)^2\right]$$

est un estimateur sans biais de $V(\hat{Y}_{2m}^{*})$, lorsque $\beta_0$ est connu, $k = (N-n)/\{n(N-1)\}$ et $k_2 = (n-m)/\{m(n-1)\}$.

**Théorème 2**

$E_1V_2[\sum_{i\epsilon s_m}\tilde{r}_i^{*}/p_1^{*}] = N(n-m)/\{nm(N-1)\}[\sigma_3^2 + \sigma_0^2 - 2\sigma_{03}]$ peut être estimé sans biais à partir de l'équation

$$\{(n-m)/n(m-1)\}\sum_{i\epsilon s_m}(\tilde{r}_i^{*}/p_i^{*} - \sum_{i\epsilon s_m}\tilde{r}_i^{*}/p_i^{*})^2\tilde{P}_i$$

où $\tilde{r}_i^{*} = \tilde{r}_i\Delta_i/p_i$, $\tilde{r}_i = y_{2i} - z_i$; $\sigma_3^2, \sigma_0^2$ et $\sigma_{03}$ sont définis respectivement en (4) et (7).

À partir du théorème 2, nous constatons que

$$\hat{\sigma}_0^2 = d\sum_{i\epsilon s_m}\left(z_i/p_i^{*} - \sum_{i\epsilon s_m}z_i\tilde{P}_i/p_i^{*}\right)^2\tilde{P}_i,$$

$$\hat{\sigma}_3^2 = d\sum_{i\epsilon s_m}\left(y_{2i}/p_i^{*} - \sum_{i\epsilon s_m}y_{2i}\tilde{P}_i/p_i^{*}\right)^2\tilde{p}_i$$

et

$$\hat{\sigma}_{30}^2 = d\sum_{i\epsilon s_m}\left(z_i/p_i^{*} - \sum_{i\epsilon s_m}z_i\tilde{P}_i/p_i^{*}\right)\left(y_{2i}/p_i^{*} - \sum_{i\epsilon s_m}y_{2i}\tilde{P}_i/p_i^{*}\right)\tilde{P}_i$$

sont des estimateurs sans biais de $\sigma_0^2$, $\sigma_3^2$ et $\sigma_{30}$, respectivement, où $d = m(N-1)/\{N(m-1)\}$.

**Estimateur de $V_{opt}(\hat{Y}_2 I\lambda)$**

Par conséquent, pour une valeur donnée de $m$ (*i.e.*, $\lambda$), nous pouvons proposer un estimateur sans biais approximatif de $V_{opt}(\hat{Y}I\lambda)$ qui corresponde à

$$V_{opt}(\hat{Y}I\lambda) = (1/\hat{V}_m + 1/\hat{V}_u)^{-1},$$

où $\hat{V}_m = \hat{V}(\hat{Y}_{2m}^{*} I\beta_0)$ et $\hat{V}_u$ = estimateur sans biais de $V(\hat{Y}_{2u}) = \{(N-u)/N(u-1)\}\sum_{i\epsilon s_u}P_i'(y_{2i}/p_i - \hat{Y}_{2u})^2$.

**Estimateur de $V_{min}(\hat{Y}_2)$**

Si l'on introduit des estimateurs appropriés de $\lambda, \zeta^{*}$ et $\sigma_2^2$ dans l'expression pour $V_{min}(\hat{Y}_2)$, nous obtenons un estimateur sans biais approximatif de $V_{min}(\hat{Y}_2)$ qui correspond à

$$\hat{V}_{min}(\hat{Y}_2) = k[1 - f + (1-\hat{\lambda})\hat{\zeta}^{*}/\hat{\lambda}]/\hat{\sigma}_2^2,$$

où

$$\hat{\zeta}^{*} = (1-\hat{\delta}^2)\hat{\zeta}, \hat{\lambda} = \sqrt{\hat{\zeta}^{*}}/(1+\sqrt{\hat{\zeta}^{*}}),$$

$$\hat{\delta} = \hat{\sigma}_{03}/(\hat{\sigma}_0^2\hat{\sigma}_3^2)^{1/2}, \hat{\zeta}^{*} = \hat{\sigma}_3^2/\hat{\sigma}_2^2,$$

$$\hat{\sigma}_2^2 = \hat{\lambda}\hat{\sigma}_2^2(m) + (1-\hat{\lambda})\hat{\sigma}_2^2(u)$$

$\hat{\sigma}_2^2(m)$ = estimateur sans biais approximatif de $\sigma_2^2$ basé sur l'échantillon apparié $s_m = \sum_{i\epsilon s_m}(y_{2i}^2\Delta_i/p_i^2)\tilde{P}_i/p_i^{*} - \{\hat{Y}_{2m}^{*2} - \hat{V}_m\}$, $\hat{\sigma}_2^2(u)$ = estimateur sans biais approximatif de $\sigma_2^2$ basé sur l'échantillon non apparié $s_u = u(N-1)/\{N(u-1)\}\sum_{i\epsilon s_u}P_i'(y_{2i}P_i'/p_i - \hat{Y}_{2u})^2$; $k$ et $f$ correspondent aux valeurs définies dans (4).

**Remarque 2.2**

Idéalement, l'estimation de $\sigma_2^2$ devrait se faire par la combinaison optimale de $\hat{\sigma}^2(m)$ et $\hat{\sigma}^2(u)$ mais, dans le cas présent, une telle combinaison ferait intervenir des paramètres inconnus. Pour éviter cette source de complexité, l'estimateur plus simple ($\hat{\sigma}^2$) de $\sigma^2$ a été proposé ci-haut.

## 2.2. Stratégie 2

La population est censée être composée de $L$ strates, $N_h$ désignant ici la taille inconnue de la $h$-ième strate ($h = 1, ..., L$; $\sum_h N_h = N$), ce qui signifie que l'on peut déterminer la strate à laquelle appartient une unité dès que sa valeur est observée au premier cycle. Au premier cycle, l'échantillon initial $s_1$ de taille $n$ a été sélectionné par PPTAR, en utilisant la taille normée $p_i$ liée à la $i$-ième unité. Supposons que les $n_h$ unités de $s_1$, qui font partie de la $h$-ième strate, sont représentées par $s_{1h}$. Supposons également que $y_{1i}(h), y_{2i}(h)$ désignent respectivement la valeur de la variable à l'étude pour la $i$-ième unité dans la $h$-ième strate, aux premier et deuxième cycles, et que $z_i(h)$ est la mesure de taille qui y correspond. Au deuxième cycle, des échantillons indépendants $s_{mh}$ de taille $m_h = m n_h/n$ (en supposant que la valeur de chaque $h$ est un nombre entier), sont sélectionnés par la méthode de RHC, en gardant la valeur de $\sum_h m_h = m$ fixe et en utilisant la taille normée $q_{hi}^{*} = [y_{1i}(h)/z_i(h)]/\sum_{i\epsilon s_1}[y_{1i}(h)/z_i(h)]$ pour la $i$-ième unité de la $h$-ième strate. L'échantillon non apparié $s_u$ a été choisi parmi l'ensemble de la population par la méthode de RHC avec la mesure de taille normée $p_i$ pour la $i$-ième unité, comme dans la stratégie 1. Les estimateurs proposés pour $Y_2$, basés sur l'échantillon

apparié $s_m$ et l'échantillon non apparié $s_u$, se définissent respectivement comme suit:

$$\hat{Y}_{2m} = \sum_h w_h \hat{Y}_{2m}(h);\ \hat{Y}_{2u} = \sum_{s_u} (y_{2i}/p_i) P_i' \qquad (9)$$

où

$$\hat{Y}_{2m}(h) = \sum_{s_{mh}} r_i(h)\, Q_{hi} / (n_{1h} p_{hi} q_{hi}^* + c_h \sum_{s_{1h}} z_j(h) /$$

$$(n_{1h} p_{hj}), w_h = n_{1h}/n, p_{hj} = z_j(h)/Z,$$

$$r_i(h) = y_{2i}(h) - c_h y_{1i}(h),$$

$Q_{hi}$ = somme de $q_{hj}^*$ pour le groupe qui contient la $i$-ième unité de la $h$-ième strate créée pour la construction de l'échantillon apparié $s_{mh}$ par la méthode de RHC. $c_h$ sont des constantes choisies pour réduire au minimum la variance de $\hat{Y}_{2m}(h)$. Selon Arnab (1991), l'expression de la variance de $\hat{Y}_{2m}$ est définie par la formule:

$$V(\hat{Y}_{2m}) = k_2 \sum_h \sum_{j=1}^{N_h} q_{hj} (r_{hj}/q_{hj} - R_h)^2 / P(h) + \sigma_2^2/n$$

où $k_2 = (n-m)/n$, $q_{hj} y_{1j}(h)/y_1(h)$, $Y_1(h) = \sum_{j=1}^{N_h}$ taille de la population dans la $h$-ième strate, $P(h) = Z_h/Z$, $Z = \sum_{j=1}^{N_h} z_j(h)$.

La valeur optimale de $c_h$ qui réduit au minimum la valeur de $V(\hat{Y}_{2m})$ et la valeur correspondante de $V(\hat{Y}_{2m})$, sont définies respectivement par

$$\text{opt } c_h = c_h(0) = \delta_{h3} = \sum_{j=1}^{N_h} q_{hj} \alpha_{hj} \beta_{hj} / (\sigma_{h0} \sigma_{h3})$$

et $[1 + (n-m)\theta/m]\,\sigma_2^2/n$, où

$$\alpha_{hj} = y_{2j}(h)/q_{hj} - Y_2(h),\ \beta_{hj} = z_{hj}/q_{hj} - Z_h,$$

$$\sigma_{h3}^2 = \sum_{j=1}^{N_h} q_{hj} \alpha_{hj}^2,\ \sigma_{h0}^2 = \sum_{j=1}^{N_h} q_{hj} \beta_{hj}^2,\ Y_2(h) = \sum_{j=1}^{N_h} y_{2j}(h)$$

et $\theta = \sum_h (1 - \delta_h^2)\,\sigma_{h3}^2 / \{P_h \sigma_2^2\}$.

L'estimateur composite proposé pour $Y_2$, la proportion optimale de l'échantillon apparié et l'expression de la variance minimale de l'estimateur composite $\hat{Y}_2$ sont représentés respectivement par

$$\hat{Y}_2 = \varphi \hat{Y}_{2m} + (1 - \varphi) \hat{Y}_{2u}$$

$$\text{opt } \lambda = \lambda_0 = [\theta - (1-f)\sqrt{\theta}\sqrt{f^*}]/[\theta + f\sqrt{\theta}\sqrt{f^*} - 1]$$

$$V_{\min}(\hat{Y}_2) = k(1/\mu_0 - f)\,\sigma_2^2 / [1 + (\lambda_0/\mu_0)\sqrt{f^*}/\sqrt{\theta}]$$

$$= M_2 \text{ (disons)}$$

où $\hat{Y}_{2m}$ et $\hat{Y}_{2u}$ sont définis en (9), $f^* = N/(N-1)$, $\mu_0 = 1 - \lambda_0$; $k, f$ et $\sigma_2^2$ sont définis en (4).

## 3. EFFICACITÉ DES STRATÉGIES PROPOSÉES

La stratégie 1 proposée est plus efficace que la stratégie de Prasad et Graham (1994), en ce qu'elle donne une variance minimale plus faible, car $\delta^2 \le 1$. Par ailleurs, l'efficacité de la stratégie 1 augmente à mesure que $\delta$, la corrélation entre $y_{2i}/q_i$ et $z_i/q_i$ augmente. L'efficacité de la stratégie 1 et de la stratégie de Prasad et Graham (1994) augmente à mesure que $\zeta$ diminue. La valeur de $\zeta = \sigma_3^2/\sigma_2^2$ varie en fonction des valeurs de $\sigma_3^2$ et $\sigma_2^2$. Ainsi, la valeur de $\sigma_3^2$ sera inférieure (ou supérieure) à $\sigma_2^2$, selon que la proportionnalité de $y_{2i}$ par rapport à $y_{1i}$ est supérieure (ou inférieure) à celle de $y_{2i}$ par rapport à $z_i$. De toute évidence, la stratégie 1 peut être utilisée dans la pratique lorsqu'on possède une bonne valeur provisoire de $\beta$ à partir des enquêtes précédentes. Si l'estimateur de la différence est utilisé dans la stratégie 1 en remplacement de l'estimateur de régression mentionné à la remarque 2.1, alors la stratégie 1 proposée donne de meilleurs résultats que celle de Prasad et Graham (1994), chaque fois que $\delta > \frac{1}{2}\sigma_0/\sigma_3$. Par ailleurs, la stratégie 1 donne des résultats meilleurs ou pires que celle de Chotai (1974), selon que $\zeta^* = (1 - \delta^2)\zeta$ <ou> $(1 - \delta^{*2})$. Ici, $\delta^*$ peut être considéré comme un coefficient de corrélation entre $y_{2i}/p_i$ et $y_{1i}/p_i$. Si $z_i$, sont des constantes, alors $\delta^*$ devient le coefficient de corrélation simple entre $y_{1i}$ et $y_{2i}$. L'expression pour la variance minimale $M_2$, selon la stratégie 2, est complexe et ne permet pas d'établir des comparaisons simples avec les autres stratégies décrites ici. Nous remarquons toutefois que l'efficacité de la stratégie 2 augmente à mesure que le coefficient de corrélation de la strate $\delta_{h3}$ augmente. Les exemples numériques qui suivent, lesquels sont basés sur les données réelles, montrent que la stratégie 2 proposée donne de meilleurs résultats que la stratégie 1 et aussi que les autres stratégies proposées par Prasad et Graham (1994) et par Chotai (1974).

Pour établir des comparaisons, trois ensembles de données sont examinés. Le premier (que nous désignerons population 1) a été étudié par Prasad et Graham (1994); il fait référence à la superficie ensemencée en blé en 1937 ($y_2$) et en 1936 ($y_1$) ainsi qu'à la superficie en culture ($z$) dans un ensemble de 34 villages de l'Inde, selon des données compilées par Sukhatme et Sukhatme (1970). La population 1 a été divisée en deux strates, selon que la

superficie en blé en 1936 était supérieure ou inférieure à 200 acres. Les paramètres pour cette population sont: $N = 34, N_1 = 20, N_2 = 14, \delta^* = 0{,}7635, \delta = 0{,}3638, \zeta = 0{,}3811, \theta = 0{,}2436$. La population 2 représente la production de céréales en Amérique du sud en 1980 ($z$), 1988 ($y_1$) et 1989 ($y_2$), selon les données compilées dans l'Annuaire statistique des Nations-Unies (1988-1989). Cette population est elle aussi stratifiée en deux strates, selon que la production en 1988 est supérieure ou inférieure à 570 (milliers de tonnes métriques). Les paramètres pour cette population sont: $N = 19, N_1 = 7, N_2 = 12, \delta^* = -0{,}6939, \delta = 0{,}7666, \zeta = 1{,}1478, \theta = 0{,}3681$. Enfin, la population 3 compilée par Singh et Chaudhuri (1986) fait référence à la superficie ensemencée en blé, en hectares, en 1979-1980 ($y_2$) et 1978-1979 ($y_1$) et à la superficie totale en culture en 1978-1979 ($z$) dans 16 villages du district de Meerut. Les paramètres pour la population 3 sont: $N = 16, N_1 = 9, N_2 = 7, \delta^* = 0{,}7729, \delta = 0{,}1057, \zeta = 0{,}3965, \theta = 0{,}2827$.

Le tableau qui suit indique l'efficacité relative des stratégies proposées 1 et 2 ainsi que de la stratégie proposée par Prasad et Graham (1994) par rapport à celle de Chotai (1974), efficacité qui est représentée respectivement par $E_1 = V_c/M_1$, $E_2 = V/M_2$ et $E_3 = V_c/V_{PG}$.

**Tableau 1**
Efficacité des diverses stratégies

| $f$ | Population 1 | | | Population 2 | | | Population 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $E_1$ | $E_2$ | $E_3$ | $E_1$ | $E_2$ | $E_3$ | $E_1$ | $E_2$ | $E_3$ |
| 0,05 | 1,0463 | 1,1033 | 1,0181 | 1,0196 | 1,0850 | 0,8262 | 1,0053 | 1,0864 | 1,0030 |
| 0,10 | 1,0479 | 1,0895 | 1,0187 | 1,0202 | 1,0711 | 0,8212 | 1,0055 | 1,0711 | 1,0031 |
| 0,15 | 1,0496 | 1,0776 | 1,0194 | 1,0209 | 1,0579 | 0,8172 | 1,0057 | 1,0577 | 0,0033 |
| 0,20 | 1,0514 | 1,0683 | 1,0200 | 1,0216 | 1,0519 | 0,8123 | 1,0058 | 1,0469 | 1,0034 |
| 0,25 | 1,0533 | 1,0622 | 1,0208 | 1,0224 | 1,0490 | 0,8071 | 1,0061 | 1,0396 | 1,0035 |
| 0.30 | 1,0554 | 1,0604 | 1,0216 | 1,0232 | 1,0530 | 0,8017 | 1,0063 | 1,068 | 1,0036 |

À la lumière du tableau qui précède, nous constatons que, pour les trois populations à l'étude, la stratégie 2 donne de meilleurs résultats que les autres. Il convient également de souligner que les deux stratégies proposées donnent toutes deux de meilleurs résultats que celle de Chotai (1974) ou encore celle de Prasad et Graham (1994). Pour la population 1, $\zeta = 0{,}3811$, ce qui favorise grandement la stratégie de Prasad et Graham (1994) et, partant, la stratégie 1 proposée. La stratégie de Prasad et Graham (1994) et la stratégie 1 ont toutes deux devancé la stratégie de Chotai (1974). Pour la population 2, $\zeta = 1{,}1478$, ce qui est une valeur à la fois élevée et défavorable en regard de la stratégie de Prasad et Graham (1994) tandis que $\delta = 0{,}7666$ favorise bien la stratégie 1. Par conséquent, pour la population 2, la stratégie de Prasad et Graham (1994) devient moins efficace que celle de Chotai (1974), mais la stratégie 1 proposée demeure préférable. Enfin, pour la population 3, $\zeta = 0{,}3965$ ce qui est assez bon pour la stratégie de Prasad et Graham (1994) mais parallèlement $\delta^* = 0{,}7729$ et ceci ($\delta^*$) favorise la stratégie de Chotai (1974). En fait, la stratégie de Chotai (1974) est légèrement inférieure à celle de Prasad et Graham (1994) mais la stratégie 2 proposée demeure supérieure aux deux. Il convient toutefois de préciser que les exemples présentés ici sont assez inhabituels, en ce qu'ils représentent de faibles corrélations entre $y_2$ et $z$ (dans l'exemple 1, $\delta = 0{,}3638$ et dans l'exemple 3, $\delta = 0{,}1057$) et qu'il y a une corrélation négative entre $y_2$ et $y_1$ ($\delta^* = -0{,}6939$) dans l'exemple 2. On s'attend à ce que les corrélations $\delta$ et $\delta^*$ soient élevées et positives. Il faudra donc mener d'autres études pour comparer le rendement des stratégies présentées au moyen de données appropriées.

**Tableau 2**
Sensibilité de l'efficacité $E^* = V_{PG}/M_\beta$

| $\lvert v \rvert$ | 0,05 | 0,1 | $f$ 0,15 | 0,2 | 0,25 | 0,3 |
|---|---|---|---|---|---|---|
| | | | Population 1 | | | |
| 0,0 | 1,028 | 1,029 | 1,030 | 1,031 | 1,032 | 1,033 |
| 0,2 | 1,027 | 1,027 | 1,028 | 1,029 | 1,031 | 1,032 |
| 0,4 | 1,023 | 1,024 | 1,027 | 1,026 | 1,027 | 1,028 |
| 0,6 | 1,017 | 1,108 | 1,019 | 1,019 | 1,020 | 1,021 |
| 0,8 | 1,010 | 1,010 | 1,010 | 1,011 | 1,011 | 1,011 |
| 1,0 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| 1,2 | 0,989 | 0,988 | 0,988 | 0,988 | 0,988 | 0,987 |
| 1,4 | 0,976 | 0,976 | 0,975 | 0,974 | 0,973 | 0,972 |
| | | | Population 2 | | | |
| 0,0 | 1,234 | 1,241 | 1,249 | 1,257 | 1,266 | 1,278 |
| 0,2 | 1,219 | 1,227 | 1,233 | 1,241 | 1,249 | 1,258 |
| 0,4 | 1,180 | 1,186 | 1,191 | 1,197 | 1,204 | 1,211 |
| 0,6 | 1,125 | 1,128 | 1,133 | 1,137 | 1,141 | 1,146 |
| 0,8 | 1,063 | 1,065 | 1,067 | 1,068 | 1,070 | 1,073 |
| 1,0 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| 1,2 | 0,939 | 0,938 | 0,936 | 0,935 | 0,933 | 0,931 |
| 1,4 | 0,883 | 0,880 | 0,877 | 0,875 | 0,871 | 0,869 |
| | | | Population 3 | | | |
| 0,0 | 1,002 | 1,002 | 1,004 | 1,003 | 1,003 | 1,003 |
| 0,2 | 1,002 | 1,002 | 1,002 | 1,002 | 1,003 | 1,002 |
| 0,4 | 1,002 | 1,002 | 1,002 | 1,002 | 1,002 | 1,002 |
| 0,6 | 1,002 | 1,002 | 1,002 | 1,002 | 1,002 | 1,001 |
| 0,8 | 1,001 | 1,001 | 1,001 | 1,001 | 1,001 | 1,001 |
| 1,0 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| 1,2 | 0,999 | 0,999 | 0,999 | 0,999 | 0,999 | 0,999 |
| 1,4 | 0,998 | 0,997 | 0,998 | 0,998 | 0,998 | 0,998 |

Pour étudier l'effet de la déviation par rapport à la valeur optimale de $\beta = \beta_0$ lorsqu'une valeur provisoire de $\beta$

est utilisée dans la stratégie 1, on peut examiner la sensibilité de l'efficacité de $\hat{Y}_2$ pour la stratégie 1, en utilisant différentes valeurs de $\beta$, selon la méthode Prasad et Srivenkataramana (1980). Lorsqu'on utilise une valeur provisoire $\beta_0 = \tilde{\beta}$ , la variance minimale de $\hat{Y}_2$ pour la stratégie 1 devient

$$V_{\min}(\hat{Y}_2 \mid \tilde{\beta}) = k(1 - f + \sqrt{\zeta^{**}})\,\sigma_2^2/2 = M_{\tilde{\beta}} \qquad (9)$$

où $\zeta^{**} = [1 - (1 - \nu^2)\,\delta^2]\,\zeta$ et $\nu = 1 - \tilde{\beta}/\beta_0$.

À partir de l'équation (9), nous constatons que la stratégie 1 proposée , basée sur une valeur provisoire pour $\tilde{\beta}$, donne des résultats meilleurs ou pires que ceux obtenus par la stratégie de Prasad et Graham (1994) selon que $|\nu| < 1$ ou $|\nu| > 1$. De même, la stratégie 1, avec $\beta = \tilde{\beta}$, est supérieure ou inférieure à celle de Chotai (1974) selon que $\nu^2 >$ ou $< (1 - 1/\delta^2)(1 - 1/\zeta)$. Le tableau 2 qui précède indique la sensibilité $E^*$ de l'estimateur $\hat{Y}_2$ par rapport à la stratégie de Prasad et Graham (1994) où $E^* = V_{PG}/M_{\tilde{\beta}}$. À partir du tableau 2, nous constatons que la perte avec $\nu > 1$ risque d'être plus élevée que le gain avec $\nu < 1$ pour la population 1 et 3, alors que l'inverse vaut pour la population 2.

## CONCLUSION

Pour l'échantillonnage sur deux cycles, il vaut mieux utiliser les données recueillies au premier cycle, pour obtenir un estimateur efficace de l'ensemble de la population au deuxième cycle. Chotai (1974) a utilisé les données recueillies au premier cycle, au stade de l'estimation, alors que Prasad et Graham (1994) les ont utilisées pour la sélection (et donc l'estimation) de l'échantillon apparié. Nous proposons ici deux stratégies. La première s'appuie sur les données recueillies au premier cycle, pour la sélection de l'échantillon apparié, comme le proposent Prasad et Graham (1994), et aussi pour la définition d'un estimateur de régression selon la méthode proposée par Chotai (1974). Il en résulte que la stratégie 1 est plus efficace que celle de Prasad et Graham (1994). La stratégie 2 utilise les valeurs du premier cycle comme variable de stratification, comme mesure de la taille pour la sélection de l'échantillon apparié au deuxième cycle et pour la formation d'un estimateur de régression utilisant la variable auxiliaire ($z$) obtenue au premier cycle. Intuitivement, on peut s'attendre à ce que la stratégie 2 donne de meilleurs résultats que les autres mentionnées ici; cependant, aucun résultat théorique n'a pu être obtenu en raison de la complexité de l'expression de la variance minimale de l'estimateur proposé. La supériorité de la stratégie 2 a cependant pu être établie grâce à des données numériques.

## REMERCIEMENTS

L'auteur aimerait remercier l'examinateur, le rédacteur adjoint et le rédacteur en chef pour leurs précieux commentaires qui ont permis d'améliorer sensiblement la version antérieure du présent article. Ces travaux ont été réalisés grâce à une aide du FRD, en Afrique du sud.

## BIBLIOGRAPHIE

ARNAB, R. (1991). On sampling over two occasions using varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 43(3), 282-290.

AVADHANI, M.S., et SUKHATME, B.V. (1970). A comparison of two sampling procedures with an application to successive sampling. *Applied Statistics*, 19, 251-259.

CHOTAI, J. (1974). A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhyā*, Series C, 36, 173-180.

PRASAD, N.G.N., et SRIVENKATARAMANA, T. (1980). A modification to the Horvitz-Thompson estimator under the Midzuno sampling scheme. *Biometrika*, 67, 709-711.

PRASAD, N.G.N., et GRAHAM, J.E. (1994). Échantillonnage avec PPT en deux occasions. *Techniques d'enquête*, 20, 63-68.

RAJ, D. (1965). On sampling over two occasions with probabilities proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.

RAO, J.N.K., HARTLEY, H.O., et COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-491.

SINGH, D., et CHAUDHURI, F.S. (1986): *Theory and Analysis of Sample Survey Designs*. India: Wiley Eastern Limited, 166.

SINGH, M.P. (1967). The relative efficiency of some two-phase sampling schemes. *Annals of Mathematical Statistics*, 38, 937-940.

SUKHATME, P.V., et SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State University Press, 185.

UNITED NATIONS (1992). *Statistical Year Book*, (1988/89). New York: United Nations, 356.

Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 209-218
Statistique Canada



# Intervalles de confiance pour les proportions à petit nombre d'événements positifs prévus estimées au moyen des données d'enquête

EDWARD L. KORN et BARRY I. GRAUBARD[1]

RÉSUMÉ

En dehors des enquêtes complexes, on applique fréquemment des intervalles de confiance «exacts» aux proportions obtenues par distribution binomiale au lieu d'intervalles reposant sur une normalité approximative, quand les événements positifs sont peu nombreux. Malheureusement, il est impossible d'en faire autant avec les données des enquêtes complexes, en sorte qu'on utilise des intervalles reposant sur la normalité approximative hypothétique de la proportion pondérée selon l'échantillon, même quand il existe peu d'événements positifs. Les auteurs proposent une simple modification aux intervalles binomiaux dans une situation de ce genre. Les simulations restreintes qu'ils présentent indiquent que la couverture probable de tels intervalles dépasse celle des intervalles reposant sur la normalité, des intervalles venant de la transformation par logit et des intervalles résultant de l'approximation de Poisson. Ils appliquent ensuite ces intervalles à la prévalence du virus de l'immunodéficience humaine (VIH) selon les données de la troisième Enquête nationale sur la santé et la nutrition et à la proportion de cocaïnomanes selon les données de l'Enquête sur la santé et la nutrition des personnes hispaniques.

MOTS CLÉS: Intervalle de confiance binomial; intervalle de confiance exact; transformation par logit; intervalle confiance de Poisson.

## 1. INTRODUCTION

Avec les données d'une enquête complexe, l'intervalle de confiance de niveau $1 - \alpha$ typique d'une proportion d'événements positifs pour une variable de type 0-1 ressemble à

$$\hat{p} \pm t_d\,(1 - \alpha/2)\,[\hat{\text{var}}(\hat{p})]^{1/2} \qquad (1.1)$$

où $\hat{p}$ représente l'estimateur de la proportion pondéré pour l'échantillon, où $\hat{\text{var}}(\hat{p})$ est l'estimateur de la variance de $\hat{p}$, et où $t_d(1 - \alpha/2)$ correspond au $1 - \alpha/2$ quantile d'une distribution $t$ à $d$ degrés de liberté. On calcule l'estimateur $\hat{\text{var}}(\hat{p})$ par linéarisation ou par répétition au moyen d'une méthode tenant compte du plan d'échantillonnage, notamment le fait que $\hat{p}$ est un estimateur pondéré. Par données d'une enquête complexe, on entend les données recueillies en vertu d'un plan d'échantillonnage à plusieurs degrés, avec sélection stratifiée de grappes au premier degré. Avec un plan d'échantillonnage de ce genre, $d$ correspond habituellement au nombre de grappes échantillonnées moins le nombre de strates (Korn et Graubard 1990). L'intervalle de confiance (1.1), que nous appellerons «intervalle linéaire» repose sur l'hypothèse que $\hat{p}$ a une distribution proche de la normale. L'hypothèse se vérifie pour diverses asymptotes raisonnables (Krewski et Rao 1981). L'emploi du quantile $t$ au lieu d'un quantile à distribution normale dans (1.1) s'appuie sur des preuves empiriques (Frankel 1971, ch. 7), mais il est possible de le justifier formellement à partir d'hypothèses fermes (Korn et Graubard 1990).

Avec un nombre restreint d'événements positifs cependant, $\hat{p}$ perd sa normalité approximative (Cochran 1977, p. 58). On peut éviter l'hypothèse de la normalité en utilisant l'intervalle de confiance de Clopper et Pearson (1934) reposant sur la distribution binomiale lorsqu'on est en présence d'un échantillon aléatoire simple (ou qu'il ne s'agit pas d'une enquête complexe); Vollset (1993) donne une analyse complète des intervalles de confiance des proportions dans un tel contexte. Quand un échantillon aléatoire simple de taille $n$ comporte $x$ valeurs positives, l'intervalle de confiance Clopper et Pearson $(p_L(x, n), p_U(x, n))$ de niveau $1 - \alpha$ peut s'exprimer comme suit (Johnson, Kotz et Kemp 1993, p. 130):

$$p_L(x, n) = \frac{\nu_1 F_{\nu_1,\nu_2}(\alpha/2)}{\nu_2 + \nu_1 F_{\nu_1,\nu_2}(\alpha/2)}$$

$$p_U(x, n) = \frac{\nu_3 F_{\nu_3,\nu_4}(1 - \alpha/2)}{\nu_4 + \nu_3 F_{\nu_3,\nu_4}(1 - \alpha/2)} \qquad (1.2)$$

où $\nu_1 = 2x$, $\nu_2 = 2\,(n - x + 1)$, $\nu_3 = 2(x + 1)$, $\nu_4 = 2\,(n - x)$ et $F_{d_1, d_2}(\beta)$ correspond au quantile $\beta$ d'une distribution $F$ à $d_1$ et $d_2$ degrés de liberté. Pour un intervalle de confiance unilatéral, on utilise $\alpha$ au lieu de $\alpha/2$ dans les expressions ci-dessus. On sait qu'avec un échantillon aléatoire simple, la probabilité de couverture de tels intervalles est supérieure ou égale à leur niveau nominal, peu importe le nombre d'événements positifs. On parle parfois d'intervalles de confiance «exacts», mais nous dirons plutôt «intervalles binomiaux».

[1] Edward L. Korn, Biometric Research Branch, EPN-739, National Cancer Institute, Bethesda, MD 20892, U.S.A.; Barry I. Graubard, Biostatistics Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A.

Nous proposons une simple modification aux intervalles binomiaux susceptible d'en permettre l'application aux proportions estimées à partir des données d'une enquête complexe. Nous nous intéressons particulièrement au cas où un petit nombre d'événements positifs est prévu. Beaucoup d'analystes d'enquête ne fourniront pas de proportions estimées en pareil cas, car elles manqueraient de fiabilité. Si on appliquait le critère de l'erreur-type relative aux proportions de l'Enquête-ménages nationale sur la toxicomanie de 1996 (SAMHSA 1998), par exemple, la proportion estimée de femmes cocaïnomanes n'apparaîtrait pas au tableau 7. Or, nous pensons que de telles proportions véhiculent une information intéressante pourvu qu'on en précise de façon explicite l'imprécision en donnant les intervalles de confiance. On trouvera à la partie 2 les intervalles que nous proposons et ceux reposant sur la transformation par logit et la distribution de Poisson suggérés par d'autres auteurs. À la partie 3, les résultats d'une simulation comparent l'efficacité des intervalles en question. On constate que les intervalles proposés donnent de bons résultats pour ce qui est de la probabilité de couverture de la proportion réelle et de la largeur moyenne de l'intervalle. La partie 4 présente deux applications des intervalles dans une enquête majeure où l'on s'attend à un faible nombre d'événements positifs. L'article se termine par une analyse de travaux connexes débouchant sur des intervalles de confiance qui atteindront leur probabilité de couverture nominale quelle que soit la configuration des groupes au sein de la population.

## 2. LIMITES DES INTERVALLES PROPOSÉS ET DES AUTRES INTERVALLES DE CONFIANCE

Pour l'intervalle de confiance $1-\alpha$ d'un échantillon de taille $n$, définissons d'abord la taille efficace de l'échantillon par

$$n^* = \frac{\hat{p}(1-\hat{p})}{\text{vâr}(\hat{p})} \tag{2.1}$$

et la taille efficace de l'échantillon ajustée selon le nombre de degrés de liberté par

$$n^*_{df} = \frac{\hat{p}(1-\hat{p})}{\text{vâr}(\hat{p})}\left(\frac{t_{n-1}(1-\alpha/2)}{t_d(1-\alpha/2)}\right)^2 \tag{2.2}$$

Les valeurs $n^*$ et $n^*_{df}$ sont égales à $n$ quand $\hat{p}=0$. Les limites proposées remplacent $n$ par $n^*_{df}$, et $x$ par $\hat{p}n^*_{df}$ dans (1.2), *viz.* $p_L(\hat{p}n^*_{df}, n^*_{df})$ et $p_U(\hat{p}n^*_{df}, n^*_{df})$. (Quand $n$ est élevé, on peut se servir du quantile $1-\alpha/2$ de la distribution normale au lieu de $t_{n-1}(1-\alpha/2)$ dans (2.2).) Pour estimer l'intervalle de confiance d'une proportion dans un sous-groupe de la population, on suppose que la taille de l'échantillon $n$ est égale à celle de l'échantillon restreint au sous-groupe.

Voici la justification heuristique de la procédure qui précède. La taille efficace de l'échantillon (2.1) est $n$ divisée par un estimateur de l'effet du plan de sondage de l'enquête. La chose paraît raisonnable si on veut tenir compte de la variabilité supplémentaire de $\hat{p}$ attribuable à l'échantillonnage complexe. La variabilité de l'estimateur de la variance présente elle aussi de l'importance lorsqu'on construit les intervalles de confiance. La deuxième fraction de (2.2) tient compte du fait que $\text{vâr}(\hat{p})$ variera typiquement plus qu'un estimateur de la variance utilisé avec un échantillonnage aléatoire simple. Si $d$ est élevé, ce facteur prend presque la valeur un et est inutile. Avec une faible valeur $d$ et une valeur $n$ et $\hat{p}n^*_{df}$ élevées, il serait préférable que l'intervalle proposé se rapproche de l'intervalle (1.1), qui convient en pareil cas. Puisque $F_{u,w}(\beta) \cong 1 + z(\beta)\sqrt{2(1/u + 1/w)}$ pour les valeurs $u$ et $w$ élevées (Johnson et Kotz 1970, p. 81), c'est effectivement ce qui se produit, à savoir $\hat{p} - p_L(\hat{p}n^*_{df}, n^*_{df}) \cong p_U(\hat{p}n^*_{df}, n^*_{df}) - \hat{p} \cong t_d(1-\alpha/2)[\text{vâr}(\hat{p})]^{1/2}$.

Breeze (1990) a élaboré une méthode analogue à celle que nous suggérons pour l'Enquête générale sur les ménages du Royaume-Uni. Cette méthode repose sur l'intervalle de confiance $1-\alpha$ de l'échantillonnage aléatoire simple $(po_L(x), po_U(x))$ pour une variable de Poisson $x$, s'exprimant comme suit (Johnson et coll. 1993, p. 171):

$$po_L(x) = 0.5\,\chi^2_{\nu_1}(\alpha/2) \text{ et } po_U(x) = 0.5\,\chi^2_{\nu_2}(1-\alpha/2)$$

où $\nu_1 = 2x$, $\nu_2 = 2(x+1)$ et $\chi^2_\nu(\beta)$ correspond au quantile $\beta$ d'une distribution $\chi^2$ à $\nu$ degrés de liberté. On suppose que l'intervalle de confiance est égal à $(po_L(\hat{p}n^*)/n^*, po_U(\hat{p}n^*)/n^*)$ avec les données d'une enquête complexe.

Une troisième façon de bâtir un intervalle de confiance consiste à recourir à la transformation par logit. L'intervalle de confiance de niveau $1-\alpha$ correspond à

$$\left(\frac{1}{1+\exp(-LLOGIT)}, \frac{1}{1+\exp(-ULOGIT)}\right)$$

où

$$LLOGIT = \log\frac{\hat{p}}{1-\hat{p}} - t_d(1-\alpha/2)\frac{[\text{vâr}(\hat{p})]^{1/2}}{\hat{p}(1-\hat{p})} \tag{2.3}$$

et

$$ULOGIT = \log\frac{\hat{p}}{1-\hat{p}} + t_d(1-\alpha/2)\frac{[\text{vâr}(\hat{p})]^{1/2}}{\hat{p}(1-\hat{p})} \tag{2.4}$$

On a suggéré d'utiliser ces intervalles avec un quantile à distribution normale plutôt qu'à distribution $t$ pour l'Enquête-ménages nationale sur la toxicomanie de 1996 (SAMHSA 1998). En dehors des conditions d'enquête complexe, quand $\hat{p}=0$, on pourrait ajouter une petite constante au nombre d'événements et de non-événements observés, par ex. 1/2, afin de calculer l'intervalle de

confiance de la transformation par logit (Agresti 1990, pp. 249-250). Dans le cas présent, quand $\hat{p} = 0$, l'intervalle de confiance a été fixé à l'intervalle binomial $(p_L(0,n), p_U(0,n))$.

Lorsqu'on sait que l'effet (véritable) du plan d'échantillonnage dépassera 1 avant l'échantillonnage, il est possible d'apporter diverses modifications aux procédures qui précèdent. Nous préconisons pour notre part de tronquer à $n$ la taille efficace de l'échantillon ajustée en fonction du nombre de degrés de liberté. Bref, si $n^*_{df}$ est plus grand que $n$, nous en fixons la valeur à $n$ et prenons comme limites supérieure et inférieure de l'intervalle $p_L(\hat{p}n, n)$ et $p_U(\hat{p}n, n)$. Pour les intervalles de Breeze, on pourrait établir que $n^*$ est égal à $n$ si $n^* > n$. Avec l'intervalle linéaire ou à logit, il est possible de prendre l'estimateur de la variance de l'échantillon aléatoire simple $\hat{p}(1-\hat{p})/n$ au lieu de $\hat{\text{var}}(\hat{p})$ dans (1.1), (2.3) et (2.4) si $n^* > n$; on trouvera d'autres troncations dans SAMHSA (1998). On justifie ces troncations par le fait que l'instabilité de l'estimateur de la variance $\hat{\text{var}}(\hat{p})$ peut donner un effet de plan estimé inférieur à un. L'instabilité en question peut s'avérer particulièrement importante puisque $\hat{p}$ est petit (SAMHSA 1998). Ces méthodes de troncation ont pour conséquence d'élargir les intervalles de confiance et de les rendre plus conservateurs. En théorie, on pourrait aussi ajuster la taille réelle estimatée de l'échantillon si on sait que l'effet (véritable) du plan de sondage est inférieur à un avant l'échantillonnage. Par mesure de prudence, nous recommandons toutefois de l'éviter.

Dans cet article, nous nous concentrons sur les intervalles de confiance de la probabilité de «superpopulation» que $Y = 1$ et n'égale pas la proportion de la population finie. En d'autres termes, le paramètre visé est $p = \sum_{u=1}^{N} p_u / N$ et non pas $P = \sum_{u=1}^{N} Y_u / N$, où $Y_u$ présente une distribution de Bernoulli de paramètre $p_u$, et $N$ correspond à la taille de la population. Les probabilités de couverture obtenues par simulation qui apparaissent à la section suivante se rapportent donc à la couverture de $p$. Ce paramètre étant établi, nous ne recourrons pas aux facteurs de correction de la population finie au moment d'estimer $\hat{\text{var}}(\hat{p})$ utilisé dans (2.2); nous ne procéderons pas ici à d'autres ajustements de la variance $\hat{\text{var}}(\hat{p})$ attribuable au plan d'échantillonnage pour l'inférence à la superpopulation (Korn et Graubard 1998). Un examinateur a suggéré qu'on recoure à un modèle pour estimer l'intervalle de confiance de $p$. Bien que limités, nos essais indiquent néanmoins qu'une approche de ce genre donne des estimateurs semblables aux estimateurs pondérés et ne présente aucun avantage sur le plan de l'inférence (Pfeffermann et LaVange 1989; Graubard et Korn 1996).

Ceux qui s'intéressent à un intervalle de confiance pour $P$ devraient se servir des intervalles proposés en intégrant des facteurs de correction pour la population finie à $\hat{\text{var}}(\hat{p})$, dans (2.2). On pourrait établir un intervalle de confiance pour $\sum_{u=1}^{N} Y_u$ en multipliant les extrémités de l'intervalle de $P$ par $N$, s'il est connu, ou par l'estimateur $\hat{N}$ de $N$, dans le cas contraire. (Théoriquement, on pourrait aussi tenir compte de la variabilité de $\hat{N}$, mais la variabilité supplémentaire sera faible.) Une autre façon d'estimer l'intervalle de confiance de $P$ consisterait à modifier les limites usuelles (Guenther 1983) d'un échantillon aléatoire simple (d'après la distribution hypergéométrique), un peu comme les intervalles suggérés modifient les intervalles binomiaux.

## 3. SIMULATIONS

Les principaux résultats de la simulation apparaissent aux tableaux 1 à 5. Le tableau 1 donne les résultats des simulations pour 32 grappes de 100 unités chacune. Dans la grappe $i$, on a simulé le nombre d'événements positifs au moyen d'une distribution binomiale assortie de $p_i$ comme paramètre probabiliste. Les valeurs $\{p_i, i = 1, ..., 32\}$ du tableau 1 correspondent à la probabilité dans les grappes. Dans le tiers supérieur du tableau, cette probabilité est égale à la constante $p$ = 0,1, 0,02, 0,01 ou 0,0025, ce qui correspond à un nombre d'événements positifs prévu de 320, 64, 32 ou 8 sur un total possible de 3 200. Dans le tiers intermédiaire du tableau, la probabilité associée aux grappes est égale à $p/2$ avec une probabilité de 1/2 ou à $3p/2$ avec une probabilité de 1/2, $p$ ayant la même valeur que dans le premier tiers. En faisant varier $p_i$ d'une grappe à l'autre, une corrélation se développe entre les observations de la grappe. Cette corrélation est égale à 0,00278, 0,0051, 0,0025 et 0,0006 dans le tiers intermédiaire (quand on néglige les poids de l'échantillonnage) et correspond à 320, 64, 32 ou 8 événements positifs prévus, respectivement. Pour le tiers inférieur du tableau, la probabilité est égale à $p/2$ avec une probabilité de 3/4 ou à $5p/2$, avec une probabilité de 1/4, ce qui correspond à une corrélation intraclasse de 0,0833, 0,0153, 0,0076 et 0,0019. Un poids d'échantillonnage de 1 ou de 10 est attribué au hasard dans les simulations du tableau 1, avec une probabilité de 1/2 pour chaque observation (poids sans valeur informative).

Les résultats du tableau 1 conviennent aux limites inférieure et supérieure d'un intervalle de confiance unilatéral de 95 %; idéalement, le pourcentage représentant la non-couverture dans le tableau devrait être inférieur ou égal à la valeur nominale de 5,0. Les résultats conviennent aussi aux intervalles bilatéraux de 90 %, pour lesquels, idéalement, les valeurs supérieure et inférieure du tableau devraient être ≤5,0 (Jennings 1987). Pour chaque ligne du tableau, on a simulé 100 000 jeux de données au moyen du générateur de nombre aléatoire du SAS (1990, p. 631), de manière à estimer la probabilité de non-couverture des limites de l'intervalle.

**Tableau 1**

Simulation de la non-couverture (en pour cent) des limites supérieure et inférieure d'un intervalle de confiance unilatéral de 95 % comptant 32 grappes de 100 observations chacune; poids d'échantillonnage de 1 ou de 10 avec probabilité de 1/2 (poids non informatifs)

| Distribution des proportions pour les grappes[a] | Proportion globale | Nombre prévu positif | Méthode de calcul des limites de l'intervalle | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linéaire | | Logit | | Breeze | | Proposée | |
| | | | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. |
| (1) | | | | | | | | | | |
| 0,1 | 0,1 | 320 | 4,6 | 5,5 | 5,3 | 4,6 | 4,5 | 4,1 | 4,8 | 4,4 |
| 0,02 | 0,02 | 64 | 3,4 | 7,1 | 5,2 | 4,6 | 4,5 | 4,7 | 4,2 | 4,4 |
| 0,01 | 0,01 | 32 | 2,9 | 8,0 | 5,4 | 4,5 | 4,4 | 4,5 | 4,0 | 4,1 |
| 0,0025 | 0,0025 | 8 | 1,6 | 9,5 | 5,5 | 1,8 | 3,6 | 2,2 | 3,3 | 1,8 |
| (1/2, 1/2) | | | | | | | | | | |
| 0,05, 0,15 | 0,1 | 320 | 4,3 | 5,8 | 5,5 | 4,3 | 4,3 | 3,8 | 4,7 | 4,1 |
| 0,01, 0,03 | 0,02 | 64 | 3,1 | 7,5 | 5,2 | 4,8 | 4,3 | 4,8 | 4,0 | 4,5 |
| 0,005, 0,015 | 0,01 | 32 | 2,7 | 8,6 | 5,2 | 4,7 | 4,1 | 4,9 | 3,7 | 4,4 |
| 0,00125, 0,00375 | 0,0025 | 8 | 1,5 | 9,9 | 5,4 | 2,0 | 3,4 | 2,3 | 3,1 | 2,0 |
| (3/4, 1/4) | | | | | | | | | | |
| 0,05, 0,25 | 0,1 | 320 | 3,1 | 7,8 | 4,7 | 5,6 | 3,4 | 5,0 | 3,6 | 5,3 |
| 0,01, 0,05 | 0,02 | 64 | 2,7 | 8,6 | 5,1 | 5,3 | 4,0 | 5,4 | 3,7 | 5,0 |
| 0,005, 0,025 | 0,01 | 32 | 2,2 | 9,8 | 5,0 | 5,3 | 3,7 | 5,5 | 3,3 | 5,0 |
| 0,00125, 0,00625 | 0,0025 | 8 | 1,3 | 10,7 | 5,3 | 2,2 | 3,3 | 2,5 | 3,0 | 2,2 |

(a) Les fractions entre parenthèses indiquent la probabilité que la proportion de grappes corresponde à la valeur mentionnée.

**Tableau 2**

Simulation de la non-couverture (en pour cent) des limites supérieure et inférieure d'un intervalle de confiance de 95 % pour un échantillon de 32 grappes de 100 observations chacune; les poids d'échantillonnage sont informatifs et correspondent à 1 ou 10 (voir le corps du texte)

| Distribution des proportions pour les grappes[a] | Proportion pondérée globale | Nombre prévu positif | Méthode de calcul des limites de l'intervalle | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linéaire | | Logit | | Breeze | | Proposée | |
| | | | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. |
| (1) | | | | | | | | | | |
| 0,1 | 0,1 | 191,0 | 4,3 | 5,9 | 5,1 | 4,9 | 4,2 | 4,4 | 4,6 | 4,6 |
| 0,02 | 0,02 | 36,9 | 3,3 | 7,3 | 5,3 | 4,3 | 4,4 | 4,4 | 4,1 | 4,1 |
| 0,01 | 0,01 | 18,4 | 2,8 | 8,7 | 5,5 | 4,0 | 4,3 | 4,3 | 3,9 | 3,7 |
| 0,0025 | 0,0025 | 4,6 | 1,3 | 18,7 | 6,1 | 4,8 | 3,2 | 4,8 | 2,8 | 4,8 |
| (1/2, 1/2) | | | | | | | | | | |
| 0,05, 0,15 | 0,1 | 191,0 | 5,0 | 5,0 | 6,4 | 3,7 | 5,1 | 3,2 | 5,4 | 3,4 |
| 0,01, 0,03 | 0,02 | 36,9 | 3,0 | 7,9 | 5,4 | 4,5 | 4,3 | 4,6 | 4,0 | 4,3 |
| 0,005, 0,015 | 0,01 | 18,4 | 2,5 | 9,2 | 5,4 | 4,2 | 4,1 | 4,4 | 3,7 | 3,9 |
| 0,00125, 0,00375 | 0,0025 | 4,6 | 1,3 | 19,0 | 6,1 | 4,9 | 3,2 | 4,9 | 2,8 | 4,9 |
| (3/4, 1/4) | | | | | | | | | | |
| 0,05, 0,25 | 0,1 | 191,0 | 4,7 | 5,7 | 7,1 | 4,1 | 5,1 | 3,6 | 5,5 | 3,8 |
| 0,01, 0,05 | 0,02 | 36,9 | 2,6 | 8,9 | 5,2 | 5,2 | 4,0 | 5,3 | 3,7 | 4,9 |
| 0,005, 0,025 | 0,01 | 18,4 | 2,1 | 10,1 | 5,3 | 4,8 | 3,8 | 5,1 | 3,4 | 4,5 |
| 0,00125, 0,00625 | 0,0025 | 4,6 | 1,2 | 19,8 | 5,9 | 5,3 | 3,2 | 5,3 | 2,8 | 5,3 |

(a) Les fractions entre parenthèses indiquent la probabilité que la proportion de grappes pondérée corresponde à la valeur mentionnée.

**Tableau 3**
Simulation de la non-couverture (en pour cent) des limites supérieure et inférieure d'un intervalle de confiance unilatéral de 95 % pour un échantillon de 32 grappes de 100 observations chacune; analyses non pondérées

| Distribution des proportions pour les grappes[a] | Proportion globale | Nombre prévu positif | Méthode de calcul des limites de l'intervalle | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linéaire | | Logit | | Breeze | | Proposée | |
| | | | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. |
| (1) | | | | | | | | | | |
| 0,1 | 0,1 | 320 | 5,0 | 4,9 | 5,7 | 4,2 | 4,9 | 3,8 | 5,2 | 4,1 |
| 0,02 | 0,02 | 64 | 3,8 | 6,3 | 5,2 | 4,5 | 4,7 | 4,8 | 4,4 | 4,4 |
| 0,01 | 0,01 | 32 | 3,5 | 6,8 | 5,6 | 4,4 | 4,7 | 4,4 | 4,3 | 4,0 |
| 0,0025 | 0,0025 | 8 | 2,5 | 8,8 | 5,6 | 3,8 | 4,1 | 3,9 | 3,9 | 3,9 |
| (1/2, 1/2) | | | | | | | | | | |
| 0,05, 0,15 | 0,1 | 320 | 4,5 | 5,6 | 5,6 | 4,2 | 4,5 | 3,7 | 4,8 | 4,0 |
| 0,01, 0,03 | 0,02 | 64 | 3,4 | 7,0 | 5,1 | 4,8 | 4,5 | 4,9 | 4,1 | 4,6 |
| 0,005, 0,015 | 0,01 | 32 | 3,0 | 7,6 | 5,2 | 4,8 | 4,4 | 4,8 | 3,9 | 4,4 |
| 0,00125, 0,00375 | 0,0025 | 8 | 2,2 | 9,2 | 5,4 | 4,3 | 3,8 | 4,3 | 3,5 | 4,3 |
| (3/4, 1/4) | | | | | | | | | | |
| 0,05, 0,25 | 0,1 | 320 | 3,3 | 7,7 | 4,8 | 5,6 | 3,5 | 5,1 | 3,7 | 5,3 |
| 0,01, 0,05 | 0,02 | 64 | 2,9 | 8,1 | 5,1 | 5,2 | 4,1 | 5,3 | 3,8 | 4,9 |
| 0,005, 0,025 | 0,01 | 32 | 2,5 | 9,2 | 4,9 | 5,6 | 3,9 | 5,6 | 3,5 | 5,2 |
| 0,00125, 0,00625 | 0,0025 | 8 | 2,0 | 10,4 | 5,3 | 5,1 | 3,8 | 5,1 | 3,3 | 5,1 |

(a) Les fractions entre parenthèses indiquent la probabilité que la proportion de grappes corresponde à la valeur mentionnée.

**Tableau 4**
Simulation de la non-couverture (en pour cent) des limites supérieure et inférieure d'un intervalle de confiance unilatéral de 95 % pour un échantillon de 32 grappes de 10 observations chacune; poids d'échantillonnage de 1 ou 10 avec probabilité de 1/2 (poids non informatifs)

| Distribution des proportions pour les grappes[a] | Proportion globale | Nombre prévu positif | Méthode de calcul des limites de l'intervalle | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linéaire | | Logit | | Breeze | | Proposée | |
| | | | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. |
| (1) | | | | | | | | | | |
| 0,2 | 0,2 | 64 | 4,0 | 6,6 | 5,2 | 4,7 | 3,1 | 4,7 | 4,2 | 4,3 |
| 0,1 | 0,1 | 32 | 3,2 | 7,8 | 5,3 | 4,4 | 3,6 | 3,8 | 3,9 | 4,0 |
| 0,025 | 0,025 | 8 | 1,7 | 10,2 | 5,5 | 2,1 | 3,4 | 2,1 | 3,2 | 2,4 |
| (1/2, 1/2) | | | | | | | | | | |
| 0,1, 0,3 | 0,2 | 64 | 3,6 | 7,0 | 5,0 | 4,9 | 2,8 | 3,4 | 3,9 | 4,4 |
| 0,05, 0,15 | 0,1 | 32 | 3,0 | 8,1 | 5,1 | 4,6 | 3,4 | 4,0 | 3,7 | 4,2 |
| 0,0125, 0,0375 | 0,025 | 8 | 1,6 | 10,6 | 5,4 | 2,1 | 3,3 | 2,1 | 3,1 | 2,5 |
| (3/4, 1/4) | | | | | | | | | | |
| 0,1, 0,5 | 0,2 | 64 | 3,1 | 7,8 | 4,6 | 5,3 | 2,4 | 3,9 | 3,3 | 4,8 |
| 0,05, 0,25 | 0,1 | 32 | 2,5 | 9,2 | 4,8 | 5,2 | 3,0 | 4,6 | 3,3 | 4,8 |
| 0,0125, 0,0625 | 0,025 | 8 | 1,5 | 11,5 | 5,3 | 2,4 | 3,2 | 3,5 | 3,0 | 2,8 |

(a) Les fractions entre parenthèses indiquent la probabilité que la proportion de grappes corresponde à la valeur mentionnée.

**Tableau 5**

Simulation de la non-couverture (en pour cent) des limites inférieure et supérieure d'un intervalle de confiance unilatéral de 95 % pour un échantillon de 32 grappes de 10 ou de 100 observations chacune avec probabilité de 1/2; poids d'échantillonnage de 1 ou 10 avec probabilité de 1/2 (poids non informatifs)

| Distribution des proportions pour les grappes[a] | Proportion globale | Nombre prévu positif | Méthode de calcul des limites de l'intervalle | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linéaire | | Logit | | Breeze | | Proposée | |
| | | | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. | Infér. | Sup. |
| (1) | | | | | | | | | | |
| 0,1818 | 0,1818 | 320 | 5,1 | 6,0 | 5,7 | 5,2 | 4,2 | 4,1 | 5,2 | 5,0 |
| 0,0364 | 0,0364 | 64 | 4,1 | 7,6 | 5,7 | 5,2 | 5,0 | 5,2 | 4,8 | 4,9 |
| 0,0182 | 0,0182 | 32 | 3,4 | 8,5 | 5,7 | 5,0 | 4,7 | 5,1 | 4,4 | 4,7 |
| 0,0045 | 0,0045 | 8 | 2,0 | 12,7 | 5,9 | 3,4 | 4,0 | 4,3 | 3,6 | 3,8 |
| (1/2, 1/2) | | | | | | | | | | |
| 0,0909, 0,2727 | 0,1818 | 320 | 5,0 | 6,4 | 6,1 | 4,8 | 4,2 | 3,6 | 5,2 | 4,4 |
| 0,0182, 0,0545 | 0,0364 | 64 | 3,9 | 8,1 | 6,0 | 5,1 | 4,9 | 5,0 | 4,7 | 4,8 |
| 0,0091, 0,0273 | 0,0182 | 32 | 3,1 | 9,3 | 5,8 | 5,2 | 4,5 | 5,3 | 4,2 | 4,9 |
| 0,0023, 0,0068 | 0,0045 | 8 | 1,8 | 13,2 | 5,9 | 3,6 | 3,9 | 4,5 | 3,5 | 4,0 |
| (3/4, 1/4) | | | | | | | | | | |
| 0,0909, 0,4545 | 0,1818 | 320 | 3,1 | 9,9 | 4,6 | 7,6 | 2,5 | 6,3 | 3,3 | 7,1 |
| 0,0182, 0,0909 | 0,0364 | 64 | 2,8 | 10,9 | 5,3 | 7,3 | 3,9 | 7,3 | 3,7 | 7,0 |
| 0,0091, 0,0455 | 0,0182 | 32 | 2,4 | 11,5 | 5,4 | 6,8 | 3,9 | 6,9 | 3,6 | 6,5 |
| 0,0023, 0,0114 | 0,0045 | 8 | 1,6 | 14,5 | 5,7 | 4,0 | 3,7 | 5,0 | 3,3 | 4,4 |

(a) Les fractions entre parenthèses indiquent la probabilité que la proportion de grappes pondérée corresponde à la valeur mentionnée.

La limite supérieure de l'intervalle de confiance linéaire tombe plus en-dessous de sa valeur véritable que le niveau nominal de 5 %, même avec un compte prévu de jusqu'à 320 événements positifs, en particulier quand il y a une corrélation positive dans la grappe (tiers intermédiaire et inférieur du tableau), ce qui est quelque peu étonnant. Dans le cas de l'intervalle de confiance de la transformation par logit, la non-couverture semble dépasser légèrement le niveau nominal, surtout pour les limites inférieures. Les limites des intervalles de Breeze et des intervalles proposés paraissent généralement conservatrices. Celles des intervalles binomiaux avec échantillonnage aléatoire simple ne conviennent pas aux cas simulés au tableau 1, à cause des poids d'échantillonnage et de la corrélation dans la grappe (dans les deux tiers inférieurs du tableau). Une non-couverture de plus de 8 % pour les limites inférieure et supérieure de l'intervalle binomial pour tous les cas du tableau, en est la preuve (résultats non indiqués).

Discuter de la «longueur» des intervalles de confiance unilatéraux s'avère un peu compliqué aussi nous restreindrons-nous à parler de la longueur des intervalles bilatéraux à 90 %. Pour l'ensemble des simulations du tableau 1, les intervalles de Breeze et ceux que nous proposons ont une largeur moyenne de 3,3 % et de 4,9 % supérieure à celle des intervalles de transformation par logit.

Le tableau 2 présente les résultats d'une simulation semblable à celle du tableau 1, si ce n'est que les poids de l'échantillon sont informatifs. Pour cela, on a fixé le poids à 10 avec probabilité de 2/3 si l'événement était positif et de 1/3 dans le cas contraire, sinon le poids était de 1. La probabilité que chaque grappe comprenne un événement positif a été ajustée à la baisse jusqu'à ce que les proportions pondérées générales soient identiques à celles du tableau 1. Les résultats du tableau 2 ressemblent à ceux du tableau 1 mais les intervalles linéaire et à logit ont tendance à donner de moins bonnes probabilités de couverture.

Au tableau 3, on peut voir les résultats d'une simulation identique à celle du tableau 1, à l'exception du fait que l'analyse n'est pas pondérée. Comme le tiers supérieur du tableau 3 n'indique aucune corrélation entre les éléments des grappes, on pourrait aussi bien recourir aux limites binomiales de l'échantillon aléatoire simple. Si on effectue la moyenne des quatre possibilités dans ce tiers du tableau, on constate que les limites proposées donnent une largeur de 2,5 % supérieure à celle de l'intervalle binomial (résultats non indiqués). Puisque l'effet véritable du plan de sondage dans le tiers supérieur du tableau 3 est égal à 1,0, les simulations permettent d'examiner les conséquences de la troncation de $n_{df}^{*}$ dans la méthode proposée. (Il y a rarement troncation dans les simulations du tableau 1, car l'incidence véritable du plan de sondage est toujours >1.) Lorsqu'on applique la méthode suggérée avec troncation, la simulation donne des intervalles plus larges et plus conservateurs que ceux du premier tiers du tableau 3. Après calcul de la moyenne des quatre cas, on se

rend compte que les limites proposées créent un intervalle de 4,0 % plus large avec troncation, comparativement à sans (résultats non indiqués pour les limites tronquées).

Le tableau 4 montre les résultats d'une simulation identique à celle du tableau 1, sauf pour le nombre d'observations par grappe, qui passe de 100 à 10. Les résultats ressemblent fort à ceux du tableau 1 pour les simulations comprenant le même nombre prévu d'événements positifs. Seuls les intervalles de Breeze, plus conservateurs que ceux de la méthode préconisée, font exception à la règle. On le doit au fait que la proportion globale est plus importante au tableau 4 qu'au tableau 1, pour le même nombre d'événements prévus (l'échantillon du tableau 4 étant plus petit). Les intervalles de Breeze modifiés par l'approximation de Poisson ne donnent pas de bons résultats avec les grandes proportions. Pour le vérifier, nous avons effectué une simulation identique à celle du tiers supérieur du tableau 1 mais en prenant $p$ = 0,5 comme proportion pour 1 600 événements positifs prévus. Les pourcentages inférieur et supérieur de non-couverture était de 1,2 % et 1,3 %, respectivement, contre 4,6 % et 4,7 % avec la méthode proposée. En moyenne, les intervalles de Breeze sont 37 % plus larges que les intervalles proposés.

Les intervalles de Breeze laissent aussi à désirer avec un très petit nombre de grappes, car ils négligent le nombre de degrés de liberté dans l'estimation de la variance. Nous avons, par exemple, effectué une simulation correspondant au premier tiers du tableau 1 en ne prenant les données que de 8 grappes (100 observations par grappe), et $p_i \equiv 0{,}1$, de telle sorte que le nombre d'événements positifs prévus se situait à 80. Les pourcentages inférieur et supérieur de non-couverture s'établissaient à 6,1 % et à 5,4 %, respectivement, pour les intervalles de Breeze, comparativement à 4,7 % et à 4,0 % pour la méthode préconisée.

Le tableau 5 donne les résultats d'une simulation identique à celle du tableau 1, mais avec des grappes de 10 ou de 100, selon une probabilité de 1/2. Les probabilités de non-couverture dépassent la valeur nominale de 5 % du tiers inférieur du tableau, peu importe la méthode employée. En outre, les intervalles à logit ne se comportent pas aussi bien qu'au tableau 1 dans les deux premiers tiers du tableau.

On a procédé à une dernière simulation reproduisant deux grappes (de 50) pour chaque strate. Le nombre d'événements positifs prévu était identique à celui du tableau 1, les poids ont été fixés au hasard à 1 ou à 10 et la probabilité d'un événement positif variait avec les strates, cela afin d'imiter une corrélation à l'intérieur de la grappe. Les résultats (non indiqués) se rapprochaient beaucoup de ceux du tableau 1.

## 4. APPLICATIONS

Nous examinerons maintenant deux applications caractérisées par un petit nombre d'événements positifs. La première concerne l'estimation de l'infection par le VIH au sein d'une population non sélectionnée. Le nombre de réagissants est peu élevé, car le taux d'infection par le VIH est relativement faible. Dans le deuxième cas (estimation de la cocaïnomanie), les taux sont élevés mais le nombre d'événements positifs reste faible parce que l'analyse ne s'applique qu'à de petits sous-groupes. Dans les deux cas, on a utilisé SUDAAN (Shah, Barnwell et Bieler 1995) pour calculer l'erreur-type (attribuable au plan de sondage) des proportions et la fonction «FINV» du SAS (1990, p. 547) pour établir les quantiles de la distribution $F$ en (1.2).

### 4.1 Estimation de la séroprévalence du VIH dans le cadre de la troisième Enquête nationale sur la santé et la nutrition (NHANES III)

L'enquête NHANES III s'est déroulée de 1988 à 1994 et couvrait les membres de la population civile non intitutionnalisés de 2 mois et plus aux États-Unis (National Center for Health Statistics 1994). Les participants de 18 ans et plus ont subi un test de dépistage du VIH. McQuillan, Khare, Karon, Schable et Vlahov (1997) se sont penchés sur la prévalence du virus chez les personnes de moins de 60 ans et dans divers sous-groupes, dont certains sont reproduits au tableau 6. Sur les 11 202 personnes testées, 59 portaient le virus. La prévalence estimée au tableau 6 (0,32 %) s'écarte considérablement de la proportion non pondérée (0,53 % = 59/11 202) parce qu'elle correspond à une proportion pondérée par les poids de l'échantillon. Les tests de dépistage préservant l'anonymat, les poids d'échantillonnage ont été dérivés des poids originaux de l'échantillon de la NHANES III pour les personnes testées présentant les mêmes caractéristiques, soit mêmes îlot (lieu de l'enquête), race ou groupe ethnique, sexe et groupe d'âge (18-39 et 40-59 ans) (communication personnelle de M. Khare). Le pseudo-plan de sondage utilisé pour estimer la variance consistait à prélever 2 pseudo-UPE dans chacune des 23 strates (communication personnelle de M. Khare), ce qui ne correspond pas au pseudo-plan de sondage typique servant à estimer la variance dans la NHANES III.

Les intervalles de confiance linéaires de 90 % s'appliquant à la prévalence du VIH dans les différents groupes, au tableau 6, sont décalés à gauche et sont plus étroits que les autres intervalles, qui se ressemblent. Les intervalles proposés sont légèrement plus larges que ceux de Breeze ou à logit. La taille efficace de l'échantillon établie en (2.1) est nettement plus faible que celle de l'échantillon, à cause des effets du plan de sondage; les intervalles de confiance obtenus après troncation seront donc identiques à ceux du tableau 6. L'écart entre $n^*$ et $\hat{n}_{df}$ s'avère relativement faible. Pour cet événement positif relativement rare, les simulations décrites à la partie 3 donnent à penser que les intervalles de confiance de Breeze et ceux proposés maintiennent mieux la probabilité de couverture nominale de 90 % que les autres intervalles.

**Tableau 6**
Séroprévalence du VIH chez les adultes de 18 à 59 ans selon la troisième Enquête nationale sur la santé et la nutrition

| | Total | Sexe | | Race/ascendance | | |
|---|---|---|---|---|---|---|
| | | Masculin | Féminin | Blanc | Noir | Mex.-Amér. |
| Taille de l'échantillon | 11202 | 5142 | 6060 | 4128 | 3579 | 3495 |
| Nombre infectés | 59 | 44 | 15 | 9 | 38 | 12 |
| Prévalence (%) ± ET | 0,320 ± 0,076 | 0,519 ± 0,130 | 0,127 ± 0,053 | 0,203 ± 0,071 | 1,100 ± 0,247 | 0,368 ± 0,134 |
| Taille efficace de l'éch. | | | | | | |
| $n^*$ | 5588 | 3056 | 4433 | 3976 | 1779 | 2039 |
| $n^*_{df}$ | 5148 | 2816 | 4084 | 3664 | 1640 | 1880 |
| Int. conf. linéaire 90 % | (0,19, 0,45) | (0,30, 0,74) | (0,04, 0,22) | (0,08, 0,33) | (0,68, 1,52) | (0,14, 0,60) |
| Int. conf. logit 90 % | (0,21, 0,48) | (0,34, 0,80) | (0,06, 0,26) | (0,11, 0,37) | (0,75, 1,62) | (0,20, 0,69) |
| Int. conf. Breeze 90 % | (0,21, 0,48) | (0,32, 0,79) | (0,05, 0,26) | (0,10, 0,37) | (0,73, 1,61) | (0,18, 0,68) |
| Int. conf. proposé 90 % | (0,20, 0,48) | (0,32, 0,80) | (0,05, 0,26) | (0,10, 0,37) | (0,71, 1,63) | (0,17, 0,69) |

**Tableau 7**
Adultes de 12 à 44 ans qui ont essayé la cocaïne d'après la population d'hispanophones de 16 ans et plus de scolarité échantillonnée dans le cadre de l'Enquête sur la santé et la nutrition des personnes d'origine hispanique

| | Total | Sexe | |
|---|---|---|---|
| | | Masculin | Féminin |
| Taille de l'échantillon | 123 | 69 | 54 |
| A essayé | 13 | 10 | 3 |
| Proportion (%) ± ET | 11,6 ± 2,5 | 14,3 ± 3,4 | 7,0 ± 4,8 |
| Taille efficace de l'éch. | | | |
| $n^*$ | 1671 | 1050 | 282 |
| $n^*_{df}$ | 1328 | 844 | 229 |
| Int. conf. linéaire 90 % | (7,0, 16,2) | (8,0, 20,7) | (-1,9[a], 15,9) |
| Int. conf. logit 90 % | (7,8, 17,1) | (9,1, 21,9) | (1,9, 22,8) |
| Int. conf. Breeze 90 % | (7,7, 17,0) | (8,3, 23,2) | (0,9, 24,8) |
| Int. conf. proposé 90 % | (7,4, 17,2) | (8,5, 22,1) | (0,9, 22,7) |
| Avec troncation | | | |
| Int. conf. linéaire 90 % | (6,3, 17,0) | (6,5, 22,2) | voir ci-dessus |
| Int. conf. logit 90 % | (7,2, 18,2) | (8,1, 24,1) | " |
| Int. conf. Breeze 90 % | (7,1, 18,1) | (7,7, 24,4) | " |
| Int. conf. proposé 90 % | (7,2, 17,5) | (8,0, 23,2) | " |

(a) En pratique, ces intervalles auraient pour limites (0, 15,9) puisque les proportions ne peuvent être négatives.

### 4.2 Cocaïnomanie chez les diplômés de collège échantillonnés lors de l'Enquête sur la santé et la nutrition des personnes d'origine hispanique (HHANES)

La HHANES s'est déroulée en 1982-1983 et portait sur trois groupes hispanophones des États-Unis (National Center for Health Statistics 1985). Nous ne nous sommes intéressés qu'à l'échantillon Mexicain-Américain. On a demandé aux personnes de 12 à 44 ans à quel âge elles avaient essayé de la cocaïne pour la première fois. Les réponses possibles étaient l'âge du répondant lorsqu'il avait essayé de la cocaine pour la première fois (en années), «n'a jamais essayé» et «ne sait pas». Nous avons tenté d'estimer la proportion de ceux qui avait déjà fait l'expérience de la drogue parmi les personnes comptant 16 années ou plus de scolarité (pour lesquelles la réponse «ne sait pas» n'existait pas).

Sur les 123 personnes échantillonnées, 13 avait déjà absorbé de la cocaïne, soit une proportion de 11,6 % après pondération (tableau 7). L'erreur-type attribuable au plan de sondage (2,5 %) n'est estimée qu'avec 8 degrés de liberté parce qu'on peut reproduire approximativement le plan de sondage de l'enquête en prélevant 2 UPE dans chacune des 8 strates (Kovar et Johnson 1986). La taille efficace de l'échantillon s'établit à $n^* = 167{,}1$ et $n^*_{df} = 132{,}8$, soit plus que l'échantillon, dans les deux cas. On le doit à l'effet estimé du plan de sondage (0,73, de sorte que $n^* = 123/0{,}736 = 167{,}1$). (Le deuxième facteur de (2.2) est égal à 0,794.) Malgré la stratification, nous

croyons que l'effet véritable du plan de sondage dépasse 1 pour l'enquête, en raison des grappes et de la pondération de l'échantillon. (Il est difficile d'estimer l'effet du plan de sondage étant donné le nombre restreint de degrés de liberté.) C'est pourquoi les méthodes de troncation paraissent raisonnables pour cette application, à notre avis.

En raison des degrés de liberté peu nombreux et parce que l'usage de la cocaïne n'est pas rare, on note plus de variations entre les intervalles de confiance à logit, de Breeze et proposés, au tableau 7. Les simulations décrites à la partie 3 nous portent à recommander les intervalles de confiance proposés (avec troncation).

Notre approche pourrait donner l'impression d'être légèrement incohérente dans le cas de cette enquête, en ce sens que nous tolérons des échantillons d'une taille efficace estimée inférieure à la taille de l'échantillon, et les tronquons davantage. Nous pensons toutefois qu'il s'agit d'une approche raisonnablement conservatrice lorsqu'on soupçonne que l'effet véritable du plan de sondage dépasse 1.

## 5. DISCUSSION

Quoique les intervalles de confiance proposés présentent une couverture probable adéquate pour toutes les simulations ou presque, on ne peut garantir ce résultat pour l'ensemble des configurations de la population (voir le dernier tiers du tableau 5). Il est facile d'illustrer une non-couverture plus importante. Supposons, par exemple, que la population se compose de grappes de 100 éléments et que 10 % des grappes ne renferment que des éléments positifs et les 90 % restants, aucun. En prélevant 10 grappes par échantillonnage aléatoire simple et en prenant toutes les unités des grappes retenues, dans 35 % (= $(1\text{-}0,1)^{10}$) des cas, l'échantillon de 1 000 ne comprendra aucune unité positive. Dans une situation de ce genre, les intervalles proposés se résument aux intervalles binomiaux usuels, c'est à-dire la limite supérieure de l'intervalle de confiance de 95 % est 0,003 ( = $1\text{-}0,05^{1/1000}$). Il s'ensuit que la limite supérieure de l'intervalle de confiance de 95 % est inférieure à la valeur véritable de 0,10 dans au moins 35 % des cas, signe d'une sérieuse non-couverture.

Dans les cas d'échantillonnage simples, on peut bâtir des intervalles de confiance qui respectent au moins la probabilité de couverture nominale en tenant compte de tous les agencements possibles de population et en retenant la configuration la moins favorable pour calculer la couverture probable. Dans le cas hypothétique de l'échantillon à un degré mentionné ci-dessus, par exemple, la limite binomiale calculée à partir de 0 unité positive sur 10, soit 0,26 (=$1\text{-}0,05^{1/10}$) pourrait servir de limite supérieure à l'intervalle de confiance de 95 %. Gross et Frankel (1991) ont étudié de tels intervalles, dont le calcul peut devenir très laborieux, et suggèrent eux aussi des approximations réclamant des opérations moins nombreuses.

Comparativement aux approches de ce genre, les intervalles proposés ont pour avantage (1) d'être faciles à calculer, (2) de se plier aux plans d'échantillonnage complexes, y compris à l'ajustement des poids pour la non-réponse et par stratification a posteriori, (3) de généralement garder leur probabilité de couverture nominale, (4) d'être moins conservateurs que les intervalles qui préservent leur couverture nominale sans égard à la configuration de la population et (5) de présenter des propriétés plus intéressantes que les intervalles linéaires, à logit ou de Breeze. Les conclusions (2) et (5) reposent sur les résultats des simulations qui, il va de soi, ne couvrent pas toutes les situations imaginables. Il conviendrait d'entreprendre d'autres recherches en la matière.

## REMERCIEMENTS

Les auteurs désirent remercier M. Khare pour leur avoir fourni les estimations sur la prévalence et l'erreur-type attribuable au plan de sondage qui s'y rapporte apparaissant au tableau 6, ainsi que le rédacteur associé et les examinateurs pour leurs précieux commentaires.

## BIBLIOGRAPHIE

AGRESTI, A. (1990). *Categorical Data Anlaysis*. New York: Wiley.

BREEZE, E. (1990). *General Household Survey: Report on Sampling Error*. London: Her Majesty's Stationery Office (Office of Population Censuses and Surveys).

CLOPPER, C.J., et PEARSON, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413.

COCHRAN, W.G. (1977). *Sampling Techniques*. Troisième édition. New York: Wiley.

FRANKEL, M.R. (1971). *Inference from Survey Samples: An Empirical Investigation*. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

GRAUBARD, B.I., et KORN, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, 5, 263-281.

GROSS, S.T., et FRANKEL, M.R. (1991). Confidence limits for small proportions in complex samples. *Communications in Statistics – Theory and Methods*, 20, 951-975.

GUENTHER,, W.C. (1983). Hypergeometric distributions. Dans *Encyclopedia of Statistical Sciences*, Volume 3, (Éds. S. Kotz et N.L. Johnson). New York: Wiley, 707-712.

JENNINGS, D.E. (1987). How do we judge confidence-interval adequacy? *American Statistician*, 41, 335-337.

JOHNSON, N.L., et KOTZ, S., (1970). *Continuous Univariate Distributions – 2*. New York: Wiley.

JOHNSON, N.L., KOTZ, S. et KEMP, A.W. (1993). *Univariate Discrete Distributions*. Deuxième édition. New York: Wiley.

KORN, E.L., et GRAUBARD, B.I. (1990). Simultaneous testing or regression coefficients with complex survey data: use of Bonferroni *t*-statistics. *American Statistician*, 44, 270-276.

KORN, E.L., et GRAUBARD, B.I. (1998). Variance estimation for superpopulation parameters. *Statistica Sinica*, 8, 1131-1151.

KOVAR, M.G., et JOHNSON, C. (1986). Design effects from the Mexican American portion of the Hispanic Health and Nutrition Examination Survey: a strategy for analysts. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-399.

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.

McQUILLAN, G.M., KHARE, M., KARON, J.M., SCHABLE, C.A., et VLAHOV, D. (1997). Update on the seroepidemiology of Human Immunodeficiency Virus in the United States household population: NHANES III, 1988-94. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 14, 355-360.

NATIONAL CENTER FOR HEALTH STATISTICS (1985). Plan and operation of the Hispanic Health and Nutrition Examination Survey, 1982-84. *Vital and Health Statistics* 1(19). Hyattsville, MD: National Center for Health Statistics.

NATIONAL CENTER FOR HEALTH STATISTICS (1994). Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. *Vital and Health Statistics*, 1(32). Hyattsville, MD: National Center for Health Statistics.

PFEFFERMANN, D., et LAVANGE, L. (1989). Regression models for stratified multi-stage cluster samples. Dans *Analysis of Complex Surveys*, (Éds. C.J. Skinner, D. Holt, et T.M.F. Smith). New York: Wiley, 237-260.

SAMHSA (1998). National Household Survey on Drug Abuse: Main Findings 1996. (DHHS Publication No. (SMA) 98-3200). Rockville, MD: SAMHSA.

SAS (1990). *SAS Language: Reference, Version 6.* Première édition. Cary, NC: SAS Institute Inc.

SHAH, B.V., BARNWELL, B.G., et BIELER, G.S., (1995). *SUDAAN User's Manual, Release 6.40.* Research Triangle Park. NC: Research Triangle Institute.

VOLLSET, S.E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, 12, 809-824

## REMERCIEMENTS

*Techniques d'enquête* désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 1998. Un astérisque indique que la personne a participé plus d'une fois.

B.M. Balk, *Statistics Netherlands*
* D.R. Bellhouse, *University of Western Ontario*
J. Bethel, *Westat, Inc.*
P. Biemer, *Research Triangle Institute*
* D.A. Binder, *Statistique Canada*
J.-R. Boudreau, *Statistique Canada*
K. Brewer, *Australian National University*
J.M. Brick, *Westat, Inc.*
A. Brinkley, *U.S. Bureau of the Census*
T.W. Broene, *Energy Information Administration, U.S.A.*
N. Buck, *University of Essex*
P. Cantwell, *U.S. Bureau of the Census*
* R. Chambers, *University of Southampton*
* M. Cohen, *National Center for Education Statistics*
B. Cox, *Mathematica Policy Research*
* J. Denis, *Statistique Canada*
* J.-C. Deville, *Institut national de la statistique et des études économiques*
* P. Dick, *Statistique Canada*
A. Dorfman, *U.S. Bureau of Labor Statistics*
* J.D. Drew, *Statistique Canada*
J. Dufour, *Statistique Canada*
* J. Eltinge, *Texas A&M University*
R. Evans, *University of Nebraska - Lincoln*
M. Feder, *University of Southampton*
* W.A. Fuller, *Iowa State University*
* J. Gambino, *Statistique Canada*
B. Graubard, *U.S. National Cancer Institute*
* R.M. Groves, *University of Maryland*
S.J. Haslett, *Massey University, New Zealand*
L. Hattersley, *Office for National Statistics, U.K.*
* M.A. Hidiroglou, *Statistique Canada*
* D. Holt, *Office for National Statistics, U.K.*
K. Humphreys, *University of Glasgow*
J.-S. Hwang, *Academia Sinica*
W. Jocelyn, *Statistique Canada*
E. Johnson, *Educational Testing Service, U.S.A.*
D. Judkins. *Westat, Inc.*
* G. Kalton, *Westat, Inc.*
B. King, *Case Western Reserve University*
* P.S. Kott, *National Agricultural Statistics Service*
M. Kovačević, *Statistique Canada*
M. Kramer, *U.S. Bureau of the Census*
A. Krieger, *University of Pennsylvania*
* R. Lachapelle, *Statistique Canada*
* P. Lahiri, *University of Nebraska - Lincoln*
M. Larson, *Harvard University*
* P. Lavallée, *Statistique Canada*
J. Lawless, *University of Waterlōō*
G. Lee, *Australian Bureau of Statistics*
* S. Linacre, *Australian Bureau of Statistics*
* R. Little, *University of Michigan*
* S. Lohr, *Arizona State University*
* H. Mantel, *Statistique Canada*
P.L. do Nascimento Silva, *IBGE, Brasil*
* G. Nathan, *Central Bureau of Statistics, Israel*
D. Paton, *Statistique Canada*
* D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
* B. Quenneville, *Statistique Canada*
T.E. Raghunathan, *University of Michigan*
E. Rancourt, *Statistique Canada*
* J.N.K. Rao, *Carleton University*
J.O. Ramsey, *McGill University*
* L.-P. Rivest, *Université Laval*
G. Roberts, *Statistique Canada*
G. Robinson, *Commonwealth Scientific and Industrial Research Organisation*
* I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
* N. Schenker, *University of California - Los Angeles*
* F.J. Scheuren, *George Washington University*
A.J. Scott, *University of Auckland*
* J. Sedransk, *Case Western Reserve University*
* M.P. Singh, *Statistique Canada*
* R. Sitter, *Simon Fraser University*
* C.J. Skinner, *University of Southampton*
K.P. Srinath, *ABT Associates, Inc.*
L. Stokes, *University of Texas - Austin*
* D. Stukel, *Statistique Canada*
A. Théberge, *Statistique Canada*
S. Thivierge, *Statistique Canada*
R. Thomas, *Carleton University*
I. Thomsen, *Statistics Norway*
R. Tortora, *Gallup Organization*
C. Tucker, *U.S. Bureau of Labor Statistics*
* R. Valliant, *U.S. Bureau of Labor Statistics*
* V.K. Verma, *University of Essex*
* P.J. Waite, *U.S. Bureau of the Census*
* J. Waksberg, *Westat, Inc.*
* K.M. Wolter, *National Opinion Research Center*
F. Yu, *Australian Bureau of Statistics*
E. Zanutto, *University of Pennsylvania*
* A. Zaslavsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1998: J. Beauseigle (Division de la diffusion) et L. Perreault (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à D. Blair, C. Larabie, D. Lemire, C. Marleau et G. Ray de la Division des méthodes des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

CONTENTS TABLE DES MATIÈRES

**Volume 26, No. 3, September/septembre 1998**

Geoffrey WATSON
On the role of statistics in the paleomagnetic proof of continental drift

L. BILLARD
Some statistical opportunities in the agricultural sciences

Noel CRESSIE
Transect-spacing design of ice cores on the Antarctic continent

Daniel GERVINI and Victor J. YOHAI
Robust estimation of variance components

David R. BRILLINGER and Brent S. STEWART
Elephant seal movements: modelling migration

Phaik Mooi LEONG and Stephan MORGENTHALER
Mutability of DNA base pairs: A statistical approach based on linear discrimination

Deborah L. HALL, Karen KAFADAR and Alvin M. MALKINSON
Statistical methodology for assessing homology of intronic regions of genes

Raymond J. CARROLL, Laurence S. FREEDMAN, Victor KIPNIS and Li LI
A new class of measurement error models, with applications to dietary data

S. DURHAM, N. FLUORNOY and W. LI
A sequential design for maximizing the probability of a favorable response

Frank HAMPEL
Is statistics too difficult?

**Volume 26, No. 4, December/decembre 1998**

Kiros BERHANE and Robert J. TIBSHIRANI
Generalized additive models for longitudinal data

J.V. ZIDEK, N.D. LE, H. WONG and R.T. BURNETT
Including structural measurement errors in the nonlinear regression analysis of clustered data

J.F. LAWLESS and M. ZHAN
Analysis of interval-grouped recurrent event data using piece-wide constant rate functions

Ming Gao GU and Shaolin LI
A stochastic approximation algorithm for maximum likelihood estimation with incomplete data

Jiahua CHEN
Penalized likelihood ratio test for finite mixture models with multinomial observations

E. SUSKO, J.D. KALBFLEISCH and J. CHEN
Constrained nonparametric maximum likelihood estimation for mixture models

Qiqing YU, Anton SCHICK, Linxiong LI and George Y.C. WONG
Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times

Claude BÉLISLE
Slow convergence of the Gibbs sampler

Kenny CRUMP and Daniel KREWSKI
Estimation of the number of studies with positive trends when studies with negative trends are present

Marlos A.G. VIANA, André ROGATKO and Timothy R. REBBECK
Statistical assessment of multi-valued screening tests

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents
## Volume 14, Number 2, 1998

The Sixth Morris Hansen Lecture: Opening Remarks
*Daniel Levine* ..... 115
The Hansen Era: Statistical Research and Its Implementation at the Census Bureau, 1940 - 1970
*Joseph Waksberg* ..... 119
Discussion
*Margo Anderson* ..... 137
Discussion: The Irresistible Appeal of the Hansen Years at the U.S. Census Bureau
*Robert M. Groves* ..... 143
Sequential Poisson Sampling
*Esbjörn Ohlsson* ..... 149
Estimating the Sampling Variance of the UK Index of Production
*P.N. Kokic* ..... 163
Bias Correction in the Balanced-Half-Sampled Method if the Number of Sampled Units in Some Strata is Odd
*Ger T. Slootbeek* ..... 181
Response Burden and Panel Attrition
*Adriaan W. Hoogendoorn and Dirk Sikkel* ..... 189
London Plague Statistics in 1665
*David R. Bellhouse* ..... 207
**In Other Journals** ..... 235

## Volume 14, Number 3, 1998

The Time Series Analysis of Compositional Data
*Teresa M. Brunsdon and T.M.F. Smith* ..... 237
Linking of Classifications by Linear Mappings
*Beat Hulliger* ..... 255
Determining Sample Sizes for Surveys with Data Analyzed by Hierarchical Linear Models
*Michael P. Cohen* ..... 267
Design and Analysis of Experiments Embedded in Sample Surveys
*Jan van den Brakel and Robbert H. Renssen* ..... 277
Maximizing and Minimizing Overlap When Selecting a Large Number of Units per Stratum Simultaneously for Two Designs
*Lawrence R. Ernst* ..... 297
Variance Estimation Using List Sequential Scheme for Unequal Probability Sampling
*Yves G. Berger* ..... 315
**Book and Software Reviews** ..... 325
**In Other Journals** ..... 335

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

# DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

## 1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

## 2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

## 3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc.
3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; l, 1).
3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. 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.