Statistique Canada Statistics Canada

TECHNIQUES D'ENQUÊTE

UNE REVUE
DE
STATISTIQUE CANADA

VOLUME 13, NUMÉRO 1
JUIN 1987

Canada

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

JUIN 1987

Publication autorisée par
le ministre des Approvisionnements
et Services Canada

©Ministre des Approvisionnements
et Services Canada 1987

Le lecteur peut reproduire sans autorisation des extraits de cette publication à des fins d'utilisation personnelle à condition d'indiquer la source en entier. Toutefois, la reproduction de cette publication en tout ou en partie à des fins commerciales ou de redistribution nécessite l'obtention au préalable d'une autorisation écrite des Services d'édition, Agent de droit d'auteur, Centre d'édition du gouvernement du Canada, Ottawa, Canada K1A 0S9.

Novembre 1987

Prix: Canada, $20.00 par année
Autres pays, $23.00 par année

Paiement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 13, n° 1

ISSN 0714-0045

Ottawa

STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA
1010495178

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

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

COMITÉ DE DIRECTION

Président G.J. Brackstone

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

COMITÉ DE RÉDACTION

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

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*
D.R. Bellhouse, *University of Western Ontario*
L. Biggeri, *Université de Florence*
D. Binder, *Statistique Canada*
E.B. Dagum, *Statistique Canada*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistique Canada*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*
W.M. Podehl, *Statistique Canada*
J.N.K. Rao, *Carleton University*
I. Sande, *Statistique Canada*
C.E. Särndal, *Université de Montréal*
F.J. Scheuren, *U.S. Internal Revenue Service*
V. Tremblay, *Statplus, Montréal*
K.M. Wolter, *U.S. Bureau of the Census*

Rédacteurs adjoints

J. Armstrong, *Statistique Canada*
H. Lee, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistique 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 statistique, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociale, Statistique Canada, 4e étage, Édifice Jean-Talon, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 20.00$ par année au Canada, et de 23.00$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit, soit $10.00 (É.-U.) ($14.00 Can.) est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 13, numéro 1, juin 1987

TABLE DES MATIÈRES

K.M. INGLIS, R.M. GROVES, et S.G. HEERINGA
Plans de sondage d'enquêtes téléphoniques auprès de ménages noirs aux États-Unis 1

W.M. SYKES et M. COLLINS
Comparaison entre l'interview téléphonique et l'interview sur place au Royaume-Uni 19

G.J. BRACKSTONE
Utilisation des dossiers administratifs à des fins statistiques 35

C.A. FRANCISCO, W.A. FULLER, et R. FECSO
Propriétés statistiques des estimateurs de la production végétale 53

E.B. DAGUM
Problèmes courants sur la désaisonnalisation 71

S. KUMAR et A.C. SINGH
Sur l'estimation efficace des taux de chômage à l'aide de données de l'enquête sur la population active 83

A. DEY et A.K. SRIVASTAVA
Méthode d'échantillonnage avec probabilités de sélection proportionnelles à la taille 93

D. DOLSON, K. McCLEAN, J.-P. MORIN, et A. THÉBERGE
Plan d'échantillonnage pour l'enquête sur la santé et les limitations d'activités .. 101

F.C. OKAFOR
Comparaison d'estimateurs de totaux de population obtenus par sondages successifs à deux degrés à l'aide de l'information auxiliaire 119

Rectification 133

Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 1–17
Statistique Canada

Plans de sondage d'enquêtes téléphoniques auprès de ménages noirs aux États-Unis[1]

KATHRYN M. INGLIS, ROBERTS M. GROVES, et STEVEN G. HEERINGA[2]

RÉSUMÉ

Le plan de sondage d'enquêtes téléphoniques à deux degrés avec règles de rejet élaboré par Waksberg en 1978 a été modifié pour améliorer l'efficacité des enquêtes téléphoniques auprès de la population noire aux États-Unis. L'application expérimentale d'autres plans de sondage a montré que: a) l'utilisation d'une stratification approximative fondée sur les noms des circonscriptions téléphoniques et des États; b) l'utilisation au premier degré de définitions de séries de 200 et 400 numéros consécutifs déterminant des grappes de grande taille; et c) l'application de règles de rejet fondées sur la race des ménages sont trois éléments dont la réunion améliore la précision relative d'un échantillon, étant donné des ressources fixes. Des modèles de coûts et des modèles d'erreur ont aussi été examinés pour simuler l'application d'autres plans de sondage.

MOTS CLÉS: Échantillons par CNH; enquêtes téléphoniques; échantillons de population rare.

1. INTRODUCTION

Les enquêtes effectuées auprès de populations rares et qui n'ont pas de base spéciale coûtent souvent plus cher par unité que celles qui visent une population complète. Quand la population rare est un petit sous-groupe d'une population facile à définir, l'échantillon de ce sous-groupe est souvent obtenu par filtrage de la population plus grande. Les enquêtes auprès de ménages de sous-groupes démographiques comme la population noire des États-Unis, procèdent en général de cette façon pour trouver les unités de sondage susceptibles d'être incluses dans l'échantillon; toutefois l'utilisation à grande échelle de techniques de filtrage pour constituer des échantillons de population rare coûte cher par interview. Ces dernières années, des méthodes de sondage par téléphone ont été proposées comme moyens peu coûteux pour échantillonner et interviewer des populations rares. Le coût d'une interview téléphonique est souvent inférieur à celui d'une interview en personne (Groves et Kahn 1979) et lorsqu'il faut absolument appliquer une méthode de filtrage pour déterminer les répondants susceptibles d'entrer dans un échantillon, l'efficacité des interviews téléphoniques en termes de coûts est encore plus marquée. Tout de même, les coûts de filtrage dans les enquêtes téléphoniques auprès de populations rares peuvent être élevés en chiffres absolus.

Dans le présent document, nous présentons diverses façons d'affiner la méthode de filtrage dans les enquêtes téléphoniques pour réduire les coûts tout en obtenant le degré de précision voulu. Pour cela, nous avons examiné divers plans de sondage qui pourraient s'appliquer à des enquêtes téléphoniques auprès de la population noire des États-Unis.

Les essais d'enquête téléphonique décrits ici ont été effectués dans le cadre d'un sondage d'opinion mené auprès des Noirs au moment de la campagne électorale de 1984 pour la présidence des États-Unis afin de connaître leurs attitudes politiques et leurs intentions de vote.

---

[1] Une version revisée d'un article qui a été présenté au congrès de 1985 de l'American Statistical Association. La recherche a été partiellement subventionnée par l'U.S. Bureau of the Census et le Survey Research Center. La discussion ne représente par nécessairement les opinions de ces organismes-là.

[2] Kathryn M. Inglis, McNair Anderson and Associates, Australie. Robert M. Groves et Steven G. Heeringa, Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48106-1248, États-Unis.

L'utilisation de techniques d'échantillonnage et d'interview par téléphone exclut de l'enquête les Noirs faisant partie de ménages qui n'ont pas le téléphone (soit 15% environ de toute la population noire aux États-Unis). Dans la plupart des cas, les personnes non abonnées au téléphone sont plus pauvres et plus jeunes que celles qui vivent dans les ménages abonnés au téléphone (Thornberry et Massey 1983). Dans la mesure où il est possible que les Noirs non abonnés au téléphone n'aient ni les mêmes attitudes politiques ni les mêmes intentions de vote que ceux qui sont abonnés, les estimations produites à partir des résultats de l'enquête risquent de différer des vrais paramètres de l'ensemble de la population noire des États-Unis. Sans chercher à minimiser l'erreur de non couverture associée aux enquêtes téléphoniques auprès de la population noire, nous nous sommes principalement attachés aux différences au niveau de l'efficacité sur le plan des coûts et de l'erreur d'échantillonnage pouvant résulter de l'utilisation d'approches alternatives dans la formation des échantillons de ménages noirs pour des enquêtes téléphoniques.

Les plans de sondage pour enquêtes téléphoniques présentés ici ont été élaborés à partir d'un plan défini par Waksberg (1978). Ce plan de sondage par composition de numéros de téléphone au hasard (CNH), communément appelé méthode de Waksberg-Mitofsky, est un plan d'échantillonnage de numéros de téléphone par grappes à deux degrés. Les numéros de téléphone aux États-Unis comprennent 10 chiffres, soit un code régional à 3 chiffres, un code ou indicatif de central à trois chiffres et un numéro de ligne à 4 chiffres de 0000 à 9999, par exemple 313-764-4424. À la première étape, un échantillon stratifié de numéros de téléphone à dix chiffres est produit au hasard et chaque "numéro primaire" ainsi produit est lié à un bloc de 100 numéros consécutifs (par exemple le numéro 313-764-4424 serait lié à la série de 100 numéros consécutifs 313-764-4400 à 313-764-4499). Si, dans le cas des enquêtes-ménages, il se trouve que le numéro primaire est un numéro de téléphone de ménage valide, alors la série de 100 numéros de téléphone consécutifs dont il fait partie est retenue à la première étape pour une deuxième étape d'échantillonnage. Sinon, sa série de 100 numéros consécutifs est rejetée. Par conséquent, la probabilité de sélection d'une série de 100 numéros consécutifs à la première étape est proportionnelle au nombre de numéros de téléphone de ménage valides dans cette série. À la deuxième étape de l'échantillonnage, des numéros de téléphone de ménage valides sont choisis en nombre égal dans chacune des séries de 100 numéros consécutifs retenues à la première étape. Par conséquent, l'échantillonnage des ménages, à la deuxième étape, est un échantillonnage avec probabilité conditionnelle de sélection inversement proportionnelle au nombre de numéros de téléphone de ménage valides dans chaque série de 100 numéros consécutifs. Le plan produit ainsi un échantillon de numéros de téléphone de ménage avec équiprobabilité de sélection ("epsem sampling") et regroupe ces numéros par grappes de sorte que la proportion du nombre total de numéros de téléphone sélectionnés qui rejoignent un ménage est plus grande que celle qui aurait été obtenue au moyen d'un échantillonnage stratifié par CNH. Par souci de clarté dans l'exposé qui va suivre, nous appellerons les banques de 100 numéros de téléphone consécutifs, unités de sondage du premier degré ou unités primaires d'échantillonage (UPE) du plan de sondage par CNH à deux degrés. Le terme "grappe" est réservé pour désigner l'ensemble comprenant un nombre fixé de numéros de téléphone de ménage valides choisis à partir des UPE à la deuxième étape du plan de sondage.

Dans la présente étude, les modifications apportées au plan de sondage pour réduire les coûts de filtrage prennent trois formes: a) stratification des circonscriptions téléphoniques selon la proportion de Noirs et prélèvement non proportionnel (c'est-à-dire d'une proportion plus élevée) d'une partie de l'échantillon dans les strates à forte densité de population noire; b) utilisation de règles de rejet à deux degrés fondées sur le fait que le numéro de téléphone composé est un numéro résidentiel ou non et sur la race du ménage; et c) augmentation de la taille des UPE (de 100 numéros consécutifs à 200 et 400).

La stratification de la population des ménages abonnés au téléphone selon la race a pour objet d'isoler les circonscriptions (zones de rattachement à un central) à forte proportion d'abonnés de race noire. On applique ensuite à ces strates des fractions d'échantillonnage plus grandes que celles qui sont appliquées aux strates qui ont une plus faible proportion de ménages noirs. Selon ce plan de sondage à répartition non proportionnelle, le nombre total de ménages qu'il faut rejoindre pour obtenir une interview avec un ménage noir admissible est plus petit que le nombre de ménages qu'il faudrait rejoindre si l'on utilisait un échantillon "epsem" de la population des ménages. Les coûts de filtrage qu'entraîne la constitution d'un échantillon de ménages noirs sont ainsi diminués. Dans les échantillons d'enquêtes téléphoniques, l'unité géographique de base est la circonscription correspondant à un central téléphonique et à laquelle un ou plus d'un indicatif à trois chiffres est attribué. En général, le compte des abonnés n'est pas ventilé selon les caractéristiques raciales. Il faut par conséquent utiliser des indicateurs approximatifs des circonscriptions à forte densité de population noire. Les expériences décrites dans le présent document avaient pour objet notamment d'examiner la valeur de ces indicateurs d'approximation.

Blair et Czaja (1982) ont élaboré une variante du plan de sondage par CNH de Waksberg-Mitofsky qui intègre des règles de rejet à deux degrés fondées aussi bien sur le fait que le numéro composé est un numéro résidentiel ou non et sur la race du ménage. Appliquée à la population noire, cette méthode inclut à la première étape seulement des séries de 100 numéros consécutifs dont le numéro primaire a été attribué à un ménage noir et prélève ensuite dans chacune des UPE un nombre fixé d'avance de numéros de téléphone de ménages noirs. Dans une enquête nationale par sondage aux États-Unis, Blair et Czaja ont constaté qu'en utilisant ce plan, le pourcentage de ménages noirs rejoints à l'aide de tous les numéros de téléphone choisis augmentait à 25 pour cent à la deuxième étape, tandis qu'il était de 9 pour cent à la première étape. Étant donné les probabilités de sélection qui se compensent dans les deux étapes, ce plan de sondage "epsem" réduit considérablement le travail de filtrage requis pour obtenir un échantillon de taille donnée de ménages noirs. Une variante analogue des règles de rejet utilisées dans le plan de sondage à deux degrés de Waksberg-Mitofsky a été utilisée dans les expériences décrites dans le présent document.

Dans le plan de sondage de Blair et Czaja, certaines séries de 100 numéros consécutifs de la première étape ne contenaient pas assez de numéros de téléphone de ménages noirs pour produire le nombre d'éléments par grappe (10 dans ce cas-là) qu'il faut pour constituer un échantillon "epsem" de ménages noirs. En outre, ce plan comporte des coûts assez élevés de filtrage au premier degré de sélection; il faut composer plus de 44 numéros primaires pour rejoindre un ménage noir. La solution commune à ces deux problèmes est d'augmenter la taille de l'UPE et de choisir plus d'éléments dans les UPE à la deuxième étape. Les analyses présentées ici ont utilisé des unités de sondage du premier degré de 100, 200 et 400 numéros consécutifs chacune. L'extension de la définition des UPE pour qu'elles comprennent plus de numéros que la norme habituelle de 100 numéros consécutifs est suggérée par les observations sur la répartition des numéros de téléphone à l'intérieur d'un indicatif donné. Voici les principales constantes qui semblent se dégager de nos observations: 1) presque tous les numéros de ménages dont le numéro de téléphone commence par un indicatif donné desservent des unités situées à l'intérieur des limites géographiques de la circonscription téléphonique correspondant à l'indicatif en question; 2) il y a peu de regroupements géographiques par grappes des numéros attribués à l'intérieur des limites géographiques des circonscriptions téléphoniques (c'est-à-dire qu'en général deux ménages voisins n'ont pas deux numéros consécutifs et n'ont pas nécessairement deux numéros comportant le même indicatif); et 3) le pourcentage des numéros rejoignant des ménages varie plus d'une série de 1,000 numéros consécutifs à une autre qu'entre les séries de 100 numéros consécutifs d'une même série de 1,000 numéros consécutifs. Ces constatations sont le fruit de plusieurs années d'enquêtes

téléphoniques effectuées par le Survey Research Center. Elles donnent à penser que l'extension de la définition des UPE de 100 numéros consécutifs à un nombre plus élevé pourrait nous permettre d'utiliser des grappes de numéros secondaires plus grandes sans beaucoup réduire la proportion de ces numéros qui rejoignent des ménages noirs.

2. L'ÉTUDE PILOTE

Dans deux expériences intégrées effectuées dans le cadre d'une enquête pilote, plusieurs plans de sondage différents ont été testés. Un des objectifs de l'étude pilote était de déterminer dans quelle mesure on peut produire des séries de numéros de téléphone à forte densité de numéros d'abonnés noirs au moyen d'une stratification fondée sur les régions administratives et d'une correspondance approximative entre les limites des régions administratives et celles des circonscriptions téléphoniques. Pour cela, trois strates de circonscriptions ont été définies:

1) "À forte densité" - Circonscriptions téléphoniques corespondant aux villes centrales des grandes régions métropolitaines statistiques normalisées (par exemple la ville de Chicago pour la RMSN de Chicago). Cette définition était fondée sur le nom des circonscriptions de ces régions.
2) "À densité moyenne" - Toutes les autres circonscriptions dans certains États choisis du sud des États-Unis (Virginie, Caroline du Nord, Caroline du Sud, Floride, Georgie, Alabama, Mississippi et Louisiane). La très grande majorité des circonscriptions de ces États sont comprises dans un seul État; celles qui desservaient deux États en même temps ont été associées à l'État dont le nom apparaît dans le nom de la circonscription.
3) "À faible densité" - Toutes les autres circonscriptions des États limitrophes des États-Unis.

Un échantillon avec équiprobabilité de sélection de 1,400 combinaisons à six chiffres dont trois pour l'indicatif régional et trois pour l'indicatif du central a ensuite été prélevé systématiquement à partir des 34,389 combinaisons de ce genre qui sont inscrites comme étant en service dans une base qu'on peut acheter de la société American Telephone & Telegraph (AT&T). Des numéros aléatoires à 4 chiffres ont été ajoutés à chacun des radicaux à six chiffres choisis pour produire un échantillon de 1,400 numéros primaires à dix chiffres.

Les résultats de l'étude pilote ont montré que la proportion de numéros de téléphone rejoignant des ménages noirs variait beaucoup selon la strate. On a constaté qu'il fallait faire six fois plus de filtrage pour trouver un ménage noir dans la strate à faible densité que dans la strate à forte densité. (Ce résultat a été confirmé avec plus de précision dans l'étude de production, dont nous parlerons dans la prochaine section).

Un autre objectif de l'étude pilote était de tester l'utilisation de règles de rejet fondées sur la composition raciale des ménages correspondant aux numéros de téléphone de l'échantillon tiré des UPE de différentes tailles et sur la validité/non validité de ces numéros (suivant qu'ils rejoignent ou non un ménage). Pour augmenter la précision des analyses relatives à cet objectif, 500 autres numéros primaires ont été choisis dans les strates à forte et à moyenne densité. Les 1,900 numéros primaires de l'échantillon combiné de l'étude pilote ainsi prélevé ont ensuite été composés pour repérer les ménages noirs. Quand un numéro primaire échantillonné rejoignait un ménage noir, il correspondait en même temps à trois UPE différentes. Comme on peut le voir au tableau 1, chaque numéro individuel peut être considéré comme faisant partie en même temps d'une série de 100 numéros consécutifs, d'une série de 200 numéros consécutifs et d'une série de 400 numéros consécutifs. Par exemple, le numéro 313-764-4424 fait partie de la série des 100 numéros consécutifs 4400-4499, de la série des 200 numéros consécutifs 4400-4599 et de la série des 400 numéros consécutifs 4400-4799.

Pour tester la possibilité d'augmenter la taille des UPE, les responsables de l'étude pilote ont prélevé des numéros secondaires dans chacune de ces trois séries. La taille des grappes de ménages noirs du second degré a été fixée à 3 pour les grappes constituées à partir de chaque série de 100 numéros consécutifs associée à un numéro primaire, à 6 pour les grappes constituées à partir de chaque série de 200 numéros consécutifs et à 9 pour les grappes constituées à partir de chaque série de 400 numéros consécutifs. Au premier comme au second degré de sélection, on supposait, si la race du ménage n'était pas connue, qu'il ne s'agissait pas d'un ménage noir.

Le tableau 1 présente la répartition des numéros secondaires selon le type d'UPE et la strate. Ce qui nous intéresse le plus, c'est la proportion de numéros secondaires attribués aux ménages noirs pour les différentes définitions de l'UPE. Pour les séries de 100 numéros consécutifs, .134 de tous les numéros secondaires sont des numéros rejoignant des ménages noirs. Cela entraîne que .223 de tous les ménages échantillonnés étaient noirs, comparativement à la proportion de .25 obtenue par Blair et Czaja. Pour les UPE de 200 numéros consécutifs, .124 de tous les numéros secondaires étaient des numéros rejoignant des ménages noirs, tandis que pour les UPE de 400 numéros consécutifs, .115 de tous les numéros de téléphone prélevés à la deuxième étape étaient attribués à des ménages noirs. L'écart entre ces proportions est toujours inférieur à l'erreur d'échantillonnage (l'écart-type de chaque estimation est d'au moins .02). Cela signifie qu'en portant de 100 à 400 numéros consécutifs la longueur des UPE, on ne diminue pas sensiblement la proportion des numéros rejoignant des ménages noirs (numéros admissibles). Ces taux entraînent qu'étant donné que les UPE de 100 numéros consécutifs permettent d'obtenir en moyenne au second degré des grappes de 13 ou 14 ménages noirs, les séries de 400 numéros consécutifs devraient permettre en moyenne d'obtenir des grappes de 46 ménages noirs. La possibilité d'augmenter la taille des grappes de ménages noirs au second degré d'échantillonnage permet aux chercheurs de réduire considérablement les coûts de filtrage.

Le tableau 1 permet également de comparer les proportions de numéros secondaires admissibles échantillonnés à partir des trois différentes strates utilisées dans l'étude pilote et selon les trois définitions des UPE retenues. Pour toutes les définitions des UPE retenues (100, 200 et 400 numéros consécutifs), on obtient le même résultat - les circonscriptions téléphoniques des grandes RMSN auxquelles correspond la strate à forte densité de population noire produisent des taux d'admissibilité presque deux fois plus élevés que le taux obtenu pour l'ensemble de la population (.21 contre .12 ou .13). La strate à densité moyenne, comprenant les ménages des circonscriptions téléphoniques hors des RMSN de certains États choisis du sud des États-Unis, produit des taux d'admissibilité inférieurs au taux national (soit des taux variant entre .08 et .10). La strate à faible densité, correspondant aux autres États limitrophes des États-Unis, produit également des taux inférieurs à la moyenne (oscillant entre .07 et .085). Comme la strate à forte densité couvre environ 36 pour cent des ménages noirs abonnés au téléphone, la stratification choisie conjuguée avec la répartition non proportionnelle de l'échantillon au premier degré est un moyen efficace pour réduire les coûts de filtrage.

3. L'ÉTUDE DE PRODUCTION

L'étude de production a utilisé le plan de stratification élaboré et testé dans l'étude pilote. Un échantillon de 11,223 numéros primaires répartis de façon non proportionnelle a été tiré des trois strates selon la densité de la population noire à l'aide de fractions d'échantillonnage dans des proportions de 3, 2 et 1 (3 pour la strate à forte densité, 2 pour la strate à densité moyenne et 1 pour la strate à faible densité). Même si l'étude pilote n'a pas révélé

Tableau 1

Étude pilote
Répartition des numéros secondaires choisis dans les séries de 100, 200 et 400 numéros consécutifs par strate

Strate et répartition	Proportion de tous les numéros choisis		
	Séries de 100	Séries de 200	Séries de 400*
Strate à forte densité de Noirs			
Ménages noirs	.205	.201	.214
Ne connaît pas la race	.028	.029	.032
Ménages non noirs	.316	.279	.275
Numéros non résidentiels/non valides	.451	.491	.479
Nombre de cas	(395)	(806)	(1163)
Strate à densité moyenne de Noirs			
Ménages noirs	.104	.080	.076
Ne connaît pas la race	.030	.018	.020
Ménages non noirs	.494	.443	.420
Numéros non résidentiels/non valides	.372	.459	.484
Nombre de cas	(231)	(560)	(878)
Strate à faible densité de Noirs			
Ménages noirs	.085	.084	.069
Ne connaît pas la race	.014	.028	.027
Ménages non noirs	.532	.577	.607
Numéros non résidentiels/non valides	.369	.311	.297
Nombre de cas	(141)	(286)	(491)
Total			
Ménages noirs	.134	.124	.115
Ne connaît pas la race	.024	.025	.026
Ménages non noirs	.442	.431	.448
Numéros non résidentiels/non valides	.400	.420	.411
Nombre de cas	(767)	(1652)	(2532)

* Estimation pondérée pour tenir compte du fait que les 9 numéros secondaires des grappes obtenues à partir des séries de 400 numéros consécutifs ne sont pas répartis proportionnellement entre chacune des 4 tranches de 100 numéros consécutifs.

de différence appréciable entre les taux de validité pour les UPE de 200 et de 400 numéros consécutifs, on a opté pour la prudence en choisissant les séries de 200 numéros consécutifs dans l'étude de production. La taille attendue des grappes au second degré a été fixée à 5.5 ménages noirs (sans compter le numéro primaire). Les règles de rejet au premier et au second degré appliquées dans la version modifiée du plan de sondage à deux degrés de Waksberg-Mitofsky ayant servi dans l'étude pilote ont aussi été appliquées dans l'étude de production. Comme des tailles d'échantillon beaucoup plus grandes ont été utilisées dans l'étude de production, on peut répondre aux questions concernant la précision et l'efficacité relative du plan utilisé avec plus d'assurance.

Le tableau 2 présente les résultats du filtrage des numéros primaires et secondaires effectué dans l'étude de production. D'après les taux estimatifs pondérés non biaisés obtenus suivant un plan de sondage "epsem" par CNH à deux degrés, 13 pour cent de tous les numéros

secondaires étaient des numéros de ménages noirs (l'écart-type pour cette estimation est de .6 pour cent). Cela est très proche du taux d'admissibilité des numéros secondaires de 12 pour cent enregistré dans l'étude pilote. La comparaison des résultats obtenus au premier degré de sélection avec les résultats obtenus au second degré de sélection montre les gains importants qu'on peut réaliser en utilisant un plan de sondage à deux degrés pour faire des enquêtes téléphoniques auprès des ménages noirs. Les gains à tirer d'un plan de sondage à deux degrés sont le plus marqués dans la strate à faible densité de population noire, où la proportion de ménages rejoints augmente de presque neuf fois entre le premier et le second degré (de .011 à .090). Dans la strate à forte densité, l'augmentation est plus proche du double (.072 à .190). Quant au cas où le plan de sondage utilisé est un plan de sondage à répartition non proportionnelle, les proportions non pondérées de ménages noirs rejoints aux deux degrés de sélection sont de 3 pour cent (au premier degré) et de 15 pour cent (au second degré). Si l'on compare ces chiffres aux estimations obtenues selon le plan de sondage "epsem" (c'est-à-dire à 2 et à 13 pour cent), on constate que les coûts de filtrage sont moins élevés dans le premier cas.

Tableau 2

Étude de production
Répartition des numéros choisis selon la strate

Strate et répartition	Primaires	Secondaires
Strate à forte densité de Noirs		
Ménages noirs	.072	.190
Ne connaît pas la race	.035	.027
Ménages non noirs	.219	.352
Numéros non résidentiels/non valides	.674	.431
Nombre de cas	(3,128)	(6,671)
Strate à densité moyenne de Noirs		
Ménages noirs	.032	.141
Ne connaît pas la race	.020	.018
Ménages non noirs	.188	.469
Numéros non résidentiels/non valides	.760	.372
Nombre de cas	(1,879)	(2,375)
Strate à faible densité de Noirs		
Ménages noirs	.011	.090
Ne connaît pas la race	.019	.023
Ménages non noirs	.199	.505
Numéros non résidentiels/non valides	.771	.382
Nombre de cas	(6,116)	(3,987)
Plan de sondage "epsem"*		
Ménages noirs	.021	.129
Ne connaît pas la race	.021	.023
Ménages non noirs	.200	.454
Numéros non résidentiels/non valides	.758	.394
Proportion de ménages noirs selon un plan de sondage à répartition non proportionnelle	.031	.150
Nombre de cas	(11,123)	(13,033)

* Estimations pondérées des taux obtenus selon un "plan de sondage epsem". Les coefficients de pondération compensent le fait que des fractions de sondage non proportionnelles ont été utilisées pour choisir l'échantillon de l'étude de production à partir des trois strates de densités différentes.

Comme dans l'étude pilote, le pourcentage de ménages noirs rejoints est différent pour chacune des trois strates, encore que la distinction entre la strate à densité moyenne et la strate à faible densité soit la plus intéressante. Le pourcentage de ménages noirs rejoints à l'aide des numéros secondaires varie approximativement dans des proportions de 2, 1.5 et 1 entre les trois strates. Les trois strates diffèrent également entre elles par la proportion, dans chacune, de numéros de téléphone attribués à des résidences. La strate à forte densité de population noire a proportionnellement plus de numéros secondaires attribués à des unités non résidentielles, ce qui reflète probablement le fait qu'une plus grande partie des circonscriptions téléphoniques de cette strate est urbaine.

Chaque UPE de 200 numéros consécutifs peut être considérée comme deux moitiés d'UPE de 100 numéros consécutifs chacune. Le tableau 3 montre que la proportion de numéros non résidentiels qu'on trouve dans les 100 premiers numéros consécutifs des UPE dans lesquelles le numéro primaire tombait est plus faible que celle qu'on trouve dans l'autre moitié (.378 comparativement à .409), mais cette différence n'est pas statistiquement significative au niveau de .05 (écart-type de .02 environ). De même, la proportion de ménages noirs rejoints à l'aide des numéros secondaires est un peu plus élevée lorsque les numéros secondaires tombaient dans la première moitié de 100 numéros consécutifs du numéro primaire que lorsqu'ils tombaient dans la série de 100 numéros consécutifs adjacente (.133 comparativement à .125). Mais il est peu probable qu'on observe la même différence dans la plupart des cas où cette expérience est répétée. Le tableau 3 donne un autre point de vue sur les résultats du tableau 2; il montre que la réduction observée entre la proportion des ménages admissibles obtenue à partir des séries de 100 numéros consécutifs dans lesquelles le numéro primaire tombe et la proportion obtenue à partir des séries de 100 numéros consécutifs adjacentes est très négligeable.

Le taux d'admissibilité moyen obtenu sur l'ensemble des UPE, qui correspond à la proportion de ménages noirs rejoints, ne devrait pas être le seul critère d'évaluation des plans de sondage. Pour appliquer un plan de sondage "epsem" stratifié, chaque UPE retenue doit avoir un nombre suffisant de ménages noirs pour permettre de fixer le nombre de ménages noirs souhaité au second degré. Par conséquent, la distribution de la proportion des numéros admissibles sur l'ensemble des UPE est également importante. Les histogrammes des figures 1, 2 et 3 montrent la distribution de la proportion des ménages noirs par strate sur l'ensemble des UPE. La stabilité des trois distributions varie parce qu'il y a environ quatre fois plus d'UPE dans la strate à forte densité que dans les deux autres (224 UPE dans la strate à forte densité contre 60 environ dans la strate à densité moyenne et la strate à faible densité). Par ailleurs, la forme de la courbe de distribution semble très différente selon la strate.

Tableau 3

Étude de production
Répartition des numéros secondaires selon qu'ils tombent dans la même série de 100 numéros consécutifs que le numéro primaire ou dans la série de 100 numéros consécutifs adjacente

Situation	Répartition	
	Même série de 100 numéros consécutifs que le numéro primaire	Série de 100 numéros consécutifs adjacente
Ménages noirs	.133	.125
Ne connaît pas la race	.024	.022
Ménages non noirs	.465	.444
Numéros non résidentiels/non valides	.378	.409
Nombre de cas	(6,522)	(6,511)

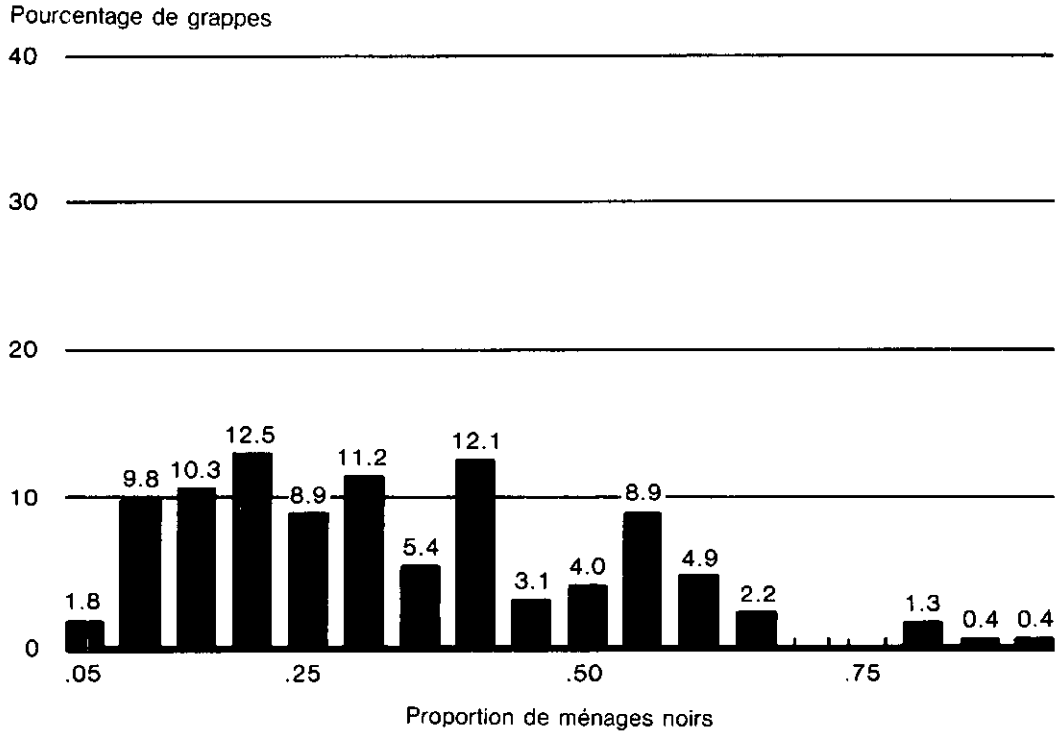

Figure 1. Pourcentage de grappes à forte densité selon la proportion de ménages noirs

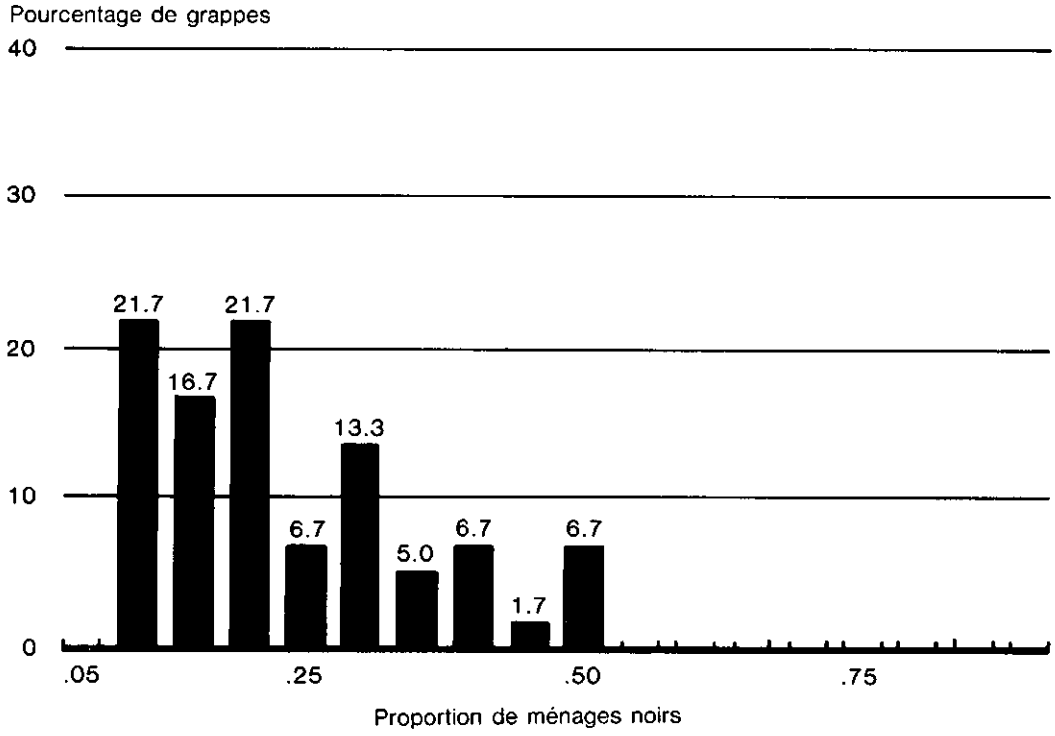

Figure 2. Pourcentage de grappes à densité moyenne selon la proportion de ménages noirs

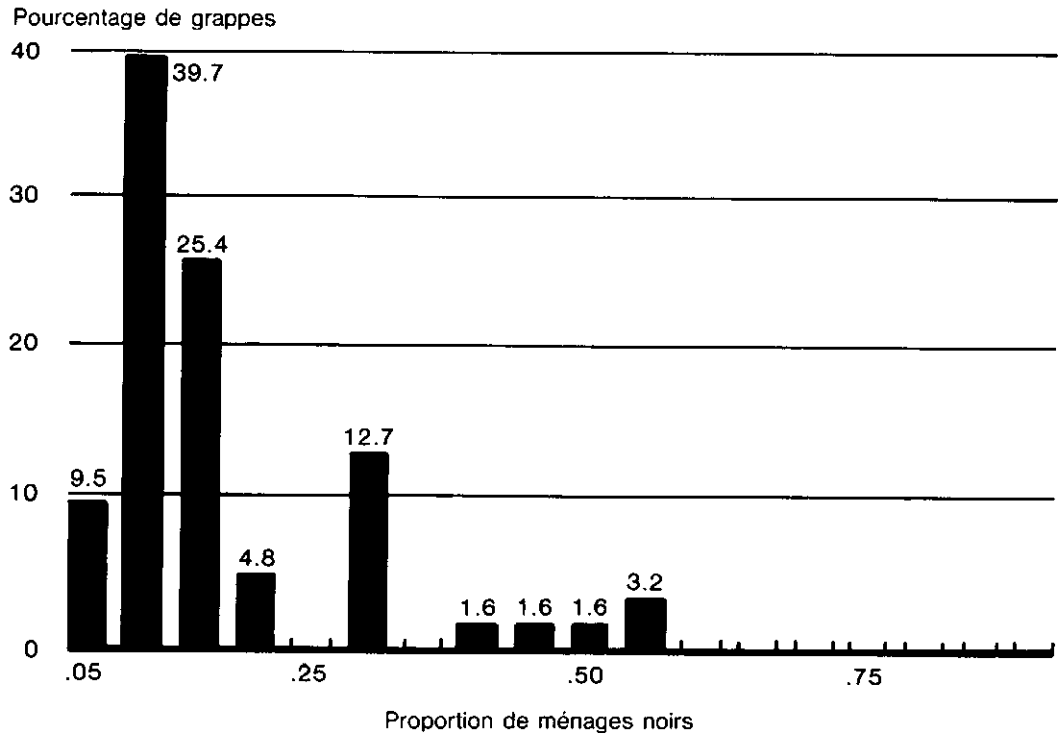

Figure 3. Pourcentage de grappes à faible densité selon la proportion de ménages noirs

La distribution des grappes dans les strates à faible densité et à densité moyenne est très asymétrique, 60 pour cent des UPE de la strate à densité moyenne et 75 pour cent des UPE de la strate à densité moyenne ayant de 5 à 20 pour cent de ménages noirs. Ces taux d'admissibilité correspondent à un maximum de 10 à 40 ménages noirs pour les UPE de 200 numéros consécutifs prélevées dans la strate à faible densité et la strate à densité moyenne. Dans l'étude de production, la strate à faible densité contenait plusieurs UPE qui n'auraient pas permis de former des grappes de cette taille. (On a estimé que 6 des 63 UPE de cette strate avaient moins de 10 ménages noirs.) La distribution des grappes dans la strate à forte densité est beaucoup plus uniforme. (On a estimé que seulement 4 des 224 UPE de cette strate avaient moins de 10 ménages noirs.)

Ces distributions du pourcentage de ménages noirs par UPE méritent plus d'explications. Compte tenu de ce que nous savons déjà de l'attribution des numéros résidentiels aux banques de numéros disponibles, il n'y a pas lieu de croire qu'à l'intérieur d'une circonscription téléphonique donnée (ou d'un indicatif de central donné), il y ait une tendance générale à attribuer des zones résidentielles différentes à des séries de 100 numéros consécutifs différentes. Cela signifie qu'à l'intérieur d'une circonscription desservant aussi bien des ménages noirs que des ménages non noirs, tout semble indiquer qu'il faut presque nécessairement faire l'hypothèse que les numéros sont attribués sans égard pour la race de l'abonné. Autrement dit, à moins que les circonscriptions ne soient subdivisées en zones rattachées à des noeuds de câbles correspondant à des zones résidentielles de ménages noirs, il n'y a pas de raison *a priori* de supposer que les grappes de ménages noirs dans les séries de 200 numéros consécutifs sont grandes. Suivant ce raisonnement, la distribution plus uniforme des grappes dans la strate à forte densité reflète, croyons-nous, le fait que la proportion de ménages noirs a tendance à varier entre les populations d'abonnés au téléphone des différentes circonscriptions téléphoniques composant la strate.

4. CARACTÉRISTIQUES DE LA VARIANCE D'ÉCHANTILLONNAGE

Pour réduire les coûts de l'échantillonnage de ménages noirs par CNH, il est avantageux d'utiliser de grandes grappes de ménages par UPE (c'est-à-dire d'utiliser moins d'UPE pour une taille d'échantillon donnée) et de tirer proportionnellement plus d'UPE des strates composées des circonscriptions où la proportion de ménages noirs abonnés au téléphone varie. Bien que l'utilisation de plus grosses grappes et la répartition non proportionnelle de l'échantillon améliorent l'efficacité en termes de coûts, la précision globale de l'échantillon est diminuée par l'effet de l'utilisation de plus grosses grappes et par les effets du plan de sondage rendus plus marqués par la nécessité d'effectuer une pondération non optimale pour tenir compte des probabilités inégales de sélection des ménages à partir des trois strates selon la densité de la population noire. Kish (1976) a décrit, les effets augmentés du plan de sondage des estimations dans les cas où la pondération n'est pas optimale. Nous montrons dans les paragraphes qui suivent comment l'utilisation de grappes influe les effets du plan de sondage dans la version modifiée de la méthode d'échantillonnage par CNH utilisée dans la présente étude.

Toutes choses étant égales par ailleurs, plus on choisit d'éléments par UPE, plus l'effet du plan de sondage est grand (l'effet du plan de sondage est le ratio de la variance d'échantillonnage d'un plan de sondage donné sur celle d'un échantillon aléatoire simple comprenant le même nombre d'éléments). Pour calculer l'effet du plan de sondage, le modèle le plus souvent utilisé est $Deff = 1 + \rho\ (b - 1)$, où Deff ("Design effect") est l'effet du plan de sondage, ρ, la corrélation intragrappe pour la statistique considérée et b, le nombre d'éléments de l'échantillon par UPE. D'autres études ont montré que, pour beaucoup de variables portant sur l'ensemble de la population aux États-Unis, la corrélation intragrappe observée avec les séries de 100 numéros consécutifs est la plupart du temps inférieure à celle qu'on trouve en général entre les grappes d'échantillons probabilistes aréolaires (voir Groves 1978). Il est possible que cela ne soit pas le cas pour la population noire avec les séries de 100 numéros consécutifs; par ailleurs aucune estimation empirique de la corrélation intragrappe n'a encore été faite avec des séries de 200 numéros consécutifs. L'hypothèse formulée avant d'estimer les erreurs d'échantillonnage était que la corrélation intragrappe devrait avoir la même valeur pour les séries de 100 comme de 200 numéros consécutifs. Cette hypothèse reflète la conception que nous avons formulée plus haut de l'attribution des numéros de téléphone à l'intérieur des circonscriptions téléphoniques.

Suivant les erreurs d'échantillonnage dont la valeur a été estimée à partir d'un ensemble de données de l'étude de production, la valeur moyenne de l'effet du plan de sondage obtenue pour sept statistiques choisies a été de 1.28 avec les séries de 100 numéros consécutifs et de 1.30 avec les séries de 200 numéros consécutifs. La valeur moyenne de l'effet du plan de sondage obtenue pour les séries de 100 numéros consécutifs a été estimée à partir des seuls cas qui tombaient dans la même série de 100 numéros consécutifs que le numéro primaire, tandis que tous les cas ont été utilisés pour calculer la valeur moyenne de l'effet du plan du sondage pour les séries de 200 numéros consécutifs. Ainsi, la taille moyenne des grappes des interviews complétées est de 2.0 dans le cas des séries de 100 numéros consécutifs (coefficient de variation de .043) et de 3.4 pour les séries de 200 numéros consécutifs (coefficient de variation de .029). Ces effets du plan du sondage reflètent la stratification, le regroupement par grappes et la pondération prévues dans le plan de sondage et aussi le fait que la taille des grappes prélevées dans les séries de 100 numéros consécutifs varie plus. (Une des règles de rejet obligeait à prélever le même nombre de ménages noirs dans les séries de 200 numéros consécutifs, mais pas nécessairement le même nombre dans les séries de 100 numéros consécutifs.) Étant donné que les effets du plan de sondage obtenues avec les séries de 100 et les séries de 200 sont très proches l'une de l'autre (1.28 et 1.30), il semble que l'élément

Tableau 4

Étude de production
Corrélations intragrappe synthétiques selon la strate et selon que les grappes sont tirées des séries de 100 ou de 200 numéros consécutifs pour sept statistiques choisies

Statistique	Corrélation intragrappe synthétique*					
	Strate à forte densité de population noire		Strate à densité moyenne de population noire		Strate à faible densité de population noire	
	Séries de 100	Séries de 200	Séries de 100	Séries de 200	Séries de 100	Séries de 200
Proportion très satisfaite de la vie en général	.021	-.002	-.172	-.042	-.238	-.116
Proportion qui estime être dans une meilleure situation financière que l'année précédente	.113	.075	.094	.069	.206	.049
Proportion qui votera pour Mondale	.189	.021	.086	-.087	-.436	-.046
Proportion qui va à l'église	.013	.017	-.009	-.078	.035	-.110
Proportion ayant passé toute sa vie dans la même ville	-.078	.001	.058	.114	.221	.248
Proportion qui a voté aux élections présidentielles de 1980	-.045	-.035	-.101	-.013	.364	.356
Proportion qui pense que Reagan sera élu président	-.045	-.045	-.545	-.078	.124	-.105
Moyenne	.024	.005	-.084	-.016	.039	.039

* Ces estimations ne sont pas pondérées.

déterminant de la variance d'échantillonnage soit la pondération non optimale qu'il faut effectuer pour tenir compte du fait que l'échantillon n'est pas réparti de façon proportionnelle, tandis que l'augmentation de la taille des UPE de 100 à 200 numéros consécutifs ne fait perdre que très peu de précision.

Le tableau 4 présente les résultats du calcul des corrélations synthétiques intragrappe par strate pour les sept statistiques de l'enquête utilisées pour estimer l'effet moyen du plan de sondage. Les estimations des corrélations synthétiques intragrappe ont été obtenues à partir de la valeur de l'effet du plan du sondage à l'aide du modèle de Kish où $Rho = (Deff - 1)/(b - 1)$ et n'ont pas été pondérées pour qu'il n'y ait pas d'effet de confusion de la pondération sur les estimations synthétiques. Les estimations du tableau ont tendance à être instables à cause du petit nombre de grappes prélevées dans chaque strate, de l'assez faible taille moyenne des grappes une fois les interviews complétées et du coefficient de variation associé. Ces caractéristiques du plan de sondage rendent plus difficile pour nous de tirer des conclusions à propos de l'effet du regroupement par grappes dans les séries de 100 et de 200 numéros consécutifs. En général, les estimations de la corrélation intragrappe sont un peu plus élevées dans les séries de 100 que dans les séries de 200. Nous croyons que cela reflète plus l'instabilité de la corrélation synthétique estimative qu'une différence réelle au niveau de l'effet du regroupement par grappes et que ces estimations ne donnent pas beaucoup d'évidence que la corrélation intragrappe soit différente selon qu'on la calcule pour des séries de 100 ou de 200 numéros consécutifs.

5. CARACTÉRISTIQUES OPTIMALES DU PLAN DE SONDAGE

Les sections précédentes portaient sur l'effet de l'utilisation de différents plans de sondage sur l'efficacité en termes de coûts et sur la variance d'échantillonnage. Dans la plupart des enquêtes par sondage, c'est à l'étape de l'élaboration du plan de sondage qu'on compare les coûts d'enquête et les erreurs pour chercher à déterminer quelles devraient être les caractéristiques "optimales" de l'enquête. Suivant cette façon de procéder, on tente d'élaborer le plan de sondage qui permettra d'obtenir la plus petite variance possible pour un ensemble donné de ressources à consacrer à l'enquête. Vu la nature de la présente étude, la recherche du choix optimal portera avant tout sur les deux caractéristiques suivantes du plan de sondage: a) nombre d'éléments par UPE et b) répartition de l'échantillon entre les trois strates définies selon la densité de la population noire.

Pour déterminer la taille optimale des grappes, nous avons utilisé le modèle de coûts totaux suivant: $C = C_o + C_a a + C_b ab$, où C_o représente les coûts fixes, où C_a représente le coût d'échantillonnage et de filtrage par grappe et a le nombre de grappes choisies et où C_b représente le coût d'échantillonnage, de filtrage et d'interview associé à chaque interview obtenue et b le nombre d'interviews obtenues dans chaque grappe. Comme la proportion de ménages noirs varie d'une strate à l'autre, la valeur des paramètres C_a et C_b varie aussi selon la strate (voir tableau 5). La taille optimale des grappes est donnée par $\sqrt{C_a(1-\rho)/(C_b\rho)}$ (Kish 1965). Le tableau 5 présente aussi les résultats des estimations que nous avons faites pour déterminer quelles seraient les tailles optimales des grappes pour les moyennes et les proportions globales et trois valeurs différentes de la corrélation intragrappe, .005, .01 et .02, en utilisant les données sur les coûts tirées de l'enquête de production. (Ces résultats sont comparables à ceux qui ont été obtenus pour les variables d'attitudes politiques et d'intentions de vote dans les enquêtes réelles.) Les estimations des coûts C_a et C_b pour chaque strate figurent également dans le tableau. Le tableau montre que c'est dans la strate à faible densité que la taille optimale des grappes est la plus grande, ce qui reflète le fait que les coûts de filtrage sont élevés dans cette strate. À noter aussi que ces tailles des grappes optimales tendent à être plus grandes que celles qui ont effectivement été utilisées dans l'enquête (soit $b = 6.5$).

À noter encore que la taille optimale des grappes est à peu près la même pour les grappes tirées des UPE de 100 comme de 200 numéros consécutifs et que la perte d'efficacité en termes de coûts par suite de l'utilisation des séries de 200 plutôt que de 100 numéros consécutifs est minime. (Les estimations de la variance d'échantillonnage signifient également que la corrélation intragrappe est à peu près la même, indépendamment du fait que les grappes sont tirées des séries de 100 ou de 200 numéros consécutifs.)

Les valeurs de la taille optimale des grappes figurant dans le tableau 5 excèdent en général la valeur de la taille des grappes que permettent d'obtenir des UPE de 100 numéros consécutifs. Autrement dit, une proportion élevée d'UPE de 100 numéros consécutifs ne contiendrait pas suffisamment de numéros de ménages noirs pour produire des grappes de la taille désignée au second degré. Cette seule raison suffit à faire préférer les séries de 200 numéros consécutifs. Même avec les séries de 200 numéros consécutifs, il n'a pas été possible d'obtenir les grappes de la taille souhaitée au second degré pour toutes les UPE de la strate à faible densité. (Cela donne à penser que la vraie solution au problème du choix de la taille optimale des grappes devrait tenir compte de la capacité des UPE de produire des grappes; par ailleurs, la méthode utilisée ici peut certainement aider à prendre des décisions pratiques au sujet de l'efficacité en termes de coûts, mais elle ne prend pas en considération certaines situations extrêmes.)

Tableau 5

Paramètres de coûts et nombre optimal d'éléments par grappe selon la strate pour les grappes tirées des séries de 100 et de 200 numéros consécutifs et différentes valeurs de ρ

Strate et définition des grappes	Taille optimale des grappes			Paramètres de coûts	
	ρ = .005	ρ = .01	ρ = .02	C_{ha}	C_{hb}
Strate à forte densité					
100	15.9	11.2	7.9	$50.81	$40.11
200	15.9	11.2	7.9		$39.78
Strate à densité moyenne					
100	22.4	15.8	11.1	$114.09	$45.18
200	21.3	15.0	10.6		$50.00
Strate à faible densité					
100	29.8	21.0	14.8	$309.98	$69.52
200	29.9	21.1	14.8		$69.18

La deuxième décision importante qu'il nous a fallu prendre au sujet du plan de sondage avait trait au choix de la façon de répartir les unités de l'échantillon entre les strates. L'enquête utilisait des fractions de sondage dans des proportions de 3, 2 et 1 de la strate à forte densité à la strate à faible densité. Nous avons cherché à déterminer quelle serait la répartition optimale de l'échantillon entre les strates en supposant que les grappes choisies dans chaque strate était de taille optimale (voir tableau 5). Étant donné une taille fixée des grappes dans chaque strate, b_h, nous posons que la fraction de sondage dans la h-ième strate, f_h, est proportionnelle à $\sqrt{(Deff_h S_h^2)/(C_{ha}/b_h)}$, où $Deff_h$ est l'effet du plan de sondage pour la statistique dans la h-ième strate, S_h^2 est la variance de l'échantillon dans la h-ième strate, C_{ha} correspond aux coûts d'échantillonnage et de filtrage des UPE dans la h-ième strate et b_h est le nombre d'unités d'échantillon par grappe dans la h-ième strate.

Le tableau 6 présente les proportions optimales entre les fractions de sondage pour différentes combinaisons de variances dans chacune des trois strates et diverses valeurs de ρ. Le tableau montre que la répartition optimale de l'échantillon entre les strates est relativement peu sensible aux variations de ρ (pour les valeurs de ρ susceptibles d'être utilisées dans le plan de sondage). Quand les deux strates à plus forte densité de ménages noirs ont une variance au moins égale à celle de la strate à faible densité, il est préférable de prélever proportionnellement plus d'unités de sondage dans ces strates. (Cela reflète les coûts beaucoup moins élevés dans ces strates.) Quand l'écart-type est dans des proportions de 1.7, 1.5 et 1 environ entre les trois strates, le meilleur rapport entre les fractions de sondage est de 3, 2 et 1. L'analyse des données obtenues à partir de l'enquête donne à penser que, pour beaucoup de variables, l'écart-type est à peu près dans des proportions de 1, 1 et 1 entre les trois strates. Dans le cas de ces variables, les fractions de sondage, pour être optimales, doivent avoir un rapport de 1.7, 1.4 et 1 pour les tailles optimales des grappes présentées dans le tableau 5. (Pour une taille de grappe de 6.5, soit la taille effectivement utilisée dans chaque strate,

Tableau 6

Répartition optimale de l'échantillon entre les strates pour les moyennes globales, compte tenu de la formation de grappes de taille optimale dans chaque strate, pour diverses ratios des écart-types à l'intérieur des strates et diverses valeurs de la corrélation intragrappe

Ratios des écarts-types à l'intérieur des strates (Forte, moyenne et faible densité)	Ratios des fractions de sondage optimales (Forte, moyenne et faible densité)
ρ = .005	
3 : 2 : 1	5.2 : 2.7 : 1
1.7 : 1.5 : 1	3 : 2 : 1
1 : 1 : 1	1.7 : 1.4 : 1
.33 : .5 : 1	.6 : .9 : 1
ρ = .01	
3 : 2 : 1	5.2 : 2.7 : 1
1.7 : 1.5 : 1	3 : 2 : 1
1 : 1 : 1	1.7 : 1.4 : 1
.33 : .5 : 1	.6 : .9 : 1
ρ = .02	
3 : 2 : 1	5.1 : 2.7 : 1
1.8 : 1.5 : 1	3 : 2 : 1
1 : 1 : 1	1.7 : 1.3 : 1
.33 : .5 : 1	.6 : .9 : 1

les fractions de sondage optimales sont dans des proportions de 2.5, 1.6 et 1). D'après ces deux résultats, on peut dire que la méthode utilisée dans l'étude de production et qui consiste à prélever proportionnellement plus d'une unités d'une strate que d'une autre a entraîné, pour un coût unitaire donné, une perte de précision par rapport à la méthode correspondant à l'utilisation de fractions de sondage optimales.

6. SOMMAIRE

L'échantillonnage de populations rares oblige le statisticien qui veut s'en servir pour faire des enquêtes à prendre en considération diverses combinaisons de définitions des UPE et des grappes, de méthodes de stratification et de méthodes de mesure de la taille qu'on ne trouve pas habituellement dans les échantillons transversaux. Notre recherche nous a amenés à conclure qu'il est possible de modifier les techniques de sondage habituelles pour augmenter l'efficacité des échantillons à deux degrés pour les enquêtes téléphoniques auprès de la population des ménages noirs abonnés au téléphone.

Premièrement, nous avons constaté que même avec une correspondance très approximative entre les limites géographiques des circonscriptions téléphoniques et celles des grandes villes et des États, il a été possible de stratifier l'univers visé par l'enquête en strates qui permettaient de distinguer des groupes de circonscriptions possédant des taux d'admissibilité très différents. La proportion de ménages noirs était plus que deux fois plus élevée dans la strate à forte densité que dans la strate à faible densité. Cela permet

de réduire les coûts de filtrage au moment de l'élaboration du plan de sondage. Avec d'autres populations rares formées de groupes dont la concentration géographique est aussi nette que celle de la population noire des États-Unis, on peut s'attendre aux résultats à peu près analogues.

Deuxièmement, l'utilisation de règles de rejet fondées sur des critères d'admissibilité des unités d'une sous-population réduit effectivement les coûts de filtrage dans les UPE. Cela augmente la proportion admissible de numéros secondaires de deux à neuf fois selon la densité de la strate considérée.

Troisièmement, on ne perd pas beaucoup de numéros admissibles en utilisant des UPE plus grandes (c'est-à-dire des UPE de 200 plutôt que de 100 numéros consécutifs). Les séries de 100 à forte densité de numéros admissibles ont tendance à précéder ou à suivre d'autres séries à forte densité. Cela est un phénomène que nous n'avions pas encore observé concernant l'attribution des numéros par les compagnies de téléphone. Ce fait permet de prélever plus de numéros par UPE, ce qui est un autre élément clé de la réduction des coûts de production de l'échantillon de la population noire.

Malgré qu'on soit fortement tenté de chercher davantage à réduire les coûts dans le cas des enquêtes par sondage auprès de populations rares, il importe d'établir explicitement un juste compromis entre les erreurs et les coûts au moment de décider du plan de sondage définitif. Le modèle de coûts et d'erreurs d'échantillonnage que nous avons adopté pour notre recherche donne à penser qu'il est préférable de ne pas répartir l'échantillon proportionnellement entre les strates de population noire et qu'il est plus efficace de choisir un nombre relativement grand de numéros secondaires par UPE. Ce nombre devrait être suffisamment grand pour obliger à utiliser des UPE de 200 ou de 400 numéros consécutifs.

Même si nous n'avons appliqué ce plan de sondage qu'à la population noire, les résultats qu'il donne devraient être à peu près les mêmes pour d'autres populations à concentration géographique nette. Ces autres populations incluent notamment les groupes selon le revenu, certaines catégories professionnelles et les groupes ethniques.

En outre, les résultats auxquels notre recherche a abouti peuvent également avoir des répercussions sur les méthodes d'échantillonnage transversal. Il se peut qu'il soit avantageux d'augmenter la taille des UPE de 100 à 200 numéros consécutifs dans un plan de sondage par CNH à deux degrés qu'on voudrait appliquer à l'ensemble de la population des ménages abonnés au téléphone. Des séries plus grandes de 200 numéros consécutifs donneraient deux fois plus de numéros parmi lesquels choisir et, comme dans le cas d'enquêtes par sondage auprès de populations rares, la proportion de numéros admissibles aurait tendance à être la même que celle qu'on obtient avec des séries de 100 numéros consécutifs. Par conséquent, avec des valeurs faibles de la corrélation intragrappe, on pourrait décider d'avoir, pour un plan donné, un nombre de numéros admissibles par grappe qui serait beaucoup plus proche de la taille optimale. Parce que toutes les UPE choisies permettraient d'obtenir le nombre choisi de numéros admissibles, la distribution des grappes selon leur taille devrait aussi être moins variable sur l'ensemble des UPE et toute pondération compensatrice devrait donc avoir peu d'effet sur la variance des estimations.

BIBLIOGRAPHIE

BLAIR, J., et R. CZAJA, (1982). Locating a special population using random digit dialing. *Public Opinion Quarterly*, 46, 585-590.

GROVES, R.M., An empirical comparison of two telephone sample designs. *Journal of Marketing Research*, 15, 622-631.

GROVES, R.M. et R.L. KAHN (1979). *Surveys by Telephone*. New York: Academic Press.

KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, Sér. A, 139, 80-95.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

THORNBERRY, O.T., et J.T. MASSEY (1983). Coverage and response in random digit dialed national surveys. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 654-659.

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

Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 19-33
Statistique Canada

Comparaison entre l'interview téléphonique et l'interview sur place au Royaume-Uni

W.M. SYKES et M. COLLINS[1]

RÉSUMÉ

Dans cette étude, les auteurs présentent les résultats des expériences de méthodologie qui ont servie à comparer l'interview téléphonique et l'interview sur place dans les enquêtes menées auprès de la population en général. Le nombre relativement élevé de personnes qui n'ont pas le téléphone au Royaume-Uni, surtout chez les plus défavorisées, est un argument qui milite en faveur de l'application d'une méthode d'interview mixte qui combinerait l'interview téléphonique et l'interview sur place, pour ceux ou celles qui n'ont pas le téléphone. L'efficacité de cette méthode repose sur l'absence, dans les réponses d'enquête, de distorsions attribuables au caractère mixte de la méthode ou sur la possibilité de tenir compte de telles distorsions s'il y en a.

MOTS CLÉS: Interview téléphonique; interview mixte; enquêtes sociales; taux de réponse; qualité des données.

1. INTRODUCTION

Le choix d'une méthode de collecte de données pour une enquête dépend des renseignements dont on dispose sur les diverses méthodes possibles. Au Royaume-Uni, l'interview téléphonique est une méthode qui commence à peine à être connue. Ce n'est que depuis deux ans que l'on compare l'interview téléphonique aux autres méthodes de collecte de données. On peut s'étonner de cette méthode relativement nouvelle vu le caractère animé du débat sur les avantages et les inconvénients de l'interview téléphonique et l'attention qu'a reçue cette question dans d'autres pays.

Le contenu du présent article est fondé principalement sur deux études qui ont été réalisées par le Centre de méthodes d'enquête du Social and Community Planning Research et qui visaient à comparer l'interview téléphonique et l'interview sur place. Réalisées en 1983 et en 1984, ces études traitent quelques-unes des questions fondamentales, à savoir le fait pour le public d'être plus ou moins disposé à participer aux enquêtes par téléphone et le genre, la qualité et la quantité de données qui peuvent être recueillies. Les deux études sont décrites dans la section 2 et leurs résultats présentés dans les sections 3 et 4. Il est également question dans ce document d'une autre étude britannique – une expérience réalisée en 1985 par le Market Research Development Fund – et des activités de recherche méthodologique plus nombreuses qui ont lieu dans d'autres pays, particulièrement les États-Unis.

2. LES ÉTUDES DU SCPR

Nos recherches ont refleté le fait que le pourcentage de ménages qui possédaient un téléphone au Royaume-Uni était faible par rapport à l'Amérique du Nord: en effet, environ 75% des ménages britanniques avaient le téléphone en 1983. La non-exhaustivité est appréciable et elle est déterminante pour les spécialistes des enquêtes sociales parce qu'elle introduit un biais à l'égard des couches moins favorisées de la société britannique. Dans les circonstances, les deux études précitées visaient essentiellement à évaluer la méthode d'interview mixte, selon laquelle les personnes qui ont le téléphone seraient interviewées par téléphone et celles qui ne l'ont pas seraient interviewées sur place.

[1] W.M. Sykes et M. Collins, Survey Methods Centre, SCPR, 35 Northampton Square, Londres, EC1V 0AX, Angleterre.

Deux comparaisons ont été faites dans la première étude: la première entre un échantillon expérimental soumis à l'interview mixte et un échantillon national, plus grand, soumis à l'interview sur place; la seconde entre deux échantillons de personnes qui ont le téléphone, le premier échantillon étant interviewé par téléphone et l'autre sur place. Notre étude porte plus spécialement sur la seconde comparaison, qui soulève la question qui est à la base de toute évaluation de la méthode d'interview mixte: les données obtenues par l'interview téléphonique et l'interview sur place sont-elles compatibles ou y a-t-il des divergences du fait de l'utilisation de deux modes d'interview différents? Dans ce dernier cas, il est impossible "d'additionner" les données et de les considérer comme une seule série sans faire certains rajustements qui, normalement, ne sont pas possibles dans le cadre d'une enquête unique. La seconde étude porte uniquement sur la comparaison directe des deux modes d'interview auxquels sont soumises les personnes qui ont le téléphone.

2.1 Étude n° 1

La première étude a été réalisée concurremment avec l'enquête de 1983 sur les attitudes sociales au Royaume-Uni (1983 British Social Attitudes Survey), qui est désignée ici comme l'enquête 'principale". Cette enquête comportait des interviews sur place d'une durée approximative d'une heure, qui couvraient un large éventail de sujets d'ordre politique, économique, social et moral.

L'échantillon de l'enquête principale était formé d'environ 1750 répondants âgés de 18 ans ou plus et membres de ménages privés. Pour des raisons pratiques, l'échantillon était limité aux personnes dont le nom figurait sur la liste électorale. Les pensionnaires d'institution (à l'exception des ménages privés vivant en institution) étaient exclus du champ de l'enquête, comme l'étaient les 4% d'adultes dont l'adresse de résidence ne figurait pas sur la liste électorale. (Todd et Butcher 1982).

Pour l'enquête, on a utilisé un plan de sondage à plusieurs degrés comportant quatre étapes de sélection: tout d'abord, 103 circonscriptions d'Angleterre et du pays de Galles et 11 districts municipaux d'Écosse ont été sélectionnés avec une probabilité proportionnelle à la taille de l'électorat; ensuite, une section de vote a été sélectionnée dans chacun de ces points d'échantillonnage avec, là encore, une probabilité proportionnelle à la taille de l'électorat; en troisième lieu, vingt-trois adresses ont été choisies dans chacune des sections de vote avec une probabilité proportionnelle au nombre d'électeurs inscrits a l'adresse. Enfin, l'intervieweur a choisi une personne à chaque adresse échantillonnée à l'aide d'une variante de la méthode de Marchant-Blyth. (Blyth et Marchant 1973).

Pour les besoins de l'étude, on a prélevé un échantillon parallèle d'environ 800 adresses des 114 points d'échantillonnages originaux (7 adresses par point d'échantillonnage). On a ensuite soumis ces adresses et les noms inscrits sur la liste électorale au service d'extraction des numéros de téléphone de British Telecom. Le service a produit des numéros de téléphone pour 65% des adresses qui lui avaient été soumises. La différence entre ce taux d'extraction et la proportion de ménages ayant le téléphone – environ 75% au moment de l'étude – s'explique principalement par les numéros qui ne sont pas inscrits dans l'annuaire; en effet, selon Collins et Sykes (1987), environ 12% des numéros de téléphone en Grande-Bretagne ne figurent pas dans l'annuaire, la proportion variant selon la région et d'autres facteurs. Les autres problèmes liés à l'extraction des numéros de téléphone semblent avoir eu peu d'effet sur la bonne marche de l'étude.

British Telecom a procédé de la façon suivante pour extraire les numéros de téléphone: après avoir déterminé la circonscription téléphonique appropriée au moyen de l'adresse, on a cherché le nom de l'abonné dans l'annuaire. Les détails de l'adresse (par exemple le nom de la rue) permettaient de distinguer les abonnés qui avaient des noms identiques. Comme la liste électorale n'indiquait pas quelle personne parmi celles demeurant à la même adresse était l'abonné du téléphone, British Telecom avait pour instruction de vérifier chaque nom avant de rejeter l'unité d'échantillonnage.

Les numéros de téléphone ainsi obtenus ont été répartis systématiquement en quatre sous-échantillons. Les membres de deux de ces sous-échantillons ont été interviewés par téléphone à l'aide d'un questionnaire qui devait être rempli en l'espace d'environ 20 minutes. Les questions étaient tirées de toutes les sections du questionnaire de l'enquête sur les attitudes sociales. Les membres des deux autres sous-échantillons ont été interviewés par téléphone à l'aide d'un questionnaire plus long (qui devait être rempli en 40 minutes environ), dont les questions provenaient également du questionnaire de l'enquête principale. Les membres des quatre sous-échantillons ont reçu une lettre les avisant qu'ils allaient recevoir un appel téléphonique pour une enquête. Les membres des autres sous-échantillons n'ont reçu aucun avis à cet effet. Dans tous les cas, le choix des répondants s'est fait selon les mêmes règles que celles utilisées pour l'enquête principale.

Les membres de l'échantillon expérimental pour lesquels il n'a pas été possible d'obtenir un numéro de téléphone ont été soumis à une interview sur place de 20 minutes. Les résultats des interviews téléphoniques et des interviews sur place (méthode d'interview mixte) ont ensuite été comparés aux résultats de l'enquête principale, qui comportait uniquement des interviews sur place (Sykes et Hoinville, 1985).

On a voulu analyser plus directement les effets du mode d'interview en soumettant au service d'extraction de numéros de téléphone de British Telecom un sous-échantillon systématique de 600 adresses choisies parmi les adresses utilisées pour l'enquête principale (5 adresses par point d'échantillonnage). British Telecom a alors produit des numéros de téléphone pour 55% des adresses qui lui avaient été soumises (la variabilité du taux d'extraction demeure inexpliquée). On a ensuite comparé les réponses fournies par les personnes interviewées par téléphone et celles fournies par les personnes qui auraient pu être interviewées par téléphone mais l'ont été sur place. En limitant la comparaison aux personnes qui peuvent être rejointes par téléphone, nous avons éliminé les effets attribuables aux différences de méthode de collecte de données et avons considéré uniquement les effets découlant des différences entre les groupes comparés.

2.2 Étude n° 2

La seconde étude portait plus spécialement sur cette comparaison directe. On a prélevé environ 2300 adresses dans la liste électorale, comme dans la première étude, et on les a soumises à British Telecom pour l'extraction des numéros de téléphone (dans ce cas, le taux d'extraction a été de 61%). Les adresses pour lesquelles British Telecom a produit des numéros de téléphone ont été réparties en trois sous-échantillons. Les membres du premier sous-échantillon ont été interviewés par téléphone suivant des méthodes classiques; les membres du deuxième sous-échantillon ont été interviewés à l'aide de la méthode ITAO (interview téléphonique assistée par ordinateur) et les membres du troisième sous-échantillon ont été interviewés sur place. L'application de la méthode ITAO a été un échec (pour diverses raisons), mais les deux autres sous-échantillons nous permettent, encore une fois, de faire une comparaison directe entre les personnes interviewées par téléphone et celles qui auraient pu l'être mais qui ont été interviewées sur place. Le questionnaire, qui devait être rempli en 25 minutes, formait un sous-ensemble de questions de l'enquête de 1983 sur les attitudes sociales au Royaume-Uni.

2.3 Limites de la comparaison des modes d'interview

Trois facteurs peuvent limiter la comparaison entre les réponses obtenues par l'interview sur place et celles obtenues par l'interview téléphonique. Premièrement, des différences de taux de non-réponse (voir section 3) peuvent avoir été à l'origine de différences dans la composition des groupes de répondants. Cette possibilité a été vérifiée à l'aide d'un certain nombre de variables démographiques et socio-économiques que l'on croit être associées à certaines variables d'attitude. Des différences significatives entre les groupes de répondants donnent

à penser qu'au-delà des différences de taux de réponse globaux entre les modes d'interview, certains types de personnes sont plus susceptibles de se prêter à une interview téléphonique qu'à une interview sur place et vice-versa. Les variables utilisées sont l'âge croisé avec le sexe, l'état matrimonial, la composition du ménage, la situation économique, le groupe socio-économique et le lieu de résidence. La première étude n'a révélé aucune différence significative sur le plan statistique entre les taux de non-réponse. Dans la seconde étude, des différences significatives entre l'échantillon interviewé par téléphone et l'échantillon interviewé sur place ont été observées pour deux variables: la composition du ménage (l'échantillon de personnes interviewées par téléphone comportait relativement plus de couples de moins de 60 ans sans enfant, tandis que l'échantillon de personnes interviewées sur place comportait relativement plus de couples avec des enfants en bas âge ou des adolescents), et le groupe socio-économique (les cols blancs de niveau subalterne ou intermédiaire et les travailleurs ou travailleuses des "autres" professions étaient plus fortement réprésentés dans l'échantillon interviewé par téléphone que dans l'échantillon interviewé sur place, tandis que les "personnes au foyer" étaient relativement plus nombreuses dans l'échantillon interviewé sur place). Ces différences ne représentent peut-être que des fluctuations d'échantillonnage mais elles devraient inspirer une certaine prudence dans l'interprétation des différences dans les réponses des deux échantillons.

Le deuxième facteur qui pourrait limiter la comparaison des deux modes d'interview sont les différences de niveau de compétence ou de supervision entre les intervieweurs affectés aux interviews téléphoniques et ceux affectés aux interviews sur place. Six intervieweurs du premier groupe ont participé à la première enquête expérimentale. Deux de ces six personnes avaient de l'expérience dans les interviews sur place tandis que les quatre autres n'avaient aucune expérience; ces dernières ont donc reçu une formation de base ainsi que la formation spéciale pour les interviews téléphoniques, qui a aussi été offerte aux deux intervieweurs expérimentés. La seconde enquête réunissait dix intervieweurs, dont trois avaient participé à l'enquête précédente. Comme dans le premier cas, un superviseur était présent pour suivre la conversation téléphonique, proposer, si nécessaire, des techniques d'interview et relever les erreurs évidentes dans les questionnaires remplis.

Pour les deux enquêtes, les intervieweurs affectés aux interviews sur place ont été choisis parmi les quelques 300 intervieweurs réguliers du Social Community Planning Research. Ces personnes ont reçu la même formation de base que les personnes affectées aux interviews téléphoniques, qui avaient généralement moins d'expérience que celles-ci. Ces différences, et particulièrement celles qui font entrevoir une qualité inférieure dans les interviews téléphoniques, devraient par conséquent être prises en considération.

Le troisième facteur concerne le questionnaire. Le questionnaire de l'enquête principale, qui comprenait environ 100 questions, couvrait cinq grands sujets: l'emploi, l'éducation, la santé et le logement, les classes sociales et la discrimination fondée sur la race ou le sexe. Les questionnaires des deux enquêtes expérimentales étaient composés des questions jugées les plus importantes dans l'enquête principale. Ces questions ont été choisies de manière à refléter tout l'éventail des questions de l'enquête principale.

Les questionnaires des enquêtes expérimentales couvraient donc toute une série de sujets (y compris certains sujets "délicats") et renfermaient des questions qui correspondaient à divers modèles de réponse et niveaux de complexité. Les questions utilisées dans les interviews de 20 et de 40 minutes de la première enquête et dans l'interview de 25 minutes de la seconde enquête ont été posées dans le même ordre qu'elles l'avaient été dans l'enquête principale. Ainsi, l'interview de 40 minutes ne consistait pas en deux interviews de 20 minutes chacune (la première d'entre elles étant la version existante) mais les questions d'une interview de 20 minutes (version existante) étaient éparpillées dans l'interview de 40 minutes. Le libellé des questions était modifié uniquement lorsque cela était vraiment nécessaire; par exemple, certaines questions ont dû être reformulées pour tenir compte de l'absence inévitable de cartons ou de fiches. Comme le questionnaire de l'enquête sur les attitudes sociales est surtout constitué de questions fermées, quelques-uns seulement des résultats de nos enquêtes expérimentales ont trait à des questions ouvertes.

Il convient de se rappeler toutes ces contraintes au moment de l'étude des résultats; néanmoins, ces contraintes sont presque inévitables dans des études comparatives. Comme nous l'avons vu ci-dessus, nous avons tenté de définir et de minimiser ces limites. Celles-ci sont préoccupantes seulement quand les résultats de nos enquêtes donnent à penser que des effets de modes d'interview pourraient brouiller les effets d'autres variables; or, la plupart de nos résultats ne laissent entrevoir rien de tel. Nous pourrions, par conséquent, considérer que les limites définies ci-dessus ne peuvent avoir que des effets qui neutraliseraient les effets de modes d'interview que nous observerions en d'autres circonstances; de sorte que la validité de nos conclusions est menacée dans une moindre mesure.

3. TAUX DE RÉPONSE

Au Royaume-Uni, on s'interroge beaucoup sur la faisibilité de l'interview téléphonique, surtout pour des enquêtes sociales. On se demande non seulement quel niveau de communication est possible et quel pourra en être l'effet sur les aspects objectifs et affectifs de l'interview, mais encore s'il est socialement acceptable d'utiliser le téléphone pour faire des enquêtes. En Grande-Bretagne, les spécialistes des sciences sociales croient généralement que les appels téléphoniques d'étrangers sont susceptibles d'être accueillis avec méfiance; un appel d'un intervieweur pourrait être considéré comme inopportun et gênant.

En contrepartie, on souligne souvent les avantages que peut présenter l'interview téléphonique par rapport à l'interview sur place, notamment dans les quartiers situés au coeur des grandes villes. L'intensification des crimes contre la personne et la propriété a amené les gens à se méfier de plus en plus des étrangers, attitude qui se traduit par une diminution des taux de réponse; en outre, l'installation de dispositifs comme des systèmes d'intercommunication rend plus difficile le contact avec des répondants éventuels. En revanche, si l'intervieweur utilise le téléphone, il pourra sans doute rejoindre un répondant éventuel s'il y a quelqu'un à la maison et, dans le cas contraire, il ne sera pas beaucoup plus coûteux de rappeler plus tard.

Tableau 1

Enquêtes expérimentales du SCPR: taux de réponse

Bases	Étude n° 1		Étude n° 2	
	Par téléphone (429)	Sur place (313)	Par téléphone (730)	Sur place (631)
	%	%	%	%
Interviews complètes	53	60	46	68
Interviews partielles	1	–	–	–
Refus (sans sélection)	5	2	21	6
Refus (personne interposée)	9	5	7	4
Refus (personne échantillonnée)	11	18	10	11
Aucun contact[a]	3	1	8	4
Personne échantillonnée toujours absente	3	3	3	2
Malade, à l'extérieur – ne comprends pas la langue de l'intervieweur	2	5	2	4
Autre[b]	13	6	4	2

Étude n° 1: $X^2 = 3.72$ d.l. = 1 $0.05 < p < 0.1$

Étude n° 2: $X^2 = 66.22$ d.l. = 1 $p < 0.001$

Étude n° 1 et 2 combinées: $X^2 = 59.46$ d.e. = 1 $p < 0.005$

Comparaisons utilisant deux catégories seulement: interviews complètes et interviews non complètes

[a] Incluant "Pas de réponse" et "Ligne constamment occupée".

[b] Incluant "Rendez-vous manqué", "Trop âgé", "Handicapé", "Pas de service au numéro composé", "Bon numéro, mauvaise adresse".

Le tableau 1 donne les taux de réponse pour les deux enquêtes réalisées par le Centre des méthodes d'enquête.

Les taux de réponse des deux enquêtes ont été relativement faibles tant pour les interviews téléphoniques que pour les interviews sur place. (Les taux de réponse pour les interviews sur place devraient normalement être supérieurs à 70% avant la relance des cas de refus). Ces taux anormalement faibles ne sont pas sans rapport avec la nature des enquêtes ; on sait combien il est difficile de convaincre d'éventuels répondants de participer à des enquêtes à caractère général. La même observation vaut pour la seule autre enquête comparative d'importance au Royaume-Uni, réalisée par Marplan pour le compte du Market Research Development Fund. Pour cette enquête, on a utilisé la même méthode d'échantillonnage que celle dont nous nous sommes servis dans nos enquêtes expérimentales et on a posé toute une série de questions d'ordre général sous le titre *Mode de vie des années 80.* Les taux de réponses ont été de 45% pour les interviews téléphoniques et de 67% pour les interviews sur place (Market Research Development Fund, 1985). Les tailles d'échantillon était de 1697 pour les interviews téléphonique et de 1233 pour les interviews sur place. Pour ce qui a trait à nos deux enquêtes expérimentales, les taux de réponse ont été moins élevés dans le cas des interviews téléphoniques. À peine la moitié de l'échantillon a produit des interviews complètes. Comme le tableau 1 le montre, la différence entre les taux de réponse était presque significative pour l'étude n° 1, et significative pour l'étude n° 2, et la combinaison les deux études.

On pourrait attribuer les différences de taux de réponse entre les deux modes d'interview à notre inexpérience en matière d'interview téléphonique; mais ces différences sont également observées dans d'autres pays. Aux États-Unis par exemple, un certain nombre d'auteurs (par exemple Hochstim 1967; Henson, Roth et Cannell 1977) ont signalé des taux de réponse moins élevés pour les interviews téléphoniques, cette situation étant surtout attribuable à l'augmentation du nombre de cas de refus. Groves et Kahn résument la situation en ces termes:

> "Le taux de réponse des enquêtes nationales demeure au moins à cinq pourcent au-dessous du taux de réponse normalement obtenu dans les interviews sur place. Ce résultat est plutôt stable en dépit des changements survenus au cours des années dans la formation des intervieweurs et du perfectionnement des méthodes de supervision, du processus de rétroaction avec les superviseurs et des méthodes de présentation de l'enquête." (Traduction) (Groves et Kahn 1979, p. 219).

Ces constatations donnent à penser que le mauvais accueil réservé à l'interview téléphonique s'explique peut-être plus par des considérations sociologiques et psychologiques que par l'inexpérience de l'intervieweur ou la nouveauté des méthodes utilisées. Cependant, la première enquête du SCPR semble avoir été mieux réussie que la seconde enquête ou l'enquête du MRDF. Certains attribuent cela à l'intérêt et à l'enthousiasme qu'a soulevé la première enquête. Cet enthousiasme peut s'être communiqué aux intervieweurs (par exemple les spécialistes assistaient souvent aux interviews), ce qui aurait influé sur le taux de réussite de l'enquête. Les résultats obtenus dans le cas des interviews sur place donnent certes à penser que l'enthousiasme et le dynamisme de l'intervieweur sont des éléments essentiels pour obtenir des taux de réponse élevés.

Dans les enquêtes du SCPR, on a modifié deux conditions pour évaluer l'effet que cela aurait sur les taux de réponse pour les interviews téléphoniques. En ce qui a trait à la première enquête, la moitié des personnes interviewées par téléphone se sont prêtées à une interview de 20 minutes et l'autre moitié à une interview de 40 minutes (on informait les répondants de la durée de l'interview vers la fin de la présentation); par ailleurs, pour les deux enquêtes, on a fait parvenir à une moitié de l'échantillon de personnes interviewées par téléphone (cette moitié étant choisie au hasard) une lettre leur annonçant l'interview.

Le tableau 2 montre que le taux de réponse est inférieur dans le cas de l'interview de 40 minutes mais qu'il n'y a pas de différence significative entre les distributions (en pourcentage) pour les deux genres d'interview. Le plus faible taux de réponse enregistré pour l'interview

Tableau 2
Enquêtes expérimentales du SCPR: effets de la durée de l'interview (étude n° 1)

Bases	40 minutes (206)	20 minutes (223)
	%	%
Interviews complètes	48	59
Refus	27	23
Autres	25	18

$\chi^2 = 4.7$ d.l. = 2 $0.10 > p > 0.05$

Tableau 3
Enquêtes expérimentales du SCPR: effet des lettres d'introduction sur les taux de réponse

	Étude n° 1		Étude n° 2	
Bases	Lettre envoyée (215)	Aucune lettre envoyée (214)	Lettre envoyée (388)	Aucune lettre envoyée (392)
	%	%	%	%
Interviews complètes	55	51	48	43
Refus	23	27	37	38
Autres	22	21	15	19

Étude n° 1: $\chi^2 = 1.09$ d.l. = 2 $p > 0.5$
Étude n° 2: $\chi^2 = 2.8$ d.l. = 2 $p > 0.2$
Étude n° 1 et 2 combinées: $\chi^2 = 3.49$ d.l. = 2 $p > 0.1$

de 40 minutes est principalement attribuable au taux de refus direct plus élevé enregistré pour cette même interview, ce qui donne à penser que les répondants étaient moins disposés à participer à de longues interviews. Néanmoins, seules quelques-unes des personnes qui ont accepté de participer à l'enquête n'ont pas complété l'interview (longue ou courte).

De longs questionnaires peuvent exiger une approche différente. Il semble raisonnable de demander à un répondant, au premier contact, de se prêter à une interview de 20 minutes, mais l'intervieweur aurait intérêt à fixer des rendez-vous pour les interviews plus longues. Wiseman et McDonald (1979) estiment que les taux de refus seront moins élevés si les intervieweurs ont reçu l'instruction de fixer une date et une heure précises avec le répondant pour un second contact si leur interlocuteur leur indique qu'il est occupé au moment du premier contact.

Dans d'autres études, on a observé que l'envoi de lettres d'introduction à des personnes choisies pour participer à une interview téléphonique avait pour effet d'accroître les taux de réponse. Par exemple, Dillman, Gallegos et Frey (1976) ont constaté des taux de refus de 6% en moyenne parmi les personnes qui avaient reçu une lettre d'introduction, comparativement à des taux de 14% pour les personnes qui n'avaient pas reçu une telle lettre. En ce qui concerne les enquêtes du SCPR, le tableau 3 montre des taux de réponse légèrement plus élevés pour les répondants à qui on avait fait parvenir une lettre d'introduction (on ignore toutefois quelle proportion de ces répondants ont vraiment reçu la lettre); néanmoins, les différences observées ne sont pas statistiquement significatives.

Pour connaître les raisons qui incitent certains répondants à refuser une interview téléphonique, on a approché 55 personnes qui avaient refusé une telle interview lors de la première enquête pour savoir si elles auraient accepté de participer à l'enquête au premier contact si ce contact s'était fait en personne plutôt qu'au téléphone. Quarante de ces personnes ont indiqué que le mode d'interview n'influençait aucunement leur décision et quelques-unes seulement de ces personnes ont accepté par la suite de se prêter à une interview. Les quinze autres personnes ont pour la plupart affirmé qu'elles auraient participé à l'enquête s'il s'était agit d'une interview sur place, et se sont effectivement prêtées à une telle interview ultérieurement (13 personnes sur 15).

Comme on n'a pas approché les personnes qui avaient refusé une interview sur place, nous ne savons pas si une proportion d'entre elles auraient préféré être interviewées par téléphone.

3.1 Divergence des réponses et qualité des données

La conception que le public se fait d'une utilisation normale du téléphone peut influer non seulement sur les taux de réponse mais aussi sur les sortes de questions auxquelles les répondants sont disposés à répondre. Ce qui importe davantage, toutefois, c'est le genre de communication qu'il est possible d'établir entre l'intervieweur et le répondant et les effets qu'il peut avoir sur les estimations.

L'interview sur place donne lieu à des échanges verbaux et non verbaux tandis que l'interview téléphonique n'a qu'une capacité d'information limitée, les échanges entre l'intervieweur et le répondant se limitant aux propos de la conversation et à ce qu'on appelle les signaux paralinguistiques: ton de voix, pauses et ainsi de suite (Miller et Cannell 1982).

La capacité d'information limitée de l'interview téléphonique peut avoir de nombreux effets sur les estimations d'enquête. Par exemple, l'absence de support visuel peut rendre plus difficile de répondre à certaines questions. Une communication uniquement verbale peut ne pas rendre entièrement le sens des paroles d'un répondant (ce qui complique, par exemple, l'analyse des réponses à des questions ouvertes) et ne permet pas toujours de vérifier si le répondant a bien compris la question. L'absence de support visuel peut également nuire à l'intervieweur dans l'accomplissement de sa tâche. Par exemple, les signaux verbaux peuvent-il remplacer les signaux non verbaux qui soutiennent l'intérêt et l'attention du répondant ou qui aident à diriger l'interview? L'intervieweur peut-il garder constamment l'attention du répondant, surtout dans les longues interviews? Inversement, est-il souhaitable de supprimer les stimuli visuels afin de réduire les sources de variabilité des données d'enquête? Enfin, le caractère impersonnel de l'interview téléphonique ne fait-il pas hésiter le répondant à fournir des renseignements personnels comme le niveau de revenu ou des renseignements fortement associés à une catégorisation sociale?

Les enquêtes expérimentales du SCPR avaient pour but de répondre à certaines de ces questions.

3.1.1 Comparaison d'ordre général

Compte tenu de la différence de taux de refus entre les deux modes d'interview, il est surprenant de constater que ces modes diffèrent peu à d'autres points de vue. Les mêmes observations ont été faites dans de nombreuses études aux États-Unis (Groves et Kahn 1979; Lucas et Adams 1977; Jordan et coll. 1980; Colombotos 1969; Wiseman, 1972) et dans d'autres pays comme le Danemark (Kormendi et coll. 1986). Après avoir utilisé le même questionnaire (questions simples) dans des interviews téléphoniques et des interviews sur place, on a obtenu des distributions de réponses similaires.

Dans les enquêtes du SCPR, on a comparé les distributions de réponses marginales produites par les deux modes d'interview et vérifié à l'aide de tests du chi-deux si les différences observées étaient statistiquement significatives. Ces tests ont été appliqués à des données non pondérées. Toutefois, les tableaux qui figurent dans le présent article contiennent, sauf avis contraire,

Tableau 4
Différences entre les distributions de réponses marginales: interview téléphonique et interview sur place

Bases	Étude n° 1 (95)	Étude n° 2 (69)
	%	%
Aucune différence significative	91	87
Différence significative à un seuil de 5%	7	9
Différence significative à un seuil de 1%	2	4

des distributions ayant trait à des données qui ont été pondérées en fonction des différences qui pouvaient exister entre le nombre de personnes inscrites à une même adresse sur la liste électorale et le nombre de personnes demeurant effectivement à cette adresse. De telles différences ont été relevées dans environ 25% des cas, et chaque fois les données ont été pondérées par le nombre de personnes âgées de 18 ans ou plus qui vivaient à l'adresse divisé par le nombre d'électeurs inscrits à cette adresse. Nous avons choisi de produire des tableaux de données pondérées au cas où le lecteur voudrait tirer des conclusions distinctes pour les données obtenues par interview téléphonique et pour celles obtenues par interview sur place en considérant que les deux séries de données ont été établies selon les procédures normales.

On a appliqué des tests de chi-deux standard même si les données provenaient d'un sondage à plusieurs degrés. Il a été démontré (voir, par exemple, Holt, Scott et Ewings 1980) que si l'on sous-estimait la variabilité réelle en ne tenant pas compte du plan de sondage, on obtiendrait en général des variables à tester trop élevées et, par conséquent, on rejetterait à tort des hypothèses nulles (les tests seraient anti-conservatifs). Pour ce qui a trait à l'enquête sur les attitudes sociales, toutefois, l'estimation des erreurs types réelles pour les variables d'attitude produit des effets du plan de sondage (rapport de l'erreur type complexe à l'erreur type de l'échantillonnage aléatoire simple) qui dépassent rarement 1.2 (Jowell et Witherspoon 1985). On soutient en outre dans les ouvrages portant sur le sujet que la répartition en grappes est susceptible d'avoir des conséquences moins sérieuses dans les tests d'indépendance bilatéraux (Holt, Scott et Ewings 1980). C'est pourquoi nous nous croyons justifiés d'utiliser des tests du chi-deux standard afin d'éviter les nombreux calculs requis pour obtenir des statistiques corrigées. Cette méthode aura plutôt pour effet de surestimer la signification des différences entre les modes d'interview.

Nous avons analysé 95 questions et sous-questions dans la première étude et 69 dans la seconde. Les résultats figurent dans le tableau 4.

Nous voyons clairement que les résultats des deux études sont conformes à ceux obtenus par d'autres spécialistes; les différences de distribution de réponses entre les modes d'interview ne sont significatives que pour un très faible pourcentage des questions. L'étude du MRDF a abouti à une conclusion semblable.

3.1.2 Comparaison de formes de questions particulières

Malgré les constatations générales, des recherches aux États-Unis ont montré qu'il existe certains types de questions qui entraînent inévitablement des différences de distributions de réponses significatives. Par exemple, Groves et Kahn (1979) ont démontré que les répondants avaient tendance à donner des réponses incomplètes à des questions ouvertes au cours d'interviews téléphoniques. Cette tendance peut être attribuable au rythme plus rapide de

Tableau 5
Durée des interviews selon le mode d'interview (étude n° 2)

Bases non pondérées	Interview téléphonique (354)	Interview sur place (360)
	%	%
Minutes		
Inférieur à 20	10	5
20–29	63	53
30–40	22	33
40+	6	8

$\chi^2 = 17.6$ d.l. = 3 $p < 0.01$

Tableau 6
Réponses à une question ouverte (enquête du MRDF): comparaison par mode d'interview

Que trouvez-vous de bon à la soupe ... ?

Bases	Interview téléphonique (700)	Interview sur place (601)
	%	%
Nombre de réponses		
Aucune	33	22
Une	58	61
Deux	7	14
Trois ou plus	1	2
Moyenne	0.77	0.96

$\chi^2 = 32.2$ d.l. = 3 $p < 0.01$

l'interview téléphonique, comme le soulignent, par exemple, Dillman (1970) et Williams (1977). Dans une conversation téléphonique, l'intervieweur et le répondant ont tous deux tendance à parler plus rapidement et à éviter les pauses. Notre seconde étude a mis en évidence cette constatation. Le tableau 5 donne la durée réelle d'interviews qui, normalement, devaient s'étendre sur 25 minutes. Nous voyons que 10% des interviews téléphoniques ont été réalisées en moins de 20 minutes, comparativement à 5% pour les interviews sur place. À l'autre extrême, 41% des interviews sur place ont duré plus d'une demi-heure, comparativement à moins du tiers pour les interviews téléphoniques.

Ball (1980) estime que l'interview téléphonique suit un rythme plus rapide parce que les règles de la conversation téléphonique exigent que l'intervieweur et le répondant s'efforcent l'un et l'autre de soutenir la conversation. Cela laisse probablement moins de temps aux répondants pour préparer leurs réponses. Sans doute, les moments de silence semblent mettre les gens mal à l'aise; dans une étude de Jordan (1980), les intervieweurs ont qualifié d'interminables les pauses qu'ils faisaient normalement dans leurs interviews. Il y a sans doute beaucoup d'autres facteurs qui entrent en ligne de compte; même l'absence de signaux visuels n'est pas à négliger.

Les enquêtes expérimentales du SCPR ne comportaient pas de question ouverte, mais l'enquête du MRDF renfermait un certain nombre de questions visant à vérifier si la compréhension était spontanée. La comparaison des résultats des interviews téléphoniques et des interviews sur place semble confirmer les observations ci-dessus. À titre d'exemple, le tableau 6 montre qu'un tiers des personnes interviewées par téléphone n'ont pas donné de réponse à une question particulière tandis que dans le cas des personnes interviewées sur place, la proportion correspondante a été inférieure à 25%. En outre, le nombre moyen de réponses données par téléphone a été beaucoup moins élevé que le nombre moyen de réponses données dans une interview sur place.

Nous pourrions supposer que des réponses plus nombreuses ou plus longues produisent des résultats plus justes, ce qui nous amènerait à souligner la nécessité de méthodes permettant de mieux adapter les questions ouvertes aux enquêtes par téléphone. À l'opposé, nous pourrions conclure que les questions ouvertes n'ont qu'un usage limité dans les enquêtes par téléphone, notamment lorsqu'on cherche à obtenir uniquement la première impression des répondants. Cette affirmation doit toutefois être vérifiée; nous devons ici nous contenter d'en décrire les effets.

On a également observé des différences de distribution de réponses pour des questions qui avaient trait à des échelles d'attitude et qui ont été incluses aussi bien dans des interviews sur place que dans des interviews téléphoniques. Les personnes interviewées par téléphone tendent à souscrire à ce qu'on leur propose ou à fournir des réponses extrêmes (Jordan, Marcus et Reeder 1980; Groves et Kahn 1979). Le modèle de réponse "d'accord/pas d'accord" utilisé dans l'enquête du MRDF a permis de constater que les personnes interviewées par téléphone étaient légèrement plus portées à être d'accord avec ce qu'on leur proposait. Par ailleurs, on n'a observé aucune variation notable de l'éventail de réponses; rien ne laissait croire à une plus forte tendance à fournir des réponses extrêmes.

3.1.3 Questions délicates

En ce qui a trait aux genres de questions que l'on peut utiliser dans les interviews téléphoniques, les spécialistes se sont intéressés longuement aux questions délicates, c'est-à-dire à celles qui ont trait à des renseignements personnels et à celles pour lesquelles il existe des réponses plus conformistes que d'autres au point de vue social. Au départ, on ne s'entendait pas sur les effets probables de questions délicates dans une interview téléphonique. D'une part, on prétendait que les répondants serait moins portés à fournir des réponses justes parce que, l'intervieweur n'étant pas sur place, ils se sentiraient moins obligés d'adopter une attitude franche et ouverte. D'autre part, on prétendait que les répondants seraient plus portés à donner des réponses justes parce que, grâce au caractère impersonnel de la conversation téléphonique, ils hésiteraient moins à divulguer des renseignements à caractère confidentiel.

La plupart des constatations tendent à corroborer la seconde thèse (Colombotos 1965; Wiseman 1972; Henson, Roth et Cannell 1974; Locander 1974; Rogers 1976). La principale exception est signalée par Groves et Kahn (1979), qui ont constaté que certaines personnes interviewées par téléphone hésitaient à fournir des renseignements sur leur situation financière et à répondre à d'autres questions délicates.

Nos études confirment l'hypothèse selon laquelle les personnes interviewées par téléphone ne craignent pas de répondre honnêtement aux questions délicates. Dans notre première étude par exemple, nous avons déterminé 14 questions qui pouvaient être jugées délicates et les avons soumises à des tests visant à comparer les effets du mode d'interview. Nous donnons ci-dessous trois exemples de questions jugées délicates:

i) Comment qualifierez-vous votre attitude à l'égard des gens d'autres races? (Lire à haute voix) ...
   ... Très empreinte de préjugés
   ... Légèrement empreinte de préjugés
   ... Libre de préjugés.

ii) Croyez-vous que dans l'ensemble, la Grande-Bretagne est trop, trop peu ou juste assez généreuse envers les Asiatiques et les Antillais établis au pays?

iii) Enfin, j'aimerais que vous me disiez si vous jugez acceptable qu'une personne homosexuelle occupe un poste de professeur dans une école.

Il n'y avait aucune question qui a donné lieu à des différences significatives dans les distributions de réponse marginales. Pour plusieurs questions, toutefois, les personnes interviewées sur place avaient un peu plus tendance à donner des réponses conformistes au point de vue social. En d'autres mots, les questions semblaient moins délicates au téléphone. Par exemple, 28% des personnes interviewées par téléphone ont déclaré avoir été interrogées par la police relativement à un crime au cours des deux dernières années alors que dans le cas des personnes interviewées sur place, la proportion correspondante a été de 20%.

On a également observé, pour les questions délicates portant sur le crime et la sexualité dans l'enquête du MRDF, que les personnes interviewées par téléphone avaient une faible tendance à donner des réponses plus "honnêtes"; en général, toutefois les différences de distributions observées n'étaient pas significatives. Par exemple, lorsqu'on a demandé aux personnes interviewées de se décrire sur un certain nombre de points, les personnes interviewées par téléphone se sont dites plus "séduisantes" (une moyenne de 2.81 sur 4 comparée à 2.72 sur place) et ils étaient plus enclin à répondre (88% ont répondu comparé à 75% pour les interviews sur place).

Les questions relatives au revenu ont toujours été considérées comme une source de problèmes dans les enquêtes par téléphone (problèmes liés à la volonté de participer des enquêtés et aux réponses fournies). La sous-déclaration du revenu est le problème le plus fréquent, bien qu'en pratique il puisse être difficile de faire la distinction entre la sous-déclaration et la sous-estimation découlant d'un taux de non-réponse plus élevé dans les tranches de revenu supérieures. Locander et Burton (1976) estiment que le degré d'exactitude des données sur le revenu peut dépendre du libellé de la question. Dans une comparaison de quatre libellés, on a constaté que la méthode qui consistait à demander premièrement au répondant si son revenu était supérieur à deux mille dollars et à accroître progressivement le montant repère jusqu'à ce que le répondant donne une réponse négative, donnait lieu à une sous-déclaration du revenu. Par contre, l'inversion de l'ordre des paliers du revenu produisait une sur-déclaration du revenu. La question qu'on utilisait pour les interviews téléphoniques dans les études du SCPR était semblable à la première méthode. C'est celle

Tableau 7

Revenu brut des ménages: études du SCPR

	Étude n° 1		Étude n° 2	
Bases[a] Revenu	Par téléphone (183)	Sur place (170)	Par téléphone (297)	Sur place (352)
Inférieur à £5,000	38	27	28	28
£5,000–£9,999	42	37	37	38
£10,000+	21	35	35	35
Étude n° 1: $\chi^2 = 10.08$ d.l. = 2 $p < 0.01$				
Étude n° 2: $\chi^2 = 0.11$ d.l. = 2 $p > 0.9$				
Bases	(217)	(199)	(344)	(405)
Ne sait pas/ N'a pas répondu	16%	15%	14%	13%

[a] "Ne sait pas" et "N'a pas répondu" sont exclues.

qui se rapproche le plus de la question sur le revenu dans l'interview sur place: la personne interviewée devait consulter une fiche sur laquelle les tranches de revenu figuraient par ordre croissant. Pour l'interview téléphonique, on énumérait les tranches de revenu en commençant par la plus basse. Les résultats sont indiqués dans le tableau 7.

Dans les différentes études, les enquêtés se sont montrés plus ou moins disposés à répondre à la question sur le revenu suivant le mode d'interview utilisé. Des différences de distribution de réponses, en l'occurence attribuables à une sous-déclaration probable de revenu, n'ont été observées que dans la première enquête.

3.1.4 Questions complexes

Dans les deux études du SCPR, on a au préalable déterminé un certain nombre de questions qui pouvaient poser des problèmes de réponse particuliers aux personnes interviewées par téléphone. Ces questions étaient celles qui pouvaient renfermer une ou plusieurs notions difficiles à saisir, les questions longues et les questions à choix multiples. Ces questions "complexes" ne semblent pas poser plus de problèmes aux personnes interviewées par téléphone qu'à celles interviewées sur place. Par exemple, des 19 questions "complexes" définies comme telles dans la première étude (12 de ces 19 questions s'accompagnaient de cartons dans les interviews sur place), une seule a eu des effets attribuables au mode d'interview.

4. SOMMAIRE ET CONCLUSIONS

Comme le nombre de personnes qui ont le téléphone au Royaume-Uni est encore relativement faible, surtout pour certaines couches de la population, il n'est pas probable que l'interview téléphonique puisse bientôt remplacer l'interview sur place pour les enquêtes qui exigent la participation des personnes moins favorisées. Néanmoins, on reconnaît de plus en plus les possibilités qu'offre ce mode d'interview lorsqu'il est combiné à l'interview sur place. Dans l'enquête sur la population active au Royaume-Uni, par exemple, la seconde interview et les interviews subséquentes se font par téléphone dans le cas des répondants admissibles qui ont accepté d'être contactés par téléphone.

Quant il y a utilisation simultanée des deux modes d'interview, il est essentiel pour le succès d'une enquête que cette circonstance ne provoque pas de distorsions. Les résultats que nous venons d'analyser sont encourageants pour l'avenir. À quelques exceptions près, on n'a relevé aucune différence statistiquement significative entre les distributions des réponses obtenues sur place et les distributions des réponses obtenues par téléphone.

Néanmoins, les taux de réponse relativement faibles des enquêtes par téléphone posent des problèmes qui doivent être résolus. Des taux de refus élevés peuvent réduire l'efficacité en termes de coûts de l'interview téléphonique. Fait plus important encore, ils augmentent la possibilité d'introduction d'un biais dans l'échantillon. Pour mettre entièrement en valeur la méthode d'interview mixte au Royaume-Uni, il est nécessaire de poursuivre la recherche pour trouver des moyens d'améliorer les taux de réponse des enquêtes par téléphone.

REMERCIEMENTS

Cette étude a été réalisée sous l'égide du Centre des méthodes d'enquête du SCPR, subventionné à titre de centre de recherche désigné par le Conseil de la recherche économique et sociale (subvention HR 3333). Les auteurs tiennent à exprimer leur reconnaissance aux rédacteurs et aux arbitres pour leurs commentaires nombreux et utiles sur les premières versions de cette étude.

BIBLIOGRAPHIE

ARONSON, S. (1971). The Sociology of the Telephone. *International Journal of Comparative Sociology*, 12, 153-167.

BALL, D.W. (1968). Towards a Sociology of Telephones and Telephoners. Dans *Sociology and Everday Life* (ed. Marcello Truzzi), Englewood Cliffs, New Jersey: Prentice Hall.

BERGSTEN, J.W. (1979). Some Methodological Results from Four Statewide Telephone Surveys Using Random Digit Dialing. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 239-243.

BISHOP, Y.M.N., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multi-variate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.

BLANKENSHIP, A.B. (1977). *Professional Telephone Surveys*. London: McGraw-Hill.

BLYTH, W.G., et MARCHANT, L.J. (1973). A self-weighting random sampling technique. *Journal of the Market Research Society*, 15, 157-162.

CANNELL, C.F., OKSENBERG, L., et CONVERSE, J.M. (1979). Experiments in interviewing techniques. Research Report, Institute for Social Research, University of Michigan.

CHRISTOFFERSEN, M.N. (1984). The quality of data collected at telephone interviews. Danish National Institute of Social Research, Copenhagen.

COLLINS, M. (1983). Telephone interviewing in consumer surveys. *Market Research Society Newsletter*, octobre.

COLLINS, M., et SYKES, W. (1987). The problems of non-coverage and unlisted numbers in telephone surveys in Britain. *Journal of the Royal Statistical Society*. Sér. A, 150, (en voie de rédaction).

COLOMBOTOS, J. (1965). The effects of personal vs. telephone interviews on socially acceptable responses. *Public Opinion Quarterly*, 29, 457-458.

COLOMBOTOS, J. (1969). Personal versus telephone interviews: effect on responses. *Public Health Reports*, 84, 773-782.

COOMBS, L., et FREEMAN, R. (1986). Use of telephone interviews in a longitudinal fertility study. *Public Opinion Quarterly*, 28, 112-117.

CZAJA, R., BLAIR, J., et SEBESTIK, J.P. (1982). Respondent selection in a telephone survey: a comparison of three techniques. *Journal of Marketing Research*, 19, 381-385.

DE MAIO, T.J. (1984). Refusals in telephone surveys: when do they occur?. Document présenté à la 39ième conférence annuelle de l'American Association for Public Opinion Research.

DILLMAN, D. (1970). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley.

DILLMAN, D., GALLEGOS, J., et FREY, J. (1976). Reducing refusal rates for telephone interviews. *Public Opinion Quarterly*, 40, 66-78.

FALTHZIK, A. (1972). When to make telephone interviews. *Journal of Marketing Research*, 9, 451-452.

FITTI, J.E. (1979). Some results from the telephone health interview survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 244-249.

FLEISHMAN, E., et BERK, M. (1979). Survey of interviewer attitudes towards methodological issues in the national medical care expenditure survey. Document présenté à la troisième conférence biennale sur la recherche et la méthodologie relatives aux enquêtes sur la santé, Reston, Virginia.

GROVES, R., et KAHN, R. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews*. New York: Academic Press.

GROVES, R.M., MAGILAVY, L.J., et MATHIOWETZ, N.A. (1981). The process of interviewer variability. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 438-443.

GROVES, R.M., et MATHIOWETZ, N.A. (1984). Computer assisted telephone interviewing: effects on interviewers and respondents. Survey Research Center, University of Michigan.

HENSON, R., ROTH, A., et CANNELL, C.F. (1974). Personal vs. telephone interviews and the effects of telephone re-interviews on reporting of psychiatric symptomatology. Document de recherche, Survey Research Center, University of Michigan.

HOCHSTIM, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-986.

HOLT, D., SCOTT, A.J., et EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society*, Sér. A, 143, 303-320.

IBSEN, C.A., et BALLWEG, J. (1974). Telephone interviews in social research: some methodological considerations. *Quality and Quantity*, 7, 181-192

JOWELL, R., et WITHERSPOON, S. (1985). *British Social Attitudes: The 1985 Report*. Aldershot: Gower.

JORDAN, L.A., MARCUS, A.C., et REEDER, L.G. (1980). Response styles in telephone and household interviewing: a field experiment. *Public Opinion Quarterly*, 44, 210-222.

KAHN, R.L., et GROVES, R.M. (1977). *Comparing telephone and personal interview systems*. Survey Research Center, University of Michigan.

KORMENDI, E., EGSMOSE, L., et NOORDHOEK, J. (1986). Datakvalitet ved Telefon-interview. Socialforskingsinstituttet, Studie 52, Copenhagen.

LOCANDER, W.B., et BURTON, J.P. (1976). The effect of question form on gathering income data by telephone. *Journal of Marketing Research*, 13, 189-192.

LOCANDER, W.B., SUDMAN, S., et BRADBURN, N. (1974). An investigation of interview method, threat and response distortion. *Proceedings of the Social Statistics Section, American Statistical Association*, 21-27.

LUCAS, W.A., et ADAMS, W.C. (1977). *An Assessment of Telephone Survey Methods*. Santa Monica, California: Rand Corporation.

McCULLAGH, P., et NELDER, J.A. (1983). *Generalised Linear Models*. London: Chapman and Hall.

MILLER, P.V., et CANNELL, C.F. (1982). A study of experimental techniques for telephone interviewing. *Public Opinion Quarterly*, 46, 250-269.

Market Research Development Fund, (1985). *Comparing telephone and face-to-face surveys*. Marplan Ltd.

OKSENBERG, L., COLEMAN, L., et CANNELL, C. (1984). Voices and refusal rates in telephone surveys. Document non publié.

O'NEIL, M., GROVES, R., et CANNELL, C. (1979). Telephone interview introductions and refusal rates: experiments in increasing respondent cooperation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 252-255.

ROGERS, T.F. (1976). Interviews by telephone and in person. *Public Opinion Quarterly*, 40, 51-65.

SCHMIEDESKAMP, J.W. (1962). Reinterviews by telephone. *Journal of Marketing*, 26, 28-34.

SYKES, W., et HOINVILLE, G. (1985). Telephone interviewing on a survey of social attitudes: a comparison with face-to-face procedures. Social and Community Planning Research Center.

TODD, J., et BUTCHER, R. (1982). *Electoral Registration in 1981*. London: OPCS.

WILLIAMS, E. (1977). Experimental comparisons of face-to-face and mediated communication. *Psychological Bulletin*, 84, 963-976.

WISEMAN, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly*, 34, 105-108.

WISEMAN, F., et McDONALD, P. (1979). Noncontact and refusal rates in consumer telephone surveys. *Journal of Marketing Research*, 16, 478-484.

Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 35-51
Statistique Canada

Utilisation des dossiers administratifs à des fins statistiques

G.J. BRACKSTONE[1]

RÉSUMÉ

Les demandes pour les statistiques sur tous les aspects de nos vies, de notre société et de notre économie continuent à augmenter. En même temps, les organismes statistiques, ainsi que beaucoup de leurs répondants, s'inquiètent de plus en plus du fardeau de réponse imposé par les enquêtes. L'utilisation des dossiers administratifs à des fins statistiques s'est présenté comme un des moyens alternatifs pour satisfaire ces demandes statistiques. Cet article expose l'expérience récente à Statistique Canada dans ce domaine, et discute des entraves à une plus grande exploitation de sources administratives. On examine aussi des méthodes possibles pour rendre les systèmes administratifs plus utiles à des fins statistiques, ainsi que certaines questions importantes liées à la protection des renseignements et au jumelage d'enregistrements.

MOTS CLÉS: Estimation indirecte; base de sondage; évaluation des enquêtes; accès; confidentialité.

1. INTRODUCTION

La demande de statistiques se rapportant à de nombreux aspects de notre vie, de notre société, de notre économie et de notre environnement ne cesse d'augmenter. Le fait que nous sommes en mesure, de nos jours, de traiter et de manipuler de vastes ensembles de données, puisque nous entrons dans ce qu'il est convenu d'appeler l'ère de l'information, explique peut-être en partie ce phénomène; celui-ci pourrait également découler de la complexité croissante des systèmes sociaux et économiques et de notre désir de mieux comprendre ces derniers. Quelle que soit la raison qui motive la demande, les organismes statistiques publics doivent y répondre, et ce dans un climat de grandes compressions budgétaires. Ces organismes sont en même temps conscients du fardeau de réponse accru que les répondants auraient à subir s'il fallait mener davantage d'enquêtes pour obtenir les renseignements voulus.

Ces facteurs nous ont amenés à chercher d'autres façons de procéder, la principale étant un plus grand recours aux systèmes administratifs existants. L'idée n'est pas nouvelle. Depuis longtemps déjà les renseignements relatifs à l'état civil, aux importations, aux exportations, aux soins de santé et à l'éducation nous parviennent sous forme de produits dérivés des processus administratifs. Nous verrons plus loin que l'utilisation des données administratives s'est étendue ces dernières années au domaine des statistiques sur les entreprises, les familles et les particuliers.

Les quatre prochaines sections de l'article décrivent les différents types de dossiers administratifs et l'utilisation qui en est faite, notamment à Statistique Canada. Il ressortira de cet exposé que l'appareil statistique canadien est fortement tributaire des dossiers administratifs. La section 6 discute de la question de l'accès aux dossiers administratifs et les moyens de rendre ces dossiers mieux adaptés aux usages statistiques. Enfin, un survol des préoccupations que l'exploitation des dossiers administratifs suscite à l'égard de la protection des renseignements personnels est fourni.

[1] G.J. Brackstone, Statisticien en chef adjoint, Statistique Canada, 26-J Immeuble R.H. Coats, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

2. TYPES DE DOSSIERS ADMINISTRATIFS

Des dossiers administratifs, il y en a de toutes sortes. Il est important de faire la distinction entre ceux qui sont tenus au niveau national (habituellement par le gouvernement fédéral) et ceux qui le sont à un niveau inférieur (par les provinces ou les municipalités, par exemple). Pour que ces derniers soient utiles au niveau national, il faut que divers secteurs de compétence s'entendent notamment sur les définitions, les normes, la présentation des dossiers et la marche à suivre. Il n'est pas toujours facile de conclure de telles ententes, surtout dans des domaines qui sont constitutionnellement du ressort des provinces.

Les dossiers administratifs diffèrent par leur objet, lequel détermine en grande partie l'univers visé et leur qualité, et par conséquent leur utilité statistique. On peut distinguer six grandes catégories de dossiers administratifs selon leur raison d'être:

(1) *Les dossiers tenus pour contrôler la circulation des biens et des personnes aux frontières.*

Ces dossiers incluent notamment les dossiers sur les importations, les exportations, l'immigration et l'émigration. L'univers visé et le contenu de ces dossiers administratifs dépendent des lois et des règlements à faire respecter et du succès avec lequel ces lois et règlements sont appliqués. Dans l'ensemble, ils sont bien appliqués. Les dossiers administratifs sur l'immigration ne contiennent pas d'information concernants les immigrants illégaux, par définition, mais sont complets à d'autres égards. Toutefois, comme l'émigration n'est pas contrôlée, il n'existe pas de dossiers administratifs concernant l'émigration. Les dossiers administratifs sur les importations canadiennes sont en général plus précis que les dossiers sur les exportations parce que le besoin d'établir les droits de douane frappant les importations requiert l'information plus détaillée.

(2) *Les dossiers tenus en vertu de l'obligation légale d'enregistrer certains événements.*

À titre d'exemple de ces événements, citons les naissances, les décès, les mariages, les divorces, les constitutions d'entreprises en société ou les fusions d'entreprises et l'octroi de licences. De façon générale, l'univers visé et la qualité des dossiers recueillis à cette fin sont très élevés au Canada, car tout citoyen doit fournir une preuve que ce genre d'enregistrement a été fait pour obtenir des droits ou des avantages.

(3) *Les dossiers nécessaires à la prestation d'avantages et à l'administration d'engagements d'assistance.*

Citons par exemple les impôts, l'assurance-chômage, les pensions, l'assurance-maladie et les allocations familiales. L'univers visé et le contenu de ces dossiers dépendent beaucoup des programmes auxquels ils se rapportent. L'univers visé peut être très bien couvert, mais, pour des raisons politiques ou administratives, sa définition risque de ne pas être la plus utile pour les travaux d'analyse.

(4) *Les dossiers nécessaires à l'administration des établissements publics.*

Il s'agit notamment des dossiers concernant les écoles, les universités, les établissements de soins de santé, les tribunaux et les prisons. De façon générale, ces dossiers portent principalement sur le nombre de cas dont l'établissement a la responsabilité et non sur chaque personne qui y fait un séjour. Par contre, ils donnent habituellement des statistiques agrégées très complètes sur la population qui fréquente ces établissements. Au Canada, beaucoup de dossiers administratifs de cette catégorie sont du ressort des provinces.

(5) *Les dossiers qui découlent de la réglementation de certaines activités économiques par le gouvernement.*

Citons les dossiers relatifs au transport, aux affaires bancaires, à la radiodiffusion et aux télécommunications ainsi qu'à la gestion des approvisionnements ou des prix de certaines marchandises, particulièrement dans le secteur agricole.

(6) *Les dossiers liés à la prestation de services publics.*

Ces services comprennent l'électricité, le téléphone et l'eau. Le taux de couverture des abonnés et la qualité des renseignements relatifs à la prestation des services et à la facturation sont bons en général. Plusieurs de ces services sont administrés au niveau provincial ou municipal.

La façon dont les dossiers administratifs sont constitués varie également beaucoup d'un organisme à un autre. La plupart des systèmes administratifs dont l'univers visé est grand sont à présent automatisés, mais le fait que le matériel utilisé et la présentation des données diffèrent non seulement d'un secteur de compétence à un autre mais aussi entre l'organisme administratif et l'organisme statistique pose un problème qu'il reste à régler. L'automatisation de plus en plus répandue entraîne également le problème des nombreuses modifications apportées aux dossiers initiaux par l'organisme administratif intéressé avant qu'ils soient transmis à l'organisme statistique. Alors qu'un meilleur contrôle de la qualité des formulaires reçus aurait peut-être de bonnes chances d'améliorer la qualité finale du fichier administratif, l'organisme statistique est obligé, lui, de procéder à des travaux supplémentaires pour connaître et évaluer l'incidence de toute opération de traitement préliminaire que l'organisme administratif a pu exécuter. Dans certains services administratifs, les dossiers individuels demeurent à l'endroit où ils ont été constitués et seules les données agrégées sont conservées dans un système central. Cette pratique restreint la capacité de l'organisme statistique d'évaluer la qualité des données et limite les possibilités en matière d'analyse.

Enfin, les dossiers diffèrent du point de vue de leur accessibilité. Il y a des lois et des règlements qui régissent l'accès à certains dossiers administratifs ainsi que leur utilisation à des fins secondaires, c'est-à-dire autres que celles pour lesquelles ils ont été constitués, notamment à des fins statistiques. Nous reparlerons de cette question à la section 6.

3. UTILISATION DES DOSSIERS ADMINISTRATIFS

Les données administratives peuvent se prêter à quatre catégories d'utilisation statistique. La plupart des applications qui sont faites de ces dossiers à des fins statistiques se rattachent à une de ces catégories, en représentent une combinaison ou en constituent une variante.

(1) *Totalisations directes*

On inclut ici le comptage des unités dans les fichiers, leur classement recoupé selon certaines caractéristiques et le regroupement des variables quantitatives associées à chaque unité. Les statistiques sur l'état civil et le commerce extérieur sont des exemples importants. Mentionnons également la publication des chiffres mensuels sur le nombre de demandeurs et de bénéficiaires de l'assurance-chômage ventilés selon la province, l'âge, le sexe, la durée des prestations et leur nature ainsi que la production des sommaires annuels sur la répartition du revenu dans chaque comté d'après les données du fichier de l'impôt sur le revenu des particuliers.

(2) *Estimations indirectes*

Cette catégorie comprend les travaux où l'on se sert de données tirées des dossiers administratifs pour effectuer une estimation. Un exemple serait le couplage de deux déclarations d'impôt consécutives d'un particulier dans le but de produire des estimations partielles de la migration, lesquelles peuvent être pondérées à l'aide des données de recensement utilisées comme données de référence. Ces estimations de la migration sont ensuite utilisées dans le programme d'estimations démographiques de Statistique Canada, au même titre que les données administratives sur les naissances, les décès et l'immigration. Un deuxième exemple serait l'utilisation des données fiscales sur

les petites entreprises au lieu de la collecte des données requises au moyen d'une enquête. Les données fiscales, corrigées s'il y a lieu, sont combinées aux données d'enquête sur les grandes entreprises pour produire des données agrégées sur la branche d'activité.

On classe également dans cette catégorie d'utilisation le couplage de différents fichiers administratifs ou statistiques dans le but de produire des estimations. Par exemple, le couplage du registre des décès avec les fichiers des personnes exposées à certains risques dans le but d'estimer les taux de mortalité différentielle, ou encore le couplage de dossiers fiscaux avec des dossiers de l'assurance-chômage et de la formation de la main-d'oeuvre dans le but d'analyser la participation et l'adaptation au marché du travail.

(3) *Bases de sondage*

Dans cette catégorie, nous incluons l'utilisation de dossiers administratifs pour créer, compléter ou mettre à jour les bases de sondage servant aux recensements ou aux enquêtes. Un bon exemple serait l'utilisation des renseignements que les employeurs fournissent à Revenu Canada au sujet des retenues sur la paye. Le questionnaire que les titulaires d'un nouveau compte de retenues sur la paye doivent remplir permet de savoir exactement si de nouvelles entreprises ont été crées ou si des modifications ont été apportées à la structure d'entreprises existantes. Même s'il n'y a pas au Canada de registre de logements, un deuxième exemple serait l'utilisation des permis de bâtir ou des listes de nouveaux abonnés au téléphone ou au service d'électricité pour déterminer qu'un logement est nouveau.

(4) *Évaluation des enquêtes*

Cette catégorie comprend l'utilisation de dossiers administratifs à des fins de vérification, de validation ou d'évaluation de données produites au moyen d'une enquête, et ce au niveau de chaque unité ou à un niveau agrégé. Dans le passé, on s'est servi des dossiers de l'immigration et de l'impôt pour évaluer les questions du recensement sur l'immigration et les revenus, et des dossiers des allocations familiales pour vérifier le taux de couverture du recensement en ce qui a trait aux enfants.

Il y a un facteur important qui permet de déterminer l'utilisation qui sera faite d'une source de renseignements administratifs, et c'est la qualité présumée des données administratives comparativement à celle des données d'enquête correspondantes. Certaines études d'évaluation utilisent les dossiers administratifs pour évaluer les réponses aux questions des enquêtes, alors que d'autres se servent de données d'enquête comme données de référence pour évaluer les estimations produites à partir de données administratives. La qualité des dossiers administratifs doit être évaluée cas à cas. Du point de vue de l'utilisation statistique qu'on va en faire, elle dépend en général d'au moins trois facteurs:

(i) les définitions utilisées pour les besoins du système administratif;
(ii) le taux de couverture voulu du système administratif;
(iii) la qualité des données recueillies et de leur traitement par les services administratifs.

Si un de ces trois facteurs présente des faiblesses, l'utilité statistique des dossiers administratifs risque d'être compromise. La rapidité avec laquelle ces dossiers peuvent être obtenus est un autre aspect important à considérer. Les contraintes qui peuvent éventuellement limiter cette utilisation et qu'il faut par conséquent considérer avant toute décision ont été traitées ailleurs (voir par exemple Brackstone, 1984). Le tableau 1 compare les points forts et les points faibles des dossiers administratifs et ceux des recensements et des enquêtes.

Pour illustrer l'utilisation qui est faite au Canada des dossiers administratifs, nous allons décrire deux domaines d'application à Statistique Canada. Le premier a trait à la production de statistiques sur les entreprises et le second, à la production de statistiques sur les particuliers et les familles.

Tableau 1

Comparaison des recensements, des enquêtes et des dossiers administratifs comme sources de données statistiques

Éléments	Recensements	Enquêtes	Dossiers administratifs
1. Couverture	Visent à une couveture à 100% de la population	Certaines enquêtes ne comprennent pas certains secteurs de la population (par exemple, les réserves indiennes, les régions éloignées)	Les populations visées sont définies par les exigences administratives
2. Contenu	La grande diversité des éléments de données permet beaucoup de classements recoupés	Couvrent habituellement un petit nombre de sujets mais de façon plus approfondie qu'un recensement	Limités aux variables requises pour des fins administratives
3. Concepts/définitions	Peuvent être établis en fonction des besoins de l'analyse sociale et économique	Peuvent être établis en fonction des besoins de l'analyse sociale et économique	Établis en fonction des exigences administratives
4. Estimations relatives aux petites régions	Disponibles par suite des efforts déployés pour arriver à une couverture complète	Ne peuvent être produites dans la plupart des cas	Peuvent être produites, pourvu que chaque dossier ait un code géographique correspondant à une petite région
5. Contrôle qualitatif	Peut être conçu pour réduire au minimum les erreurs	Étant de plus petite taille que les recensements, ils permettent un contrôle encore plus serré que dans les recensements	Les dossiers administratifs sont soumis au contrôle de l'organisme administratif et le contrôle qualitatif n'attire pas l'attention, sauf peut-être pour les variables clés
6. Coût	Très élevé	Coût relativement faible par enquête, même si le coût cumulatif d'une enquête habituelle portant sur une période intercensitaire de cinq ans peut être élevé	Relativement peu élevé si les coûts initiaux de collecte sont imputés au programme administratif
7. Fréquence	Tous les 5 ou 10 ans (selon les sujets)	Peuvent être annuelles, trimestrielles ou mensuelles selon les sujets	Peuvent être annuels ou mensuels selon le programme administratif
8. Actualité des données	Les données peuvent être obtenues six mois à 2½ ans après le jour du recensement	Les enquêtes habituelles répétées produisent des résultats en quelques semaines, alors que les enquêtes spéciales peuvent nécessiter plusieurs mois	Selon le processus administratif, alors qu'il se peut qu'une forme utilisable d'un fichier annuel ne soit pas prête avant qu'une bonne partie de l'année suivante soit écoulée
9. Stabilité	Les changements sont sous le contrôle des statisticiens qui répondent aux besoins des utilisateurs	Dans des enquêtes répétées, les changements sont peu fréquents pour permettre les comparaisons temporelles	Des changements peuvent être apportés en raison de modifications législatives ou de modifications de règlements ou en raison de modifications des pratiques administratives
10. Fardeau du répondant	Lourd mais peu fréquent; on le réduit par l'utilisation d'échantillonnage	Léger en moyenne, bien que lourd pour ceux qui ont été choisis	Aucun fardeau supplémentaire

4. DONNÉES ADMINISTRATIVES ET ENQUÊTES-ENTREPRISES

Un projet de remaniement complet de l'infrastructure et de la stratégie sur lesquelles s'appuie le programme d'enquêtes-entreprises est actuellement en cours à Statistique Canada. Il s'agit en particulier de procéder à la refonte du registre d'entreprises (qui sert de base de sondage aux enquêtes-entreprises), de repenser le rôle et l'utilisation des données fiscales et d'élaborer une stratégie cohérente pour la conception d'enquêtes-entreprises tenues une ou plusieurs fois par an. Ce remaniement a été entrepris car il fallait:

(a) éliminer les faiblesses évidentes qui existent actuellement sur le plan de la qualité des données;
(b) harmoniser les données provenant de différentes enquêtes;
(c) réduire au minimum le fardeau de réponse en utilisant autant que possible les données fiscales;
(d) réduire la quantité de ressources nécessaires à la mise à jour des bases de sondage des enquêtes.

Colledge (1987) donne une description détaillée de ce projet.

Les données fiscales et celles qui sont extraites des comptes de retenues sur la paye jouent un rôle primordial dans la réalisation des enquêtes-entreprises. En vertu de la Loi sur la statistique, Statistique Canada a accès aux déclarations d'impôt annuelles des sociétés (T2) et des particuliers (T1). Les retenues d'impôts que les employeurs font sur la paye de leurs salariés sont également à sa disposition. Ces données remplissent deux fonctions distinctes à Statistique Canada:

(i) elles servent à la mise à jour de la base de sondage (liste d'entreprises);
(ii) elles remplacent les données concernant l'impôt sur le revenu qui pourraient être recueillies au moyen d'une enquête.

4.1 Mise à jour de la base de sondage

La mise à jour de la liste d'entreprises est une tâche complexe pour deux raisons principales: premièrement, la structure d'un grand nombre d'entreprises de même que les liens qui existent entre elles sont loin d'être simples et,deuxièmement, il est difficile de se tenir au courant du nombre élevé de "naissances" et de "morts" qui surviennent dans le secteur des petites entreprises. La définition du terme "entreprise" est elle-même une tâche délicate. Il faut établir une distinction entre la structure juridique (entreprise constituée en société), la structure opérationnelle (comment l'entreprise est organisée pour mener ses activités) et la structure statistique (les unités sur lesquelles on a besoin de données pour effectuer les analyses). À l'intérieur de chaque structure, il est possible d'établir une hiérarchie. Ainsi, la structure statistique se compose, de haut en bas, des entreprises, des sociétés, des établissements et des emplacements. Pour tenir à jour la base de sondage, il ne suffit pas d'y ajouter les nouvelles unités (naissances) et de supprimer celles qui ont mis fin à leurs activités (morts); il faut également tenir compte des changements qui se produisent dans les liens entre les unités des entreprises à structure complexe, et cela veut aussi dire le lien entre la hiérarchie statistique et la hiérarchie opérationnelle.

La stratégie proposée consisterait à faire une mise à jour continue de la structure des entreprises qui dépassent une certaine taille (le seuil variant selon la branche d'activité) et du lien entre cette structure et les unités déclarantes du point de vue de l'impôt. L'activité économique des entreprises touchées représenterait environ 70% de l'activité du secteur auquel elles appartiennent. L'"établissement du profil" est un travail qui permet de déterminer la structure interne des entreprises complexes. Il consiste à interviewer des dirigeants d'entreprise pour prendre connaissance de la structure opérationnelle et établir les unités statistiques adéquates. Le système

de retenue sur la paye (RP) de Revenu Canada est une source importante de renseignements sur les changements qui surviennent dans les entreprises (naissances et réorganisations). La mise en activité d'un nouveau compte de RP par un employeur est le signal qu'il y a du nouveau. On procède alors à un suivi qui permet de savoir si une mise à jour s'impose. Les autres signaux proviennent des déclarations d'impôt annuelles, des réponses données aux enquêtes normales et de l'établissement routine des profils.

On n'essaye pas de déterminer les unités qui composent les petites entreprises ni le lien qui existe entre ces unités, car il y a trop de mouvement dans ce secteur. On se sert carrément des données administratives. Deux listes d'entreprises peuvent servir de base de sondage pour les enquêtes: la première est établie à partir des déclarations d'impôt annuelles les plus récentes et la seconde, à partir des comptes actifs de RP. Toutes les entreprises dont la taille dépasse le seuil fixé sont enlevées de ces listes. Ces dernières se superposent et, pour chaque enquête, on se sert de celle qui convient le mieux. Pour les enquêtes infra-annuelles, on privilégie la liste établie à partir des comptes de RP, qui est plus à jour car il est possible d'ouvrir ou de fermer un compte à n'importe quel moment de l'année. Par contre, on n'y trouve pas les entreprises qui n'ont pas d'employés.

4.2 Remplacement des données d'enquête

Afin de réduire le fardeau de réponse et les coûts, on remplace autant que possible les données d'enquête par les données fiscales. Les notions de base et les définitions sur lesquelles ces dernières sont fondées ne coïncident pas toujours avec celles que les enquêtes doivent utiliser pour des raisons d'uniformité avec le système de comptabilité nationale ou pour les besoins d'analyses. Il faut donc prendre soin d'extraire des déclarations d'impôt seulement les éléments d'information qui correspondent le plus aux définitions de l'enquête. En outre, les données fiscales ne couvrent pas l'éventail de variables que de nombreuses enquêtes-entreprises cherchent à étudier. Notamment, on n'y trouve pas de statistiques sur la production.

Un autre problème que pose l'utilisation des données fiscales concerne l'établissement du lien qui existe entre l'unité pour laquelle on a une déclaration d'impôt et l'unité ou les unités visées par l'enquête. On rencontre ce problème particulièrement dans le cas des grosses entreprises à structure complexe dont il a été question plus tôt.

La stratégie élaborée relativement aux enquêtes annuelles consiste à utiliser les données fiscales principalement pour les petites entreprises où il y a, en général, une déclaration d'impôt par unité d'exploitation. Le fardeau de réponse des petites entreprises est ainsi nettement réduit, et la qualité des données finales n'en souffre pas trop puisqu'on obtient la plus grande partie des renseignements dont on a besoin sur l'activité économique au moyen des déclarations que les grandes entreprises font dans le cadre des enquêtes.

Il ressort clairement de ce bref aperçu des nouvelles stratégie et infrastructure établies pour les enquêtes-entreprises que la poursuite de ce programme est fondamentalement tributaire des données fiscales. Statistique Canada et Revenu Canada doivent donc travailler en collaboration très étroite afin que l'incidence de toute réforme, administrative ou autre, du système fiscal puisse être évaluée avant qu'elle n'entre en vigueur et afin que les mesures nécessaires soient prises à temps.

5. DONNÉES SOCIO-ÉCONOMIQUES TIRÉES DE DOSSIERS ADMINISTRATIFS

À la fin des années 70, on a commençé à faire un effort concerté pour utiliser les données sur les particuliers, les familles et les ménages qui se trouvaient dans les dossiers administratifs. Cette démarche était motivée à l'origine par la croissance du coût des enquêtes et la nécessité

de trouver des solutions de rechange moins onéreuses. Il est vite devenu évident que la meilleure façon de réaliser, au Canada, le potentiel statistique des dossiers administratifs sur les particuliers serait de s'en servir non pas pour remplacer le recensement quinquennal, mais pour le compléter en allant chercher des données pour les petites régions dans ces dossiers pendant la période intercensitaire. Comparativement au recensement, les systèmes de dossiers administratifs actuels ne permettent pas d'obtenir le même niveau de couverture, une aussi grande précision géographique ni un éventail comparable de renseignements sur les caractéristiques des particuliers, des familles et des ménages. On continue néanmoins de chercher à atteindre le même taux de couverture que le recensement en combinant plusieurs systèmes de dossiers administratifs et de voir s'il est possible de remplacer les réponses fournies à certaines questions du recensement par des données tirées de sources administratives.

Nous allons examiner l'utilisation des dossiers administratifs pouvant compléter les données de recensement pendant la période intercensitaire. Les travaux expérimentaux réalisés jusqu'à présent ont porté surtout sur les systèmes administrés à l'échelle nationale (p. ex., l'impôt sur le revenu, l'assurance-chômage (A.-C.), les allocations familiales et la sécurité de la vieillesse) plutôt que sur les systèmes administrés au niveau provincial ou infraprovincial (p. ex., l'assurance-maladie, les permis de conduire et les évaluations municipales). Ces derniers posent, en plus des problèmes propres à l'utilisation de tous les dossiers administratifs, celui du manque d'uniformité entre les secteurs de compétence.

Le fichier des déclarations d'impôt annuelles des particuliers (T1) est la principale source de statistiques sur les Canadiens. Au départ, on s'en est servi pour des totalisations directes qui ont permis de produire des statistiques sur le revenu et l'activité selon l'âge et le sexe, aux niveaux provincial et infraprovincial. Le lieu de résidence des déclarants est déterminé au moyen du code postal indiqué dans le dossier. Un fichier de conversion permettant d'établir la correspondance entre les codes postaux et les diverses divisions géographiques de recensement (province, comté, municipalité, circonscription électorale, etc.) a été créé. Il est également possible de produire des totalisations spéciales pour les régions définies par les utilisateurs en fonction des codes postaux.

Les données ainsi obtenues sont, bien sûr, fondées sur des notions, des définitions et des règlements découlant de la Loi de l'impôt sur le revenu. Les définitions ne sont donc pas nécessairement les mêmes que celles qui permettent d'effectuer les analyses requises (p. ex., il n'y a pas de données sur certains régimes d'aide sociale qui ne sont pas assujettis à l'impôt). Le revenu peut être ventilé selon la source, par exemple le revenu d'emploi peut être séparé. Les variables pouvant faire l'objet d'un classement recoupé sont limitées; (p. ex., l'âge, le sexe et l'état matrimonial). Bien qu'il y ait une question sur la profession dans le formulaire de déclaration d'impôt, la qualité des réponses et de leur codage n'est pas assez bonne pour qu'on puisse s'en servir à des fins statistiques. La couverture se limite aux personnes qui sont tenues de produire une déclaration d'impôt : les personnes à faible revenu et les personnes à charge sont donc sous-représentées. Il se peut que les modifications apportées à la législation fiscale aient, avec le temps, une incidence marquée sur le taux de couverture, lequel a nettement augmenté, entre 1977 et 1978, lorsqu'il a été décidé que les personnes à faible revenu devraient faire une déclaration si elles voulaient obtenir le nouveau crédit d'impôt pour enfants.

Malgré toutes ces réserves, la totalisation directe des données tirées des fichiers fiscaux permet de produire, pendant la période intercensitaire, des données pour les petites régions sur le revenu qui sont fort utiles. Une publication récente de Statistique Canada fait état de statistiques ainsi établies pour les régions de tri d'acheminement, c.-à-d., les régions regroupant les adresses dont les trois premiers caractères du code postal sont les mêmes (Statistique Canada 1987). Étant donné qu'une principale préoccupation, lorsqu'on publie des données pour les petites régions, est de s'assurer qu'il est impossible d'en déduire des renseignements concernant des particuliers, on ne donne pas de statistiques pour les régions où moins de 100 personnes ont fait des déclarations d'impôt.

Les fichiers des déclarations d'impôt des particuliers sont également utilisés pour produire les estimations de la migration annuelle. Pour cela, les dossiers d'une même personne se rapportant à deux années consécutives sont appariés et le code attribué à la division de recensement (ou au comté) dans chaque dossier est comparé. S'il n'est pas le même, on en déduit que le déclarant a déménagé. Les données démographiques et les renseignements relatifs aux exemptions permettent d'estimer le nombre total de personnes qui ont déménagé avec le déclarant. Dans une dernière étape, comme le fichier fiscal ne couvre pas l'ensemble de la population, on procède à un redressement pour produire l'estimation annuelle du nombre total de migrants. Depuis 1981, les chiffres produits en se servant du fichier fiscal sont utilisés dans le cadre du programme des estimations démographiques de Statistique Canada. Norris et Standish (1983) décrivent en détail la méthode utilisée pour estimer la migration au moyen des dossiers fiscaux.

Nous avons vu plus tôt qu'il était possible d'établir des statistiques sur le revenu des particuliers grâce aux données fiscales; le revenu des familles présente encore plus d'intérêt tant du point de vue des possibilités d'analyse que du point de vue des programmes sociaux qui y sont liés. Pour déduire le revenu des familles au moyen des dossiers fiscaux individuels, il faut pouvoir déterminer quels titulaires de dossiers appartiennent à une même famille et ensuite apparier les données qui les concernent. Des travaux de cette nature sont en cours et les résultats sont prometteurs. Pour une description des méthodes employées et des résultats obtenus, voir Auger (1987).

Le système d'assurance-chômage (A.-C.) constitue la deuxième source importante de données administratives sur les particuliers. Statistique Canada a accès au fichier des demandeurs et à celui des bénéficiaires. Ces deux fichiers contiennent les enregistrements de personnes qui, pour diverses raisons, ont peut-être droit aux prestations. Il ne s'agit pas toujours de chômeurs au sens de la définition internationale de chômage retenue par l'enquête sur la population active (EPA), d'où sont tirées les données publiées sur le sujet. S'il y avait moyen de trouver dans le système de l'A.-C. une catégorie se rapprochant de celle de l'EPA, il serait possible de produire à partir de ses fichiers des données sur le nombre de chômeurs dans les petites régions. Cependant, comme dans le meilleur des cas les catégories de l'A.-C. ne sont pas identiques à celles de l'EPA, il faut donc plutôt chercher à harmoniser ou concilier ces deux sources de données. Notamment, les chiffres mensuels de l'A.-C. sur les petites régions pourraient être pris comme indicateurs de la variation du chômage à l'échelon local et ils pourraient ensuite être corrigés en fonction d'estimations fiables de l'EPA correspondant à un niveau géographique supérieur (p. ex. une province). Plusieurs méthodes d'estimation de ce genre ont été examinées (comme la méthode de l'estimation par régression et celle de l'estimation par le quotient avec maintien de la structure), mais il n'a pas encore été possible de déterminer laquelle est la meilleure. Trottier et Choudhry (1985) donnent une description des travaux effectués dans ce domaine et Feeney (1987) parle d'une approche semblable utilisée dans les travaux australiens. L'utilisation de modèles à séries chronologiques exploitant la structure corrélée de l'erreur dans le temps est une méthode qui semble très prometteuse (Choudhry et Hidiroglou 1987).

Les exemples que nous venons de donner montrent qu'on utilise principalement les données tirées des dossiers administratifs sur les particuliers pour effectuer des totalisations directes ou des estimations, contrairement aux données tirées des dossiers administratifs concernant les entreprises, dont on se sert surtout pour mettre à jour la base de sondage ou carrément remplacer les données d'enquête.

Bien que les domaines d'activité auxquels ces deux exemples ont trait soient en pleine expansion et revêtent beaucoup d'importance à Statistique Canada, ils ne touchent qu'une petite fraction du nombre de fichiers administratifs dont on se sert au Bureau. À titre d'exemple, les dossiers administratifs sont depuis longtemps universellement utilisés en statistique sociale

(éducation, santé, justice) pour la création de bases de sondage et pour l'obtention de données. Les responsables des recherches en cours sur les enquêtes téléphoniques et les registres d'adresses utilisent les dossiers administratifs pour dresser les listes de logements ou de ménages. Une enquête interne récente a dénombré plus d'une cinquantaine de systèmes administratifs actuellement exploités à des fins statistiques. Tous les types de dossiers et toutes les catégories d'utilisation mentionnés aux sections 2 et 3 sont représentés, y compris des registres de maladies et des dossiers renfermant des données sur les véhicules automobiles immatriculés, les mouvements d'aéronefs, la commercialisation du lait, la taxe de vente sur le carburant, les permis de construire municipaux et les droits de douane et d'accise.

6. ACCÈS AUX SYSTÈMES ADMINISTRATIFS ET DROIT DE REGARD SUR LEUR CONTENU

Le tour d'horizon que nous venons de faire de l'utilisation des dossiers administratifs à des fins statistiques montre bien que ces dossiers constituent une source d'information indispensable pour de nombreux programmes de Statistique Canada. Cela nous amène aux mesures que le Bureau peut prendre pour protéger ces sources de données et peut-être rendre ces dernières plus utiles du point de vue statistique. Cette section porte sur les deux principaux problèmes, c'est-à-dire l'accès aux dossiers administratifs et la manière dont on peut influer leur contenu, leur conception et la marche à suivre à leur égard.

6.1 Accès

Le fondement juridique de l'accès aux dossiers administratifs est l'article 12 de la Loi sur la statistique (1971) qui se lit comme suit:

> Une personne ayant la garde ou la charge de documents ou archives conservés dans un département ou dans un bureau municipal, une corporation, entreprise ou organisation et dont on pourrait tirer des renseignements que l'on cherche à obtenir pour les objets de la présente loi ou qui aideraient à compléter ou à corriger ces renseignements, doit en permettre l'accès, à ces fins, à une personne autorisée par le statisticien en chef à obtenir ces renseignements ou cette aide pour le complètement ou la correction de ces renseignements.

Cette disposition, qui semble donner un droit assez étendu à l'accès, comporte cependant des restrictions. Dans certains cas, les lois régissant le processus administratif limitent l'accès aux données administratives ou leur utilisation secondaire. Il en résulte une incompatibilité entre les lois qui, au mieux, a pour effet de retarder les négociations au sujet de l'accès. Dans d'autres cas, l'accès à des fins statistiques est permis de façon explicite.

L'adoption de lois facilitant l'accès aux dossiers administratifs est une condition nécessaire mais non suffisante de l'utilisation productive des dossiers administratifs. Il serait probablement beaucoup plus efficace, lorsqu'on cherche à obtenir l'accès aux dossiers administratifs, d'essayer d'y arriver par la voie de la collaboration (sur le plan de l'élaboration et de l'utilisation de ces dossiers dans un but statistique) que d'avoir recours à des dispositions ou à des sanctions légales. En effet, une fois l'accès obtenu, il est possible de parvenir à l'étape suivante (qui consiste à influer sur la conception ou les méthodes d'utilisation) seulement s'il existe un esprit de collaboration entre l'organisme administratif et l'organisme statistique.

L'accès aux dossiers administratifs dont jouit le Bureau est strictement un phénomène à sens unique. Les microdonnées individuelles vont de l'organisme administratif à l'organisme statistique, et seulement des données agrégées, protégées par la confidentialité, peuvent retourner en sens inverse. La seule exception technique à cette règle se produit lorsque l'organisme administratif dépend de l'organisme statistique pour organiser, préparer, mettre en forme, traiter ou restructurer ses dossiers et que les microdonnées originales sont renvoyées sous une autre forme à l'organisme fournisseur.

6.2 Droit de regard sur les changements

Nous avons déjà mentionné l'incidence que des modifications apportées aux pratiques ou règlements administratifs auraient sur les statistiques produites. Il suffirait qu'on change la clientèle d'un programme, qu'on introduise une mesure incitant les personnes visées par un programme à y participer ou au contraire à ne plus y participer ou bien qu'on modifie la procédure de manière à altérer la qualité ou l'exhaustivité des dossiers, et les séries chronologiques seraient interrompues. L'organisme statistique doit donc être sur ses gardes et réagir chaque fois qu'un changement imposé de l'extérieur risque de se produire.

Il y a, en revanche, des changements que l'organisme statistique aimerait bien voir se réaliser. Le statisticien qui souhaite utiliser des dossiers administratifs éprouve souvent un sentiment de frustration en pensant que ceux-ci seraient beaucoup plus utiles si seulement de petites modifications y étaient apportées. Par exemple, l'ajout d'une nouvelle question, l'utilisation d'une notion différente, l'inclusion de nouveaux sous-groupes ou l'exécution d'une nouvelle vérification de la qualité pourraient améliorer sensiblement la valeur statistique des dossiers. Par contre, pourquoi l'organisme administratif envisagerait-il d'apporter des changements qui ne sont pas requis par le processus administratif, quand ces changements contribuaient probablement dans une certaine mesure à augmenter les coûts et la complexité de ce processus?

L'organisme statistique a donc un défi à relever : celui de persuader les administrateurs que les avantages découlant de ces changements seraient plus grands que les coûts supplémentaires générés. La difficulté vient de ce que ce n'est pas nécessairement le ministère responsable du système administratif qui profite des avantages en question, mais plutôt les ministères décisionnaires ou d'autres utilisateurs qui se servent des données à des fins statistiques.

Il est habituellement plus facile de tenir compte des exigences statistiques au moment de concevoir un système que de modifier un système déjà en cours d'exploitation. Par conséquent, un mécanisme qui permettrait de tenir compte des besoins en données statistiques au stade de la conception ou de la restructuration en profondeur d'un système administratif serait préférable à un mécanisme qui tenterait seulement d'adapter les systèmes en place. En ce moment, au Canada, on parle beaucoup de la réforme de la taxe que le gouvernement songe à instituer. Cette réforme pourrait changer significativement la collecte des données sur les entreprises au Canada. La participation de statisticiens à la conception d'un système comme celui-là ajouterait beaucoup aux avantages statistiques qu'on pourrait en retirer. Bien entendu, l'établissement d'un nouveau système administratif est un événement relativement rare, de sorte qu'il faut également adapter les systèmes en place si l'on veut obtenir des avantages statistiques à court terme. Par ailleurs, le fait qu'il soit relativement exceptionnel qu'on conçoive ou qu'on restructure des systèmes administratifs importants renforce l'argument selon lequel il ne faut rater aucune occasion de prendre part à ce genre de travaux.

6.3 Mécanismes

Il existe divers moyens ou mécanismes permettant à l'organisme statistique d'avoir un certain droit de regard sur les sytèmes administratifs et d'y avoir accès; certains de ces mécanismes sont essentiellement bilatéraux, mettant en jeu l'organisme statistique et un service administratif quelconque, tandis que d'autres concernent l'ensemble du secteur public. En voici quelques exemples:

(i) créer des comités bilatéraux constitués de hauts fonctionnaires qui examineraient les questions intéressant les deux organismes, notamment les problèmes relatifs à la fourniture de données administratives;

(ii) fournir à l'organisme administratif les données statistiques afin de montrer à la fois l'utilité des données et, le cas échéant, les lacunes attribuables aux pratiques administratives;

(iii) offrir des conseils ou des services techniques au service statistique de l'organisme administratif;

(iv) adopter une politique officielle en matière de collecte de renseignements exigeant que tout projet de collecte de données (dans un but statistique ou administratif) soit soumis à l'examen d'un organisme central;

(v) demander que chaque nouvelle proposition de programme soit assortie d'un plan prévoyant la façon d'obtenir les renseignements statistiques nécessaires pour contrôler et évaluer le programme en question;

(vi) promouvoir l'utilisation des définitions statistiques standard (p. ex., ce que l'on entend par les termes "famille", "établissement commercial" ou "chômeur") dans les systèmes administratifs;

(vii) faire en sorte que les vérificateurs des activités gouvernementales recommandent l'utilisation des dossiers administratifs comme moyen économique par excellence de collecte de renseignements;

(viii) suivre une orientation politique favorisant un recours accru à certains systèmes administratifs ou la recherche de solutions de rechange aux enquêtes;

(ix) supprimer les obstacles législatifs qui limitent l'accès aux dossiers administratifs ou à leur utilisation à des fins statistiques.

L'expérience vécue par le Bureau dans ses relations avec les autres ministères fédéraux a été particulièrement fructueuse lorsque des ententes bilatérales étroites ont été conclues. La création de comités supérieurs bilatéraux au début des années 80 a favorisé et parfois assuré l'élaboration de telles ententes. Les mesures prises à l'échelle de l'appareil gouvernemental, par exemple la gestion de l'information et la planification statistique, ont moins bien réussi à favoriser l'utilisation des dossiers administratifs. Les vérifications des opérations gouvernementales et les directives du cabinet ont bien donné une certaine impulsion à des activités visant à accroître l'utilisation des données administratives, mais l'augmentation de cette utilisation dépend elle-même, encore une fois, de l'existence de liens de travail étroits avec certains ministères. Bien qu'il convienne de décrire l'organisme statistique comme un organisme progressif qui tente de briser les barrières irrationnelles faisant obstacle à l'utilisation des données administratives, il faut quand même également reconnaître qu'il puisse exister au sein de l'organisme même une certaine résistance au changement. Les membres du personnel dont les carrières ont été vouées à la conception et à la réalisation d'enquêtes peuvent avoir besoin d'arguments convaincants pour admettre que les restrictions budgétaires et les besoins en données nous obligent à présent à combiner les enquêtes et d'autres méthodes.

Comme les remarques qui ont précédé s'appliquent uniquement aux systèmes administrés par des ministères fédéraux, il convient d'ajouter quelques mots au sujet des dossiers provinciaux. Certaines des mesures mentionnées conviendraient aussi aux dossiers administrés par les provinces; toutefois les rapports avec les administrations infranationales posent un problème fondamental, celui de la conformité à des normes communes. La différence entre les besoins et les priorités des provinces, accentuée par des possibilités technologiques croissantes, entraînera une diversité de plus en plus grande des systèmes administratifs s'il n'existe pas de force centralisatrice. Dans le passé, le Bureau a eu recours à divers mécanismes pour essayer de favoriser une certaine uniformisation, mais avec plus ou moins de succès. Comme dans le cas des services fédéraux qui ont la garde de dossiers administratifs, l'avantage réciproque doit être l'argument principal. Il y a des comités fédéraux-provinciaux dans plusieurs domaines. Le Conseil de la statistique de l'état civil, composé d'officiers provinciaux de l'état civil et de représentants du Bureau, est un exemple d'une longue et fructueuse collaboration. Des comités comme celui-là ont, dans le passé, élaboré des conventions pour la déclaration de certains éléments de données et ils en ont suivi l'application. Par exemple, le système de déclaration des finances municipales a été conçu à la suite de réunions fédérales-provinciales sur les statistiques financières des municipalités.

7. CONFIDENTIALITÉ, PROTECTION DES RENSEIGNEMENTS PERSONNELS ET RELATIONS AVEC LE PUBLIC

Même si l'organisme statistique a le pouvoir légal d'exploiter les dossiers administratifs et même si les organismes administratifs fournissent ces dossiers dans un esprit de grande collaboration, il faut faire très attention à la perception que le public a de l'utilisation des dossiers administratifs à des fins autres que celles pour lesquelles ces dossiers ont été conçus. Comme l'efficacité, sinon la survie, d'un organisme statistique dépend en très grande partie du maintien de la collaboration et de la confiance des répondants, l'organisme en question ne doit pas s'engager à la légère dans une activité qui pourrait saper cette collaboration ou cette confiance.

Dans beaucoup de pays, le public est, depuis quelques années, de plus en plus sensible à tout ce qui concerne la protection des renseignements personnels et, par conséquent, à l'accès à l'information et au contrôle de cet accès. Au Canada, l'adoption de la Loi sur la protection des renseignements personnels en 1982 témoigne de cette préoccupation grandissante. Cette loi dit, entre autres choses et avec certaines exceptions, qu'un répertoire de toutes les banques de renseignements personnels relevant d'institutions gouvernementales fédérales doit être publié périodiquement, que toutes les personnes ont le droit d'avoir accès aux renseignements les concernant contenus dans ces banques de renseignements et que les renseignements personnels ne peuvent servir qu'à des usages qui sont compatibles avec les fins pour lesquelles ils ont été obtenus. Une des exceptions prévues à cette dernière disposition est que des renseignements personnels peuvent être divulgués

> "... à toute personne ou à tout organisme, pour des travaux de recherche ou de statistique, pourvu que ... les fins auxquelles les renseignements sont communiqués ne peuvent être normalement atteintes que si les renseignements sont donnés sous une forme qui permette d'identifier l'individu qu'ils concernent, (et) la personne ou l'organisme (s'engagent) par écrit (auprès du responsable de l'institution) à s'abstenir de toute communication ultérieure des renseignements tant que leur forme risque vraisemblablement de permettre l'identification de l'individu qu'ils concernent." (Loi sur la protection des renseignements personnels 1982 alinéa 8(2) (j)).

Cette disposition régit l'utilisation des dossiers administratifs à des fins statistiques du point de vue de la Loi sur la protection des renseignements personnels. L'article en question est toutefois formulé sous réserve de toute autre loi du Parlement, de sorte que si une disposition d'une loi régissant le processus administratif interdisait une telle utilisation, celle-ci aurait préséance.

Bien qu'il soit reconnu dans la Loi sur la protection des renseignements personnels et dans d'autres lois que les travaux statistiques constituent, dans certaines conditions, une utilisation secondaire légitime des dossiers administratifs, cela ne dissipe pas pour autant les préoccupations du public au sujet d'une utilisation possible des banques de données qui irait à l'encontre des intérêts des particuliers. Il est douteux que les citoyens moyens fassent la différence entre l'utilisation statistique, où l'identité du titulaire de chaque dossier ne présente plus aucun intérêt, et l'utilisation administrative, où il est essentiel de savoir à qui chaque dossier se rapporte. Il serait plus facile d'expliquer cette différence et de s'en servir si l'on pouvait dire sans équivoque que, pour le genre de travaux qui sont faits en statistique, on n'a jamais besoin d'identificateurs. Malheureusement, ce n'est pas le cas. Plusieurs techniques statistiques tout aussi légitimes exigent qu'on se serve d'identificateurs au cours des manipulations intermédiaires des données. Ces techniques supposent toutes une forme quelconque d'appariement des données provenant de différents fichiers ou de différentes sources, et les identificateurs permettent d'apparier les bons dossiers ensemble. Une fois cette opération effectuée, les dossiers peuvent être rendus anonymes si aucun couplage subséquent n'est prévu. Citons, à titre d'exemples de cas où une forme d'identification est nécessaire, le recensement de la population (pour veiller à ce que la couverture soit exhaustive et s'en

assurer par la suite), les études longitudinales pour lesquelles on utilise des dossiers administratifs, les enquêtes épidémiologiques et enfin les études d'évaluation visant à vérifier les réponses dans des questionnaires d'enquête en les comparant à des données de sources administratives. Le fait de devoir expliquer pourquoi on a besoin d'identificateurs alors que l'identité des personnes auxquelles les données se rapportent ne présente aucun intérêt est un défi de taille pour l'organisme statistique.

Une autre source de préoccupation a trait à l'engagement même de respecter la confidentialité. Même si Statistique Canada a toujours été irréprochable en matière de protection de la confidentialité, il y a sans doute des répondants qui sont sceptiques quant à la protection accordée aux renseignements qui les concernent. Il est possible que l'utilisation de recenseurs connus des répondants, en particulier dans les petites localités, ajoute à leur scepticisme. Certains répondants semblent occuper comme principe qu'il se fait beaucoup d'échanges de renseignements entre les ministères fédéraux et, dans certains cas, ne font de distinction entre les différents ministères.

Une autre inquiétude encore aurait trait non pas à la confiance dont seraient dignes les services ayant actuellement la garde de banques de renseignements, mais découlerait de la crainte de n'avoir aucune protection contre une violation éventuelle de la confidentialité, soit par des moyens illégaux soit par des personnes élues qui auraient un point de vue différent sur la protection des renseignements personnels. Pour protéger le public contre cette éventualité, il faudrait qu'on élimine des bases de données statistiques toute information permettant d'identifier les personnes dont il y est question.

Cette préoccupation du public au sujet de la protection et du traitement des renseignements personnels oblige l'organisme statistique à penser à des mesures susceptibles de prévenir ou de réduire les réactions défavorables du public à l'utilisation légitime des dossiers administratifs pour des travaux statistiques. Comme il s'agit là essentiellement d'une question de perception du public, il importe que l'organisme statistique ne cache rien de ses pratiques, et que le public soit bien au courant des mesures en application, notamment:

a) L'information publique destinée aux répondants et aux utilisateurs doit toujours mettre l'accent sur l'importance accordée à la confidentialité de toutes les données individuelles (microdonnées) acquises par l'organisme statistique.
b) Le fait que les microdonnées ne circulent que dans un sens devrait être souligné. L'organisme statistique reçoit des microdonnées, mais il n'en ressort que des données agrégées ou résumées, dont la confidentialité est garantie par leur nature. Cela est vrai aussi bien pour les données de recensement ou les données d'enquête que pour les données tirées de dossiers administratifs.
c) Il faudrait faire valoir les avantages qui découlent de l'utilisation des dossiers administratifs : réduction du fardeau de réponse et allègement du fardeau fiscal pour les contribuables. Ces affirmations devraient s'appuyer sur des chiffres et des faits.
d) Des règles explicites en matière de jumelage d'enregistrements énonçant les conditions dans lesquelles l'organisme statistique peut exercer cette activité permettrait, d'une part, de montrer que le jumelage d'enregistrements ne se fait ni à la légère ni sans contrôle et, d'autre part, de ne pas donner suite aux demandes de jumelage d'enregistrements qui ne rempliraient pas les conditions requises.
e) En vertu de la Loi sur la protection des renseignements personnels, les individus doivent être informés des fins auxquelles les renseignements personnels les concernant sont recueillis. Il faudrait inciter les organismes administratifs à indiquer de façon précise l'utilisation statistique dont ces renseignements pourraient faire l'objet. Même si l'utilisation de données à des fins statistiques constitue une utilisation secondaire acceptable des dossiers administratifs, la mention explicite de ces fins sur le formulaire de collecte permettrait d'éviter les mauvaises surprises.

f) Il faudrait que les moyens pris pour assurer la protection matérielle dont l'utilisation des dossiers administratifs délicats fait l'objet soient évidents, et cette protection devrait peut-être même être plus grande que celle qui est habituellement exercée au Bureau. À Statistique Canada, les divisions qui sont les principales détentrices des données fiscales sont logées dans des locaux à accès limité, situés dans des bâtiments à l'entrée desquelles on effectue des vérifications de sécurité.

g) Le fait que les fichiers statistiques ne sont pas soumis à l'examen des services de sécurité ou de renseignement permet en grande partie d'entretenir le sentiment de confiance que le public éprouve à l'égard de la confidentialité absolue des données fournies à l'organisme statistique. Lorsque le Service canadien du renseignement de sécurité a été créé en 1983, Statistique Canada a obtenu une telle exemption pour ses données (la seule accordée à un organisme fédéral).

Les points que nous venons de mentionner correspondent à des mesures précises qui peuvent être prises pour éviter que le public ne réagisse mal à l'utilisation des dossiers administratifs ou pour contrer une telle réaction une fois qu'elle a eu lieu, mais ce dont l'organisme statistique a surtout besoin, c'est d'un appui politique fort. Le capital politique à gagner d'une réduction visible des coûts et du fardeau de réponse, conjugué à des garanties politiques fermes de protection des données des particuliers, constitue une base solide sur laquelle les politiciens peuvent s'appuyer pour dissiper les inquiétudes du public au sujet de l'utilisation des dossiers administratifs à des fins statistiques. En même temps, ils doivent démentir immédiatement et de façon non équivoque toute suggestion selon laquelle les dossiers statistiques sont utilisés à des fins administratives.

8. CONCLUSION

Les dossiers administratifs sont et demeureront une source de plus en plus utile de données statistiques. Les points forts et les points faibles relatifs des données tirées des systèmes administratifs, du point de vue des coûts, de l'univers visé, et de leur qualité, utilité et actualité par rapport aux données du recensement ou aux données d'enquête dictent la façon la plus efficace d'utiliser ces différentes sources de données. Entre autres utilisations courantes, il y a les totalisations directes, les estimations indirectes, le remplacement des réponses d'enquête, la constitution et la mise à jour de bases de sondage et l'évaluation de données. Ces utilisations sont aujourd'hui assez répandues dans la plupart des programmes statistiques et elles le seront probablement encore davantage à l'avenir.

Au Canada, les dossiers administratifs font à présent partie intégrante de l'appareil statistique. C'est en partie grâce à leur utilisation que Statistique Canada a pu maintenir ses programmes malgré la compression de ses budgets. Du fait même, le fardeau de réponse a été réduit et on a commencé à produire de nouvelles séries de données ou à produire les anciennes plus souvent. Comme nous ne disposons pas de registres administratifs, il nous a fallu porter une attention très particulière aux questions de couverture et d'utilisation combinée de données administratives et de données d'enquête pour nous assurer que les estimations de totaux pour un univers soient valides. Le recours à des techniques de jumelage d'enregistrements, bien que nécessitant un contrôle rigoureux, s'est avéré d'une très grande utilité, notamment en ce qui a trait aux données sur les entreprises, aux études longitudinales du marché du travail et aux enquêtes épidémiologiques.

En raison de leur utilisation croissante des dossiers administratifs, les organismes statistiques sont de plus en plus tributaires des autres organismes pour la fourniture continue des données dont ils ont besoin pour leurs programmes. Quels que soient les textes de loi et les politiques qui régissent les activités de l'organisme statistique, l'établissement d'ententes de

collaboration étroite avec les organismes fournisseurs est extrêmement important. La capacité de l'organisme statistique d'influer sur la conception ou la refonte de systèmes administratifs repose sur une compréhension mutuelle des besoins des deux organismes intéressés. Si l'on pouvait adopter, à l'échelle de l'appareil gouvernemental, une politique ou un principe donnant à Statistique Canada voix au chapitre de la conception des systèmes administratifs ou, plus généralement, de la façon de satisfaire les besoins statistiques des nouveaux programmes, cela faciliterait la tâche de l'organisme statistique mais ne remplacerait pas la création d'un climat de collaboration étroite avec les organismes administratifs.

Il est possible d'envisager divers mécanismes qui permettraient à l'organisme statistique d'avoir accès aux dossiers administratifs et d'influer sur leur conception à l'intérieur de l'appareil gouvernemental. L'applicabilité et l'efficacité de chaque mécanisme dépendront du contexte législatif et politique ainsi que du mandat et du statut de l'organisme statistique au sein de l'appareil gouvernemental. D'après l'expérience de Statistique Canada, l'établissement de relations de travail bilatérales étroites avec les ministères administratifs, fondées sur le principe des avantages réciproques, est l'approche la plus efficace. Il importe que l'utilisation des dossiers administratifs reçoive un appui politique, et cet appui s'est manifesté dans les décisions récentes du gouvernement relativement aux compressions budgétaires.

REMERCIEMENTS

Pour cette communication, nous nous sommes inspirés des idées et des travaux d'un bon nombre de personnes à Statistique Canada. Nous y avons notamment incorporé des passages extraits du document rédigé en collaboration avec Michael Colledge, Ivan Fellegi et John Leyes.

BIBLIOGRAPHIE

AUGER, E. (1987). Family data from the Canadian personal income tax file. Article présenté à la réunion annuelle de 1987 de l'American Statistical Association.

BRACKSTONE, G.J. (1984). Répercussions des changements technologiques sur le recensement. *Estadistica*, 36, 43-60.

CANADA, Loi sur la statistique (1971). *Statuts du Canada 1970-71-72*, c.15.

CANADA, Loi sur la protection des renseignements personnels (1982). *Statuts du Canada 1980-81-82*, c.11.

CHOUDHRY, G.H., et HIDIROGLOU, M.A. (1987). Small area estimation: Some experiences at Statistics Canada. *Proceedings of the 46th Session of the International Statistical Institute*, (en voie de rédaction).

COLLEDGE, M.J. (1987). The Business Survey Redesign Project: Implementation of the new strategy at Statistics Canada. *Proceedings of the Third Annual Research Conference, US Bureau of the Census*, (en voie de rédaction).

FEENEY, G.A. (1987). The estimation of the number of unemployed at the small area level. Dans *Small Area Statistics, An International Symposium*, (eds. R. Platek, J.N.K. Rao, C. Särndal et M.P. Singh), New York: John Wiley.

NORRIS, D.A., et STANDISH, L.D. (1983). A technical report on the development of migration data from tax records. Document de travail, Statistique Canada.

ROWEBOTTOM, L.E. (1978). The utilization of administrative records for statistical purposes. *Survey Methodology*, 4, 1-15.

STATISTIQUE CANADA (1987). *Tableaux sommaires pour RTA urbaines et codes postaux ruraux.* No. de catalogue 17-602, Statistique Canada.

TROTTIER, I., et CHOUDHRY, G.H. (1985). Model based unemployment estimates for small areas. Dans *Small Area Statistics, An International Symposium ('85 Contributed Papers)*, (éds. R. Platek et M.P. Singh), Laboratory for Research in Statistics and Probability, Carleton University/ Université d'Ottawa.

Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 53–70
Statistique Canada

Propriétés statistiques des estimateurs de la production végétale

CAROL A. FRANCISCO, WAYNE A. FULLER, et RON FECSO[1]

RÉSUMÉ

Le National Agricultural Statistics Service du Département de l'agriculture des États-Unis fait des enquêtes de rendement pour diverses grandes cultures aux États-Unis. Bien que les méthodes d'échantillonnage des champs varient selon les cultures, le plan de sondage demeure le même pour toutes. Cet article présente une analyse de ce plan de sondage et des estimateurs actuellement utilisés. Les auteurs définissent également d'autres estimateurs du rendement et de la production ainsi que des estimateurs de la variance des estimateurs, puis les comparent aux estimateurs courants par une approche théorique et une simulation de Monte Carlo.

MOTS CLÉS: Enquêtes sur les cultures; estimation du rendement; échantillon à deux phases; estimation de variance.

1. INTRODUCTION

Le National Agricultural Statistics Service (anciennement le Statistical Reporting Service) du Département de l'agriculture des États-Unis procède à des enquêtes objectives sur le rendement du maïs, du coton, du soja, du riz, du sorgho à grains, du tournesol et du blé dans les principaux États producteurs. Certains autres pays réalisent le même genre d'enquêtes.

Bien que les méthodes d'échantillonnage des champs varient selon les cultures (taille des parcelles, méthodes de délimitation des parcelles et techniques de mesure appliquées aux légumes et aux fruits), le plan de sondage est le même pour toutes les enquêtes. En effet, les enquêtes objectives sur le rendement reposent sur une méthode d'échantillonnage à quatre degrés. On trouvera à la section 2 une description détaillée de ce plan de sondage. Dans la section 3, nous décrivons et évaluons les estimatuers du rendement moyen des cultures et les estimateurs de la variance. Nous y examinons également d'autres estimateurs. La section 4 renferme les conclusions de l'étude et des recommandations.

2. PLAN DE SONDAGE DE L'ENQUÊTE OBJECTIVE SUR LE RENDEMENT

Les deux premiers degrés de l'échantillonnage du plan de sondage produisent l'échantillon de segments aréolaires qui sert à l'enquête énumérative de juin du National Agricultural Statistics Service (NASS). Dans chaque État, la base aréolaire (base de sondage) est stratifiée suivant le mode d'exploitation du sol. L'État de Californie, par exemple, est divisé en 12 strates d'exploitation. Chacune de ces strates est subdivisée en territoires appelés unités de base. La superficie des unités de base varie; la taille réelle d'une unité de base dépend des renseignements qui existent sur la désignation des frontières, des renseignements complémentaires recueillis, des frontières politiques, etc.. Une fois que les unités de base sont définies, on détermine le nombre de segments aréolaires dans chacune, en divisant la superficie totale d'une unité par la taille de segment désirée. Celle-ci varie selon la strate d'exploitation dans

[1] Carol A. Francisco, Syntex Laboratories Inc., 3401 Hillview Avenue, Palo Alto, Californie 94304; Wayne A. Fuller, Département de statistique, Iowa State University, Ames, Iowa 50011; et Ron Fecso, Survey Research Branch, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C.20250.

laquelle se trouve l'unité de base. En Californie, par exemple, la taille de segment recherchée est ½ mille carré pour les strates de vergers et 1 mille carré pour toutes les autres strates de terres labourables. Une unité de base contient normalement de 1 à 30 segments aréolaires.

Chaque strate d'exploitation est sous-stratifiée en fonction du lieu géographique. Pour former les sous-strates géographiques, on classe les unités de base de chaque strate d'exploitation par comté de manière que des comtés voisins qui ont des caractéristiques agricoles comparables sont fondus en un seul (Fecso 1978), puis on forme de nouveaux groupes de segments aréolaires en suivant l'ordre de classement des unités de base. Ainsi, chaque sous-strate renferme des segments aréolaires qui ont les mêmes caractéristiques agricoles et qui sont voisins. Dans une strate d'exploitation donnée, les sous-strates comptent toutes le même nombre de segments et ont toutes la même superficie, après arrondissement. Pour obtenir des renseignements détaillés sur la conception de la base aréolaire, le lecteur est invité à consulter Fecso et Johnson (1981) et Houseman (1975).

Pour l'estimation de la variance, les sous-strates servent de strates d'échantillonnage. Aussi, les sous-strates d'exploitation seront désignées les strates.

La première étape de l'échantillonnage consiste à prélever des unités de base dans chaque strate. Le nombre d'unités de base prélevées dépend de la nature de la strate. En règle générale, on prélève de 8 à 15 unités dans les strates de terres labourables tandis qu'on en prélève 4 ou 5 dans les strates agro-urbaines, urbaines et non agricoles. Les unités sont prélevées aléatoirement selon une probabilité proportionnelle au nombre de segments aréolaires qu'elles contiennent. La seconde étape consiste à prélever au hasard un segment aréolaire dans chaque unité de base échantillonnée. Ainsi, la probabilité de sélection est la même pour tous les segments aréolaires d'une même strate.

Dans le plan de sondage qui nous intéresse, l'unité de base est l'unité primaire d'échantillonnage. Comme les unités de base sont prélevées selon une probabilité proportionnelle au nombre de segments qu'elles contiennent et qu'un segment est prélevé dans chaque unité de base échantillonnée, on peut considérer le segment comme l'unité primaire d'échantillonnage. Dans notre étude, les deux premiers degrés de l'échantillonnage ne font qu'un et l'échantillon de segments est considéré comme un échantillon aléatoire simple stratifié à un seul degré. Comme la fraction de sondage moyenne est d'environ 1%, nous n'utiliserons pas le terme correctif de la population finie dans notre analyse.

Les troisième et quatrième degrés de l'échantillonnage consistent respectivement à échantillonner des champs et à prélever des parcelles à l'intérieur de ces champs. Dans l'enquête énumérative de juin, on examine chaque segment aréolaire échantillonné pour y recenser les champs qui ont été ensemencés de la plante d'intérêt ou qui doivent l'être. Les champs ainsi recensés sont classés par numéro de segment et par ordre d'énumération à l'intérieur de chaque segment. On prélève ensuite un échantillon systématique de champs selon une probabilité proportionnelle au produit de la superficie du champ par l'inverse de la probabilité de sélection du segment aréolaire qui contient le champ. Ainsi, le nombre de champs échantillonnés par segment varie et il se peut que les grands champs soient échantillonnés plus d'une fois.

Au quatrième et dernier degré de l'échantillonnage, on définit deux parcelles de superficies comparables dans chaque champ échantillonné à l'aide d'une méthode de délimitation aléatoire fondée sur les rangées et les pas. Lorsque les rangées sont difficiles à distinguer l'une de l'autre ou lorsqu'il s'agit de blé, on délimite les parcelles à l'aide d'un nombre aléatoire de pas en bordure du champ et d'un nombre aléatoire de pas vers l'intérieur du champ. Une méthode différente s'applique aussi dans les enquêtes objectives sur le rendement du blé. Pour ce genre d'enquête, on délimite aléatoirement la première parcelle puis on situe la seconde par rapport à la première. Si un grand champ est échantillonné plus d'une fois au troisième degré de l'échantillonnage, on prélève séparément des paires de parcelles additionnelles. Comme les parcelles sont toujours échantillonnées par paire, une paire de parcelles est désignée ici l'unité secondaire. On ne peut avoir plus de huit parcelles (c'est-à-dire quatre unités secondaires) par champ.

3. MÉTHODES D'ESTIMATION

En principe, l'échantillon de l'enquête objective sur le rendement est le résultat d'un sondage à deux phases avec sous-échantillonnage dans la seconde phase. Le tableau 1 donne une description schématique de cet échantillon. L'échantillon produit à la première phase est un échantillon aléatoire simple stratifié de segments. L'échantillon produit à la seconde phase comprend tous les segments qui ne sont pas ensemencés de la culture étudiée de même qu'un échantillon de segments ensemencés de cette culture, ces segments étant prélevés selon une probabilité proportionnelle à la superficie cultivée. L'échantillon de segments est tiré des champs ensemencés de la culture étudiée qui ont été prélevés dans la première phase utilisant l'échantillonnage systématique avec probabilités de sélection proportionnelles à la superficie. Un échantillon d'unités secondaires - où chaque unité secondaire correspond à une paire de parcelles - est ensuite prélevé dans les segments ensemencés qui ont été échantillonnés à la seconde phase.

Puisque l'unité secondaire est toujours une paire de parcelles, nous ne parlerons plus désormais de parcelles mais bien d'unités secondaires. De même, nous oublierons que les champs sont les unités opérationnelles qui servent à délimiter les parcelles et nous parlerons uniquement de segments échantillonnés. Notons que l'on observe deux genres de segments dans la seconde phase; les segments qui sont ensemencés de la culture étudiée et ceux qui ne le sont pas. Le nombre total de segments dans la seconde phase est K. La superficie et la production totale d'un segment qui n'est pas ensemencé de la culture considérée sont connues (toutes deux nulles). En ce qui a trait aux segments ensemencés de la culture considérée, on utilise un sous-échantillon d'unités secondaires pour estimer la production.

Soit M_{hk} le nombre d'unités secondaires incluses dans le segment k de la h-ième strate. On peut supposer, sans perte de généralité, que M_{hk} est égal à A_{hk} (A_{hk} étant la superficie cultivée du segment hk). La validité de cette identité repose essentiellement sur le choix d'une échelle d'équivalence appropriée pour la superficie.

Tableau 1

Méthode d'échantillonnage pour l'enquête objective sur le rendement

Phase/unité d'échantillonnage	Méthode de sélection	Nombre[1] échantillonnées	Données recueillies
Première phase			
Unité primaire d'échantillonnage: segment	probabilité égale à l'intérieur des strates	n_h	superficie cultivée
Seconde phase			
Unité primaire d'échantillonnage: segment	probabilité inégale	K_h	superficie cultivée, production estimée[2]
Unité secondaire d'échantillonnage: paires de parcelles	probabilité égale	m_{hk}	production des parcelles estimée

[1] Par strate pour les unités primaires d'échantillonnage et par segment pour les unités secondaires d'échantillonnage.
[2] La production du segment est nulle si la superficie ensemencée est nulle, autrement elle est estimée à l'aide des données relatives aux parcelles.

Dans la section 3.1, nous étudions l'estimateur de rendement actuellement en usage et voyons à quelles conditions cet estimateur est sans biais pour le rendement moyen d'un État. Dans la section 3.2, nous analysons un estimateur simple de la variance du rendement estimé tandis que, dans la section suivante, nous définissons des estimateurs des variances non conditionnelles des estimateurs du rendement et de la production. Enfin, la section 3.4 décrit une simulation de Monte Carlo appliquée aux estimateurs.

3.1 Estimateur de rendement et estimateur de production actuellement en usage

À l'heure actuelle, on estime le rendement moyen d'un État comme si l'échantillon était un échantillon aléatoire simple d'unités secondaires avec probabilités égales. L'estimateur est le rendement moyen simple des unités secondaires qui sont ensemencés de la culture considérée. En d'autres termes, l'estimation du rendement moyen à l'acre est définie comme étant

$$\bar{y} = D^{-1} \sum_{h=1}^{L} \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell}\, \delta_{hk\ell}, \tag{3.1}$$

où

$$\delta_{hk\ell} = 1 \quad \text{si } A_{hk} > 0,$$

$$\delta_{hk\ell} = 0 \quad \text{si } A_{hk} = 0,$$

$$D = \sum_{h=1}^{L} \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} \delta_{hk\ell}, \tag{3.2}$$

m_{hk} est le nombre d'unités secondaires prélevées dans le segment hk, L est le nombre de strates et $Y_{hk\ell}$ est l'estimation du rendement à l'acre pour l'unité secondaire ℓ du segment hk. Si la superficie cultivée dans un segment (A_{hk}) est nulle, alors $m_{hk} = 1$ et $Y_{hk\ell} = 0$, par définition. Le nombre total d'unités secondaires observées dans les segments ensemencés de la culture en question est D.

L'équation (3.1) peut être réécrite sous la forme opérationnelle suivante:

$$\bar{y} = D^{-1} \sum_{t=1}^{D} Y_t, \tag{3.3}$$

où l'indice inférieur t remplace l'indice triple $hk\ell$ et où la sommation porte sur les unités secondaires comprises dans les segments ensemencés de la culture en question.

L'estimateur du rendement moyen à l'acre (3.1) est un genre d'estimateur par quotient combiné. On peut vérifier cela en se servant des probabilités de sélection conditionnelles pour réécrire $\bar{y}$. Dans le plan de sondage du NASS, les segments sont prélevés systématiquement avec probabilités proportionnelles à la superficie accrue et ceux dont la superficie accrue est suffisamment élevée sont inclus à coup sûr dans l'échantillon. Le nombre d'unités secondaires attribué à cette catégorie de segments est proportionnel à la taille du segment, mise à part l'erreur d'arrondissement. L'arrondissement est effectué suivant le plan d'échantillonnage systématique. Soit $\pi_{hk\ell}$ la probabilité conditionnelle que l'unité secondaire ℓ du segment k de la strate h soit échantillonnée, étant donné l'échantillon de segments prélevés à la première phase d'échantillonnage. Nous avons

$$\pi_{hk\ell} = D \left(\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk} \right)^{-1} N_h n_h^{-1} \tag{3.4}$$

pour les unités secondaires incluses dans les segments dont $A_{hk} > 0$ où N_h est le nombre global de segments inclus dans la strate h, M_{hk} est le nombre d'unités secondaires incluses dans le segment k de la strate h, et n_h est le nombre de segments inclus dans la strate h qui ont été prélevés à la première phase. La probabilité conditionnelle que l'on observe un segment non ensemencé de la culture en question dans la seconde phase est un.

On peut alors exprimer l'estimateur de la moyenne défini en (3.1) par l'équation suivante:

$$\bar{y} = \frac{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} Y_{hk\ell}}{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} \delta_{hk\ell}}, \tag{3.5}$$

où $N_h n_h^{-1}$ est l'inverse de la probabilité de sélection à la première phase, K_h est le nombre de segments prélevés dans la strate h à la seconde phase, et $K = \Sigma K_h$. Étant donné une échelle appropriée, le numérateur de (3.5) est un estimateur de la production totale et le dénominateur est un estimateur de la superficie totale. Il est possible de montrer que le numérateur est un estimateur sans biais en calculant les espérances mathématiques; pour cela, on définit tout d'abord une relation conditionnelle basée sur les unités échantillonnées de la première phase, puis on fait la moyenne pour l'ensemble de ces échantillons. Le dénominateur est un estimateur stratifié du nombre total d'unités secondaires. Par la nature du sondage, le nombre d'unités d'échantillonnage est proportionnel à la superficie, et on peut choisir l'échelle de telle sorte que le nombre d'unités secondaires soit égal à la superficie. Ainsi, on peut considérer $\bar{y}$ comme le rapport d'un estimateur sans biais de la production totale d'une culture à un estimateur sans biais de la superficie totale consacrée à cette culture.

Pour estimer la production totale d'un État, le NASS multiplie $\bar{y}$ par $\hat{A}$, où $\hat{A}$ est l'estimateur de la superficie cultivée totale, qui prend la forme

$$\hat{A} = \sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk}. \tag{3.6}$$

L'estimation de la production totale est donc

$$\hat{Y} = \hat{A}\,\bar{y}. \tag{3.7}$$

3.2 Estimateurs de variance simples

Suivant l'hypothèse d'un échantillonnage aléatoire simple d'unités secondaires parmi toutes les unités secondaires disponibles dans la seconde phase, l'estimation de la variance conditionnelle de y, étant donné les segments de la seconde phase, est

$$\hat{V}_2(\bar{y}) = D^{-1}\,(D-1)^{-1} \sum_{t=1}^{D} (Y_t - \bar{y})^2, \tag{3.8}$$

où l'indice inférieur 2 (par rapport à $\hat{V}$) sert à identifier la variance conditionnelle et l'indice inférieur t (par rapport à Y) remplace l'indice triple $hk\ell$. La somme pour t variant de 1 à D est la somme effectuée pour les D unités secondaires incluses dans les segments ensemencés de la culture en question.

À cause de la simplicité de l'expression (3.8), on a proposé de l'utiliser comme estimateur de la variance non conditionnelle. On a également proposé d'estimer la variance de l'estimation de la production totale d'un État par l'équation suivante:

$$\hat{V}_*(\hat{Y}) = \hat{A}^2 \hat{V}_2(\bar{y}) + \bar{y}^2 \hat{V}(\hat{A}) + \hat{V}(\hat{A}) \hat{V}_2(\bar{y}), \tag{3.9}$$

où $\hat{A}$ est défini en (3.6) et $V(\hat{A})$ est l'estimateur habituel de la variance d'un total estimé stratifié,

$$\hat{V}(\hat{A}) = \sum_{h=1}^{L} N_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2, \tag{3.10}$$

et

$$\bar{A}_h = n_h^{-1} \sum_{k=1}^{n_h} A_{hk}.$$

L'estimateur (3.9) est un estimateur de la variance d'un produit, qui est fondé sur l'hypothèse implicite que $\bar{y}$ et $\hat{A}$ sont non corrélés.

Il est difficile d'évaluer dans quelle mesure l'estimateur (3.9) tend à sous-estimer la variance de $\hat{Y}$. La variance non conditionnelle de $\bar{y}$ est

$$\begin{aligned} V(\bar{y}) &= V_1 \{E_2(\bar{y})\} + E_1 \{V_2(\bar{y})\} \\ &= V_1 \{\hat{A}^{-1} \sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.}\} + E_1 \{V_2(\bar{y})\}, \end{aligned} \tag{3.11}$$

où $Y_{hk.} = M_{hk} \bar{Y}_{hk.}$ est le total pour le segment k dans la strate h, et E_1 et V_1 désignent respectivement l'espérance mathématique et la variance suivant le plan d'échantillonnage de la première phase.

Dans le cas d'un échantillonnage aléatoire simple des unités secondaires, l'estimateur $\hat{V}_2(\bar{y})$ est non biaisé pour le second terme du membre de droite de l'équation (3.11). Comme le plan de sondage du NASS prévoit un échantillonnage systématique dans la seconde phase, $\hat{V}_2(\bar{y})$ est un estimateur biaisé de $V_2(\bar{y})$. La nature et l'importance du biais dépendent de la structure de corrélation des unités de la liste servant à l'échantillonnage de la seconde phase. L'estimateur de la variance $\hat{V}_2(\bar{y})$ est également biaisé parce qu'on a établi la formule de la variance en supposant un échantillonnage avec remise dans la seconde phase. Dans la mesure où l'échantillonnage à la seconde phase se fait sans remise (comme les échantillons sont prélevés systématiquement dans la liste des segments élargis, seuls les segments de grande superficie sont échantillonnés plus d'une fois), $\hat{V}_2(\bar{y})$ surestimera $V_2(\bar{y})$.

L'estimateur $\hat{V}_*(\hat{Y})$ ne renferme pas d'estimateur de $A^2 V_1\{E_2(\bar{y})\}$; cela a pour effet de créer un biais négatif. Ce terme n'est toutefois pas facile à estimer, même suivant l'hypothèse simplificatrice d'un échantillonnage avec probabilité proportionnelle à la taille dans la seconde phase. Aussi, nous analyserons l'efficacité de $\hat{V}_*(\hat{Y})$ par la méthode de Monte Carlo (section 3.4).

3.3 Autres estimateurs de variance

Une autre façon d'estimer $V(\bar{y})$ est de considérer l'échantillon comme le résultat d'un sondage à deux phases (voir tableau 1) et de supposer que la probabilité non conditionnelle qu'un segment soit prélevé à l'intérieur d'une strate pour recevoir une unité secondaire est

proportionnelle à sa probabilité conditionnelle de sélection à la seconde phase, étant donné les segments échantillonnés à la première phase.

Soit π_{hk} la probabilité conditionnelle que le segment k de la strate h soit prélevé dans la seconde phase, étant donné l'échantillon de segments de la première phase. Nous avons

$$\pi_{hk} = min\,(1,\ M_{hk}\pi_{hk\ell})\,, \tag{3.12}$$

où $N_{hk\ell}$ est une constante dans le segment hk. Si $\pi_{hk} = 1$ et que le segment est échantillonné pour recevoir plus d'une unité secondaire, on suppose que les unités secondaires sont prélevées séparément.

Soit π^*_{hk} la probabilité non conditionnelle que le segment k de la strate h fasse l'objet d'une observation dans la seconde phase. Si $A_{hk} = 0$ alors π^*_{hk} est la probabilité non conditionnelle que le segment hk soit échantillonné pour recevoir au moins une unité secondaire. Si $A_{hk} = 0$ alors π^*_{hk} est égale à la probabilité que le segment hk soit échantillonné dans la première phase d'échantillonnage. Posons

$$\begin{aligned} \pi^*_{hk} &= \frac{n_h}{N_h} && \text{si } A_{hk} = 0\,, \\ \pi^*_{hk} &= \pi_{hk}\frac{n_h}{N_h} && \text{si } 0 < \pi_{hk} < 1\,, \end{aligned} \tag{3.13}$$

où π_{hk}, définie en (3.12), est la probabilité conditionnelle que le hk-ième segment soit échantillonné à la seconde phase, étant donné l'échantillon de segments de la première phase.

Dans notre analyse, nous supposons que π^*_{hk} est fixe. Cette hypothèse se vérifiera et π^*_{hk} sera la vraie probabilité non conditionnelle si π_{hk} est un multiple déterminé de M_{hk}, qui aura été déterminé avant l'échantillonnage. L'expression (3.13) sera une approximation si π_{hk} est une fonction des segments prélevés au premier degré d'échantillonnage.

L'expression (3.13) est proportionnelle à M_{hk} pour $M_{hk}\pi_{hk\ell} \leq 1$. Si $M_{hk}\pi_{hk\ell} > 1$, le nombre d'unités secondaires échantillonnées est égal ou supérieur à 1. $M_{hk}\pi_{hk\ell}$ est le nombre exact d'unités secondaires qu'il faut attribuer aux segments pour conserver un échantillon auto-pondéré d'unités secondaires. Il n'y a jamais plus qu'une unité d'écart entre $M_{hk}\pi_{hk\ell}$ et le nombre d'unités secondaires effectivement observées par suite d'un échantillonnage systématique avec probabilité proportionnelle à la taille.

Pour simplifier le reste des calculs, nous supposons que l'échantillonnage systématique ne comporte aucune erreur d'arrondissement. En d'autres termes, nous supposons que le nombre d'unités secondaires observées par segment est exactement le nombre nécessaire pour obtenir un échantillon auto-pondéré. Par conséquent, nous supposons que le nombre d'unités secondaires observées dans un segment prélevé à la seconde phase d'échantillonnage est

$$\begin{aligned} m_{hk} &= 1 && \text{si } 0 < \pi_{hk} < 1\,, \\ m_{hk} &= M_{hk}\pi_{hk\ell} && \text{si } \pi_{hk} = 1\,. \end{aligned} \tag{3.14}$$

Suivant cette hypothèse, un estimateur par quotient combiné du rendement moyen avec probabilités inégales équivaut à l'estimateur défini en (3.1). L'estimateur par quotient combiné est

$$\bar{y}_r = \hat{M}_r^{-1} \sum_{h=1}^{L} \sum_{k=1}^{K_h} \pi^{*-1}_{hk} M_{hk}\bar{y}_{hk.}\,, \tag{3.15}$$

où
$$\bar{y}_{hk.} = m_{hk}^{-1} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell} \qquad \text{si } A_{hk} > 0,$$
$$\bar{y}_{hk.} = 0 \qquad \text{si } A_{hk} = 0,$$
$$\hat{M}_r = \sum_{h=1}^{L} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk}.$$

Dans l'équation (3.15) et les autres équations données dans le reste de cette section, $\hat{M}_r$ (nombre total d'unités secondaires) équivaut à $\hat{A}_r$ (superficie totale); le lecteur peut à son gré substituer l'un à l'autre.

Dans l'analyse qui suit, nous supposons que l'échantillonnage de segments avec remise selon une probabilité proportionnelle à la superficie ensemencée de la culture en question est une approximation de l'échantillonnage systématique avec probabilité proportionnelle à la taille effectué à la seconde phase. Suivant l'hypothèse de l'échantillonnage avec remise, un estimateur de la variance de $\bar{y}$ est défini comme étant

$$\hat{V}(\bar{y}_r) = \hat{M}_r^{-2} \sum_{h=1}^{L} K_h (K_h - 1)^{-1} \sum_{k=1}^{K_h} (\pi_{hk}^{*-1} u_{hk} - \bar{u}_{h.})^2, \tag{3.16}$$

où
$$u_{hk} = M_{hk} (\bar{y}_{hk.} - \bar{y}_r),$$
$$\bar{u}_{h.} = K_h^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} u_{hk}.$$

Un estimateur de la production totale est défini par

$$\hat{Y}_r = N \bar{M}_n \bar{y}_r, \tag{3.17}$$

où
$$\bar{M}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk}.$$

N est le nombre total de segments dans la population et $W_h = N^{-1} N_h$. La formule d'approximation de Taylor appliquée à la variance non conditionnelle de la distribution approximative de $\hat{Y}_r$ est

$$V\{\hat{Y}_r\} = N^2 [\bar{M}_N^2 V\{\bar{y}_r\} + 2\bar{M}_N \bar{\bar{y}}_N C\{\bar{y}_r, \bar{M}_n\} + \bar{\bar{y}}_N^2 V\{\bar{M}_n\}], \tag{3.18}$$

où $\bar{y}_r$ est défini en (3.15), $\bar{M}_n$ est défini en (3.17),

$$\bar{M}_N = N^{-1} \sum_{h=1}^{L} \sum_{k=1}^{N_h} M_{hk},$$
$$\bar{\bar{y}}_N = \left(\sum_{h=1}^{L} \sum_{k=1}^{N_h} M_{hk} \right)^{-1} \sum_{h=1}^{L} \sum_{k=1}^{N_h} Y_{hk.},$$

$Y_{hk.} = M_{hk}\bar{Y}_{hk.}$ est la production totale du k-ième segment de la strate h, et $C\{\bar{y}_r, \bar{M}_n\}$ est la covariance de $\bar{y}_r$ et de $\bar{M}_n$.

En ce qui a trait aux échantillons de taille fixe avec probabilités inégales, l'estimateur $\bar{y}_r(\doteq \bar{y})$ est à peu près conditionnellement non biaisé pour le rendement moyen des $n = \Sigma n_h$ segments de l'échantillon de la première phase. Le rendement moyen des n segments est

$$\bar{\bar{y}}_n = \bar{M}_n^{-1} \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.}\,.$$

Par conséquent, la covariance de $\bar{y}_r$ et $\bar{M}_n$ est la covariance de $\bar{M}_n^{-1}\,\bar{Y}_n$ et $\bar{M}_n$, où

$$\bar{Y}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.}\,.$$

À l'aide de la formule d'approximation couramment utilisée pour les rapports, la covariance de $\bar{y}_r$ et $\bar{M}_n$ est donnée, approximativement, par

$$C\{\bar{M}_n^{-1}\,\bar{Y}_n, \bar{M}_n\} \doteq C\{(\bar{Y}_n - \bar{\bar{y}}_N \bar{M}_n)\bar{M}_N^{-1}, \bar{M}_n\}$$

$$= \bar{M}_N^{-1}[C\{\bar{Y}_n, \bar{M}_n\} - \bar{\bar{y}}_N\, V\{\bar{M}_n\}]\,. \tag{3.19}$$

Si la probabilité que l'on observe le couple $(Y_{hk.}, M_{hk})$ est proportionnelle à π^*_{hk}, l'estimateur de la covariance de $\bar{Y}_n$ et de $\bar{M}_n$ est défini par

$$\hat{C}\{\bar{Y}_n, \bar{M}_n\} = \sum_{h=1}^{L} W_h^2\, n_h^{-1}\, \hat{S}_{MYh}, \tag{3.20}$$

où

$$\hat{S}_{MYh} = K_h(K_n^{-1})^{-1}\left(\sum_{j=1}^{K_h} \pi^{*-1}_{hj}\right)^{-1} \sum_{k=1}^{K_h} \pi^{*-1}_{hk}\,(M_{hk} - \bar{M}^*_h)(M_{hk}\bar{y}_{hk.} - \bar{y}^*_{h..})\,,$$

$$\bar{M}^*_h = \left(\sum_{j=1}^{K_h} \pi^{*-1}_{hj}\right)^{-1} \sum_{k=1}^{K_h} \pi^{*-1}_{hk}\, M_{hk}\,,$$

$$\bar{y}^*_{h..} = \left(\sum_{j=1}^{K_h} \pi^{*-1}_{hj}\right)^{-1} \sum_{k=1}^{K_h} \pi^{*-1}_{hk}\, M_{hk}\bar{y}_{hk.}\,.$$

L'estimateur $\hat{S}_{MYh}$ équivaut à un estimateur par quotient d'Horvitz-Thompson de la moyenne des produits $(M_{hk} - \bar{M}_h)(Y_{hk.} - \bar{Y}_{h..})$ rajusté pour tenir compte du nombre de degrés de liberté. Le facteur de rajustement $K_h(K_h - 1)^{-1}$ est rendu nécessaire parce que la construction du produit exige que l'on remplace les moyennes de la population par les moyennes d'échantillon.

En substituant les équations (3.15), (3.16) et (3.20) aux termes correspondants dans l'équation (3.18), on obtient

$$\hat{V}\{\hat{Y}_r\} = N^2[\bar{M}_n^2\, \hat{V}\{\bar{y}_r\} + 2\bar{y}_r\hat{C}\{\bar{Y}_n, \bar{M}_n\} - \bar{y}_r^2\,\hat{V}\{\bar{M}_n\}]\,, \tag{3.21}$$

où $\hat{V}[\bar{M}_n]$ est l'estimateur de la variance d'une moyenne stratifiée. L'équation (3.21) définit un estimateur de la variance de l'estimation de la production totale d'un État pour un échantillonnage double stratifié. Contrairement à l'estimateur $\hat{V}_*(\hat{Y})$ défini en (3.9), l'estimateur (3.21) ne repose pas sur l'hypothèse de l'absence de corrélation entre l'estimateur du rendement et l'estimateur de la superficie. L'équation (3.21) comporte également un estimateur non conditionnel de la variance du rendement.

3.4. Comparaison d'estimateurs par la méthode de Monte Carlo

Nous avons procédé à une étude par la méthode de Monte Carlo pour illustrer les différences entre divers estimateurs. À cette fin, nous nous sommes servis des données sur la superficie ensemencée de coton tirées de l'enquête énumérative de juin 1983 en Californie et des données de l'enquête objective sur le rendement correspondante de 1983. Pour les besoins de notre analyse, nous avons considéré que le coton était cultivé dans 28 strates.

Le tableau 2 donne la répartition de cette culture entre les 28 strates, conformément aux données de l'enquête énumérative de 1983. Fecso et Johnson (1981) ont décrit les six modes d'exploitation du sol, qui sont identifiés par les deux premiers chiffres du numéro de la strate; ces modes d'exploitation sont définis ci-dessous avec le code correspondant:

1300 - au moins 50 pour cent des terres sont cultivées; principalement des cultures générales et au plus 10 pour cent de la superficie consacrée à la culture des fruits ou des légumes;
1700 - au moins 50 pour cent des terres sont cultivées; principalement culture des fruits, des noix ou du raisin combinée à des cultures générales;
1900 - au moins 50 pour cent des terres sont cultivées; principalement culture des légumes combinée à des cultures générales;
2000 - de 15 à 50 pour cent des terres sont cultivées; culture extensive et foin;
3100 - zones résidentielles et terres agricoles; plus de 20 logements au mille carré;
4100 - moins de 15 pour cent des terres sont cultivées; principalement de grands pâturages privés.

Nous avons simulé une population à partir des résultats de l'enquête énumérative de juin 1983. Le tableau 2 permet de comparer les caractéristiques de la population simulée aux résultats de l'enquête. Dans la population simulée, on pouvait dire que du coton était cultivé dans le segment k ($k = 1, \ldots, N_h$) de la strate h ($h = 1, \ldots, 28$) si $X_{hk} = 1$, où X_{hk} est une variable (aléatoire) de Bernoulli (p_h) indépendante, p_h étant la proportion de segments de la strate h où, selon l'enquête énumérative de 1983, du coton était cultivé.

La seconde étape de la création de la population consistait à attribuer une superficie (en acres) ensemencée de coton aux segments pour lesquels $X_{hk} = 1$. On a donc calculé une série de ratios mettant en relation la superficie ensemencée de coton dans un segment et la superficie moyenne des segments pour les strates d'exploitation qui renfermaient plus d'un segment ensemencés de coton selon l'enquête énumérative de juin 1983. Cet ensemble de ratios a servi à déterminer la superficie ensemencée de coton dans les segments où du coton était cultivé. Si X_{hk} était égal à 1, un ratio (r_{hk}) était prélevé dans l'ensemble de ratios observés de telle sorte que la probabilité de sélection était la même pour tous les ratios de l'ensemble. La superficie (en acres) ensemencée de coton dans le segment hk, M_{hk}, était définie par

$$M_{hk} = r_{hk}\bar{M}_{h.}, \tag{4.1}$$

où $\bar{M}_{h.}$ était la superficie moyenne consacrée à la culture du coton pour les segments de la strate h où, selon l'enquête énumérative de juin 1983, du coton était cultivé. (Voir tableau 2.)

Les résultats de l'enquête objective sur le rendement de 1983 pour le coton ont servi à simuler les observations de rendement à l'intérieur des segments. Les estimations de rende-

Tableau 2

Estimations de la superficie ensemencée de coton selon l'enquête énumérative de juin 1983 en Californie et la superficie ensemencée de coton pour la population simulée

Strate	Taille du segment visée (acres)	Nombre de segments dans la strate	Nombre de segments échant. en 1983	Proportion de segments ensemencés de coton		Superficie moyenne ensemencée de coton dans les segments ensemencés de coton	
				1983	Population simulée	1983	Population simulée
1314	640	291	10	60	60	197	200
1315	640	291	10	100	100	354	348
1316	640	291	10	90	89	167	173
1317	640	291	10	90	92	149	148
1318	640	291	10	50	53	481	422
1319	640	291	10	20	19	249[1]	260
1320	640	291	10	90	91	154	155
1321	640	291	10	60	61	270	274
1322	640	291	10	70	71	205	210
1323	640	291	10	80	79	288	279
1713	320	432	10	30	28	125	122
1714	320	432	10	30	31	58	57
1715	320	432	10	20	22	86[2]	84
1716	320	432	10	10	8	86[2]	89
1717	320	432	10	40	38	26	27
1718	320	432	10	30	29	144	144
1719	320	432	10	30	31	65	67
1720	320	432	10	30	30	38	35
1721	320	432	10	30	29	133	138
1722	320	432	10	50	47	130	131
1723	320	432	10	40	40	76	76
1906	640	362	10	70	73	117	127
1907	640	362	10	70	74	192	194
1908	640	362	10	80	83	253	246
2010	640	649	10	30	31	303	306
2011	640	649	10	40	41	175	165
3107	160	1,847	5	20	22	25[3]	25
4110	2,560	1,044	10	10	10	178	165

[1] Moins de trois segments échantillonnés; ce chiffre représente la moyenne pour tous les segments des sous-strates de la strate d'exploitation 13.

[2] Moins de trois segments échantillonnés; ce chiffre représente la moyenne pour tous les segments des sous-strates de la strate d'exploitation 17.

[3] Moins de trois segments échantillonnés; ce chiffre représente la superficie approximative ensemencée de coton pour cette strate agro-urbaine.

ment étant difficiles à obtenir, on s'est servi d'une variable qui est une composante importante de ces estimations, soit le nombre de plants aux 100 pi^2. Dans l'enquête objective sur le rendement de 1983, le nombre moyen estimé de plants aux 100 pi^2 pour l'ensemble de la population était de 79.6. Le tableau 3 donne, pour chaque strate étudiée, le nombre moyen de plants aux 100 pi^2 selon l'enquête de 1983. La moyenne pour chaque strate est fondée sur toutes les unités secondaires de la strate qui ont été prélevées par échantillonnage avec probabilité proportionnelle à la taille estimée.

Une analyse de variance portant sur les données de 1983 (tableau 4) a montré que 28% de la variabilité totale entre les unités secondaires était attribuable à des différences entre segments à l'intérieur des strates ($s_b^2 = 378.0$), tandis que 58% de cette variabilité était attribuable à la variation entre unités secondaires à l'intérieur des segments ($s_w^2 = 776.6$). Si la

Tableau 3

Nombre moyen de plants aux 100 pi^2 selon l'enquête objective sur le rendement de 1983 pour le coton en Californie et dans la population simulée

Strate	Nombre moyen de plants aux 100 pi^2	
	Enquête objective sur le rendement de 1983	Population simulée
1314	78	76
1315	80	80
1316	67	68
1317	72	73
1318	80	80
1319	93	93
1320	92	91
1321	70	69
1322	84	84
1323	72	71
1713	118	117
1714	96[1]	95
1715	96[1]	93
1716	96[1]	86
1717	96[1]	96
1718	139	140
1719	96[1]	97
1720	96[1]	97
1721	89	86
1722	79	79
1723	84	85
1906	98	98
1907	67	67
1908	53	53
2010	118	118
2011	47	47
3107	80[2]	79
4110	60	59

[1] Moins de trois unités secondaires observées; ce chiffre représente la moyenne pour les unités secondaires de la strate d'exploitation 17.

[2] Moins de trois unités secondaires observées; ce chiffre représente la moyenne pour les unités secondaires de toutes les strates.

composante de la variance liée à la strate est tenue pour constante, 67% de la variation à l'intérieur des segments est attribuable à la variabilité entre les unités secondaires.

Lorsqu'il avait été établi qu'un segment comprenait des champs de coton, le nombre moyen de plants aux 100 pi^2 pour le segment hk était simulé par

$$\bar{c}_{hk} = \bar{c}_{h.} + e_{hk}, \tag{4.2}$$

où $\bar{c}_{h.}$ est le nombre moyen de plants aux 100 pi^2 pour la strate h, e_{hk} est distribuée suivant une loi normale $N(0, s_b^2)$, et $s_b^2 = 378.0$. Si la moyenne obtenue par la simulation ($\bar{c}_{hk}$) équivalait à moins de 10% de la moyenne pour la strate, on fixait la valeur de c_{hk} à $(.10)\bar{c}_{h.}$. Le tableau 3 permet de comparer les moyennes obtenues pour la population simulée avec celles obtenues par l'enquête objective sur le rendement de 1983. La moyenne globale pour la population simulée était $\bar{\bar{y}}_N = 79.6$.

Tableau 4

Analyse de variance pour les données de l'enquête objective sur le rendement de 1983

Source	Degrés de liberté	Sommes des carrés	Carré moyen	Composante de la variance	Pourcentage de la variabilité totale
Strate	26	80,193	3,084.3	187.3	14
Segments à l'intérieur de la strate	85	124,086	1,459.8	378.0	28
Résidu	103	79,991	776.6	776.6	58
Total	214	284,270		1,341.9	100

Cinq cents échantillons de l'enquête énumérative de juin ont été tirés de la population simulée au moyen d'un échantillonnage aléatoire stratifié. Deux cent soixante-quinze segments ont été échantillonnés dans chaque cas. Le nombre de segments prélevés dans chaque strate était identique au nombre prélevé pour l'enquête énumérative de juin 1983 (voir tableau 2). Pour chacun des échantillons simulés, on a calculé la superficie moyenne (en acres) des segments dans la population et les probabilités conditionnelles, $\{\pi_{hk}\}$, de (3.12), que les segments de l'échantillon recevraient des parcelles par suite d'un tirage. Les probabilités conditionnelles ainsi calculées ont été utilisées au second degré de l'échantillonnage systématique avec un seul départ et probabilité proportionnelle à la taille estimée (voir section 2). Afin de simuler des échantillons de l'enquête objective de rendement on a prélevé 220 unités secondaires en se servant de cette méthode d'échantillonnage systématique. Les unités secondaires ainsi prélevées constituent en soi un échantillon de l'enquête objective sur le rendement. Deux échantillons de ce genre ont été simulés pour chacun des 500 échantillons de la population simulée de l'enquête énumérative de juin.

Lorsqu'un segment était sélectionné pour recevoir une unité secondaire, on simulait le rendement (nombre de plants aux 100 pi^2) observé à l'intérieur d'un champ en supposant l'invariabilité du coefficient de variation à l'intérieur de chaque segment. Le nombre de plants observé était défini comme étant

$$y_{hk\ell} = \bar{c}_{hk} + s_w \bar{\bar{y}}_N^{-1} \bar{c}_{hk} f_{hk\ell}, \tag{4.3}$$

où $y_{hk\ell}$ est le nombre moyen estimé de plants aux 100 pi^2 pour la ℓ-ième unité secondaire du segment k de la strate h et $f_{hk\ell}$ est distribuée suivant une loi normale $N(0, 1)$. L'erreur type à l'intérieur des segments est la racine carrée de $s_w^2 = 776.6$, qui figure dans le tableau 4, et $\bar{\bar{y}}_N$ est le nombre moyen global de plants par parcelle. Si $y_{hk\ell}$ équivalait à moins de 10% de la moyenne pour la strate, on fixait sa valeur à $(.10)\bar{c}_{hk}$. De même, si $y_{hk\ell}$ équivalait à plus de 190% de la moyenne pour la strate, sa valeur était fixée à $(1.9)\bar{c}_{hk}$.

Le tableau 5 donne un résumé des résultats des simulations pour les superficies ensemencées de coton. La superficie moyenne estimée (en acres) par segment est

$$\bar{A}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk}, \tag{4.4}$$

et la variance estimée correspondante est

$$\hat{V}(\bar{A}_n) = \sum_{h=1}^{L} W_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2. \tag{4.5}$$

Tableau 5

Estimations de la superficie ensemencée de coton établies à partir de 500 échantillons simulés de l'enquête énumérative de juin

	$\bar{A}_n$	$\hat{V}(\bar{A}_n)$
Moyenne	9.93	0.64
Intervalle	8.13 - 12.21	
Variance	0.66	0.016

La superficie moyenne ensemencée de coton (par segment) pour la population simulée est de 9.94 acres tandis que la moyenne des estimations des 500 échantillons était de 9.93 acres. La variance réelle de l'estimateur stratifié $\bar{A}_n$ est 0.63 tandis que la variance moyenne estimée pour les 500 échantillons simulés était 0.64. À cause de la faible variabilité de l'estimation de la superficie consacrée à la culture du coton, π^*_{hk} produit une estimation stable de la probabilité non conditionnelle que le segment k de la strate h soit échantillonné pour recevoir au moins une unité secondaire.

Outre les estimateurs analysés ci-dessus, nous avons construit des estimateurs de la variance par les groupes aléatoires. Deux ensembles de groupes aléatoires ont été formés pour chaque échantillon de l'enquête objective sur le rendement. Le premier ensemble contenait cinq groupes ($\gamma = 5$) et l'autre, dix ($\gamma = 10$). Les groupes aléatoires ont été créés par la formation de sous-ensembles avec les unités primaires d'échantillonnage (c'est-à-dire les segments) à l'intérieur de chaque strate d'exploitation. On a formé le premier groupe de chaque ensemble en tirant un échantillon aléatoire simple sans remise de taille $K_{h(\gamma)} = n_h/\gamma$ dans chaque strate ($h = 1, \ldots, 28$) de l'échantillon parent de l'enquête énumérative de juin. Le second groupe a été formé de la même manière par tirage de $K_{h(\gamma)}$ segments parmi les $n_h - K_{h(\gamma)}$ segments qui restaient dans chaque strate. Les autres groupes aléatoires ont été formés d'une manière semblable. Comme l'échantillon de la strate d'exploitation n° 3107 ne comptait que cinq segments ($n_h = 5$), on a répété les valeurs de superficie et de rendement des cinq segments pour obtenir les dix observations nécessaires à la formation des dix groupes quand $\alpha = 10$.

Soit D_α le nombre d'unités secondaires ensemencées de coton qui ont été prélevées dans le groupe aléatoire α ($\alpha = 1, \ldots, \gamma$) au cours de l'enquête objective sur le rendement. Soit $\bar{y}_{(\alpha)}$ l'estimateur de rendement obtenu pour l'α-ième groupe aléatoire:

$$\bar{y}_{(\alpha)} = D_\alpha^{-1} \sum_{t=1}^{D_\alpha} Y_{t(\alpha)}, \tag{4.6}$$

où $\bar{y}_{(\alpha)}$ est l'équivalent de l'équation (3.3) pour l'α-ième groupe. L'estimateur de la variance de $\bar{y}$ par les groupes aléatoires prend la forme suivante:

$$\hat{V}_{g\gamma}(\bar{y}) = \gamma(\gamma - 1)^{-1} \sum_{\alpha=1}^{\gamma} (\bar{y}_{(\alpha)} - \bar{y})^2. \tag{4.7}$$

Cet estimateur est légèrement biaisé dans le cas de l'estimateur de rendement se rapportant à la série de dix groupes parce que la strate n° 3107 ne comporte que cinq observations et que celles-ci se répètent dans les divers groupes.

De même, posons $\hat{Y}_{(\alpha)}$ comme l'estimateur de la production totale pour l'α-ième groupe aléatoire:

$$\hat{Y}_{(\alpha)} = N\,\bar{M}_{n(\alpha)}\bar{y}_{(\alpha)}, \tag{4.8}$$

ou

$$\bar{M}_{n(\alpha)} = \sum_{h=1}^{L} W_h K_{h(\alpha)}^{-1} \sum_{k=1}^{K_{h(\alpha)}} M_{hk(\alpha)},$$

$M_{hk(\alpha)}$ est la superficie (en acres) consacrée à la culture du coton dans le segment k de la strate h pour le groupe aléatoire α et $K_{h(\alpha)}$ est le nombre de segments compris dans la strate h pour l'α-ième groupe. L'estimateur de la variance de $\hat{Y}$ par les groupes aléatoires est alors défini comme étant

$$\hat{V}_{g\gamma}(\hat{Y}) = \gamma(\gamma-1)^{-1} \sum_{\alpha=1}^{\gamma} (\hat{Y}_{(\alpha)} - \hat{Y})^2. \tag{4.9}$$

Les tableaux 6 et 7 présentent un sommaire des résultats de la simulation de Monte Carlo pour les estimateurs du rendement et de la production. On y trouve les valeurs moyennes des estimations et leurs estimations de variances pour les 1,000 échantillons simulés de l'enquête objective sur le rendement. En simulant deux échantillons de l'enquête objective sur le rendement pour chaque échantillon simulé de l'enquête énumérative de juin, nous avons pu estimer la variance entre enquêtes énumératives et la variance à l'intérieur des enquêtes énumératives.

L'estimateur actuellement en usage, $\bar{y}$, défini en (3.1), et l'estimateur par quotient combiné $\bar{y}_r$, défini en (3.17), qui est fondé sur les probabilités π_{hk}^* établie à partir des résultats de l'enquête énumérative de juin, produisent tous deux des estimations de précision comparable (voir tableau 6). Cette similitude est en partie attribuable à l'exactitude avec laquelle on estime les probabilités de sélection non conditionnelles dans chaque échantillon.

Comme nous l'avons vu dans la section 3.2, l'estimateur de la variance conditionnelle $\hat{V}_2(\bar{y})$ produit une sous-estimation de $V(\bar{y})$. Dans cette étude de simulation $\hat{V}_2(\bar{y})$ a produit une estimation qui est de 38% inférieure à la variance observée de $\bar{y}$. En effet, la variance observée de $\bar{y}$ était de 11.57, comparativement à une moyenne de 7.21 pour $\hat{V}_2(\bar{y})$. Cette sous-estimation de la variance a été constatée dans l'ensemble des échantillons. La variance estimée de $\hat{V}_2(\bar{y})$ était de 0.99, pour des valeurs de $\hat{V}_2(\bar{y})$ allant de 3.85 à 11.24 pour les 1,000 observations. Ainsi, la valeur maximale observée pour l'estimation de la variance conditionnelle était inférieure à la vraie variance.

Suivant l'hypothèse d'un échantillonnage de segments avec probabilité proportionnelle à la taille et remise en seconde phase, $\hat{V}_2(\bar{y})$ est un estimateur sans biais de la variance conditionnelle de $\bar{y}$, étant donné l'échantillon de segments prélevés au premier degré, comme nous l'avons vu dans la section 3.2. Selon la simulation de Monte Carlo, une estimation de l'espérance mathématique de la variance conditionnelle de $\bar{y}$, $V_2(\bar{y})$, est 3.97. L'écart appréciable entre cette estimation et la moyenne indiquée au tableau 6 (3.97 contre 7.21) peut être attribuable au fait que l'estimateur $\hat{V}_2(\bar{y})$ ne tient pas compte des effets de la stratification dans la population (voir tableaux 2 et 3) et que $\hat{V}_2(\bar{y})$ est calculée suivant l'hypothèse d'un échantillonnage de segments avec remise au second degré du sondage.

L'estimateur (3.9), $\hat{V}_*(\hat{Y})$, produit une sous-estimation de la variance non conditionnelle de $\hat{Y}$. En effet, la variance observée de $\hat{Y}$ dans la simulation de Monte Carlo est 49.69 (millions)^2 tandis que la valeur moyenne de $\hat{V}_*(\hat{Y})$ n'est que 40.85 (millions)^2. Cette sous-estimation (18%) de la vraie variance est attribuable à un certain nombre de facteurs. Nous avons vu plus tôt que $\hat{V}_2(\bar{y})$ comporte un biais négatif lorsqu'il sert d'estimateur de $\hat{V}(\bar{y})$.

Tableau 6

Propriétés des estimations du rendement à l'acre et des estimations de variances selon la simulation de Monte Carlo[1]

	Estimateur					
	$\bar{y}$	$\hat{V}_2(\bar{y})$	$\hat{V}_{g5}(\bar{y})$	$\hat{V}_{g10}(\bar{y})$	$\bar{y}_r$	$\hat{V}(\bar{y}_r)$
Moyenne	79.74	7.21	12.62	12.39	79.76	12.39
Variance totale	11.57	0.99	74.58	36.86	11.56	12.51
Variance entre les EEJ	7.60	0.48	6.10	4.56	7.64	7.61
Variance à l'intérieur des EEJ	3.97	0.51	68.48	32.30	3.92	4.90

[1] Deux échantillons de l'enquête objective sur le rendement ont été simulés pour chacun des 500 échantillons de l'enquête énumérative de juin.

Tableau 7

Propriétés des estimations de la production et des estimations de variances selon la simulation de Monte Carlo[1]

	Estimateur[2]					
	$\hat{Y}$	$\hat{V}_*(\hat{Y})$	$\hat{V}_{g5}(\hat{Y})$	$\hat{V}_{g10}(\hat{Y})$	$\hat{Y}_r$	$\hat{V}(\hat{Y}_r)$
Moyenne	73.04	40.85	48.99	48.53	73.07	48.73
Variance totale	49.69	82.52	1245.10	608.80	49.58	222.96
Variance entre les EEJ	46.35	78.17	50.82	208.48	46.30	199.58
Variance à l'intérieur des EEJ	3.34	4.35	1194.28	400.32	3.28	23.38

[1] Deux échantillons de l'enquête objective sur le rendement ont été simulés pour chacun des 500 échantillons de l'enquête énumérative de juin. Il y avait $N = 92{,}240$ segments dans la population simulée.
[2] L'estimateur $\hat{Y}$ est exprimé en millions d'unités et les variances le sont dans les unités correspondantes.

Par ailleurs, cet écart est dû au fait que $\hat{V}_*(\hat{Y})$ ne tient pas compte de la covariance de $\bar{M}_n$ et de $\bar{y}$. Dans l'exemple qui nous occupe, le biais lié au second facteur contrebalance en partie celui lié au premier.

L'utilisation de l'expression (3.16), $\hat{V}(\bar{y}_r)$, pour estimer la variance de $\bar{y}_r$ et de l'expression (3.21), $\hat{V}(\hat{Y}_r)$, pour estimer la variance de $\hat{Y}_r$ a donné des résultats beaucoup plus satisfaisants que ceux obtenus avec les estimateurs présentement en usage. La moyenne des estimations $\hat{V}(\bar{y}_r)$ a été de 12.51 dans la simulation de Monte Carlo, ce qui représente une surestimation d'environ 7% par rapport à la variance observée de $\bar{y}_r$ (11.57). Environ le tiers de cet écart (l'équivalent de 2 à 4 pour cent) peut s'expliquer par l'utilisation d'un échantillonnage sans remise aux deux premiers degrés de l'échantillonnage. Le reste (l'équivalent de 4 pour cent environ) est faible par rapport à l'erreur type de l'écart estimé. On a estimé la variance de l'écart en estimant la variance de la moyenne de z_{ij}, où

$$z_{ij} = (\bar{y}_{n,r(ij)} - 79.76)^2 - \hat{V}(\bar{y}_{n,r(ij)}), \tag{4.10}$$

pour le j-ième échantillon de rendement ($j = 1, 2$) dans le t-ième échantillon de l'enquête énumérative de juin ($t = 1, \ldots, 500$). L'erreur type estimée de l'écart était de 0.58. Ainsi, la valeur moyenne de $\hat{V}(\bar{y}_r)$ se situe à moins de 1.5 erreurs types de la variance estimée de $\bar{y}_r$. Par ailleurs, la valeur moyenne de l'estimation de la variance de $\hat{Y}_r$ se situe à moins de 2% de la variance observée dans la simulation de Monte Carlo.

Les estimateurs de la variance de $\bar{y}$ par les groupes aléatoires avaient un biais peu significatif. Dans la simulation de Monte Carlo, les moyennes des estimateurs $\hat{V}_{g5}(\bar{y})$ et $\hat{V}_{g10}(\bar{y})$ étaient respectivement de 9 et de 7% supérieures aux variances observées. Ces écarts ne sont pas significativement différents de zéro et sont comparables à ceux obtenus pour l'estimateur $\hat{V}(\bar{y}_r)$. Celui-ci, toutefois, est un estimateur de variance beaucoup plus stable, le coefficient de variation pour cet estimateur étant d'environ 30%, comparativement à 75% pour $\hat{V}_{g5}(\bar{y})$. Comme prévu (Wolter 1985), un accroissement du nombre de groupes aléatoires a amené une diminution du coefficient de variation de l'estimateur de la variance par les groupes aléatoires. En effet, le coefficient de variation pour $\hat{V}_{g10}(\bar{y})$ était de 50%. Les variations entre groupes aléatoires et entre échantillons de rendement de l'enquête énumérative de juin ont expliqué en majeure partie la variance des estimateurs de variances par les groupes aléatoires.

4. CONCLUSIONS

Les analyses montrent que les estimateurs du rendement moyen et de la production totale d'un État, qui sont actuellement utilisés par le National Agricultural Statistical Service sont satisfaisants. En revanche, nous avons vu que les estimateurs de variance simples $\hat{V}_2(\bar{y})$ et $\hat{V}_*(\hat{Y})$ étaient entachés d'un biais négatif dont l'importance dépendait de la variance à l'intérieur des segments et des fractions de sondage à l'intérieur des segments. L'estimateur $\hat{V}_2(\bar{y})$ a produit une estimation qui était près de 40% inférieure à la vraie variance de $\bar{y}$, tandis que l'estimateur $\hat{V}_*(\hat{Y})$ en a produit une qui était de 18% inférieure à la vraie variance de $\hat{Y}$ pour la population simulée.

On a élaboré les estimateurs $\bar{y}_r$ et $\hat{Y}_r$ pour l'estimation de rendement d'un échantillonnage à deux phases, selon lequel les segments qui sont identifiés comme des segments de culture du coton dans la première phase (enquête énumérative de juin) sont sous-échantillonnés dans la seconde phase afin d'estimer le rendement. On estime la probabilité non conditionnelle qu'un segment soit prélevé en vue de recevoir une unité secondaire à l'intérieur d'une strate, π^*_{hk}, en supposant que cette probabilité est proportionnelle à la probabilité conditionnelle de sélection de segments dans la seconde phase de l'échantillonnage. Suivant cette hypothèse, on a défini l'estimateur par quotient combiné du rendement moyen avec probabilités inégales, $\bar{y}_r$, et l'estimateur de sa variance, $\hat{V}(\bar{y}_r)$. L'estimateur de l'agrégat $\hat{Y}_r$ est un estimateur par produit de la production moyenne d'un segment pour un sondage à deux phases, où l'estimateur de la moyenne de la variable auxiliaire (superficie consacrée à la culture étudiée) est tiré de l'enquête énumérative de juin (première phase du sondage). L'estimateur de variance $\hat{V}(\hat{Y}_r)$ est un estimateur de la variance de $\hat{Y}_r$ pour un échantillonnage double stratifié (sondage à deux phases).

La simulation de Monte Carlo a montré que $\bar{y}_r$ et $\hat{Y}_r$ produisaient des estimations comparables à celles découlant des estimateurs courants $\bar{y}$ et $\hat{Y}$. $\hat{V}(\bar{y}_r)$ et $\hat{V}(\hat{Y}_r)$ sont tous deux des estimateurs de variance précis pour des échantillons de la taille de ceux qu'utilise normalement le NASS. Ces résultats sont attribuables en partie à l'exactitude avec laquelle les superficies cultivées moyennes sont estimées au moyen de l'enquête énumérative de juin. Des estimations de superficie justes produisent des estimations de probabilités de sélection qui se rapprochent des probabilités de sélection non conditionnelles. En outre, le fait que l'estimateur soit sous forme de quotient amenuise les effets de la substitution d'estimateurs aux probabilités non conditionnelles réelles.

Par ailleurs, les estimateurs de variance pour les groupes aléatoires sont essentiellement des estimateurs non biaisés de la variance de l'estimation du rendement et de la production. Cependant, ces estimateurs sont beaucoup moins stables que $\hat{V}(\bar{y}_r)$ et $\hat{V}(\hat{Y}_r)$. C'est pourquoi ces derniers sont préférés aux estimateurs pour les groupes aléatoires.

L'enquête énumérative de juin constitue la première phase de l'enquête objective sur le rendement. Les méthodes d'échantillonnage utilisées dans cette enquête sont simples et, comme en fait foi la simulation de Monte Carlo, produisent des estimations de superficie justes. Par conséquent, il ne convient pas d'apporter des modifications à la première phase de l'enquête objective sur le rendement.

On peut toutefois envisager un certain nombre de modifications pour la seconde phase de cette enquête. À l'heure actuelle, on estime le rendement au moyen d'un sondage à deux phases, dans lequel est utilisé un estimateur par quotient combiné. Pour les États où l'échantillon est relativement grand, il conviendrait d'envisager un échantillonnage indépendant à la seconde phase pour chaque strate ou des groupes de strates, ainsi que l'utilisation d'un estimateur par quotient distinct.

Pour obtenir des estimateurs non biaisés de la variance, il conviendrait de remplacer l'échantillonnage systématique effectué dans la seconde phase. À l'heure actuelle, l'échantillonnage des segments à la seconde phase pour l'estimation du rendement se fait par ordinateur et vise tous les États américains. Il serait donc relativement facile d'adopter une méthode d'échantillonnage qui utiliserait des probabilités conjointes connues. Des estimateurs semblables à ceux qui sont recommandés pour le plan de sondage qui nous intéresse pourraient toujours convenir si les probabilités de sélection demeuraient les mêmes. Fuller (1970) décrit une méthode d'échantillonnage qui peut être informatisée et qui comprend le calcul de probabilités conjointes de sélection; en outre, cette méthode utilise des probabilités de sélection déterminées et assure un degré de contrôle comparable à celui qu'offre l'échantillonnage systématique.

REMERCIEMENTS

Cette étude a été réalisée en partie grâce à un contrat de recherche coopérative (n° 58-319T-1-0054X) passé avec le National Agricultural Statistics Service du Département de l'agriculture des États-Unis. Nous tenons à exprimer notre reconnaissance aux arbitres pour leurs commentaires utiles.

BIBLIOGRAPHIE

FECSO, R. (1978). Cluster analysis as an aid in creating paper strata. Statistical Reporting Service, U.S. Department of Agriculture.

FECSO, R., et JOHNSON, V. (1981). The new California area frame: A statistical study. Statistical Reporting Service, U.S. Department of Agriculture.

FULLER, W.A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, Sér. B, 32, 209-226.

HOUSEMAN, E.E. (1975). Area frame sampling in agriculture. Statistical Reporting Service, U.S. Department of Agriculture.

PRATT, W.L. (1984). The use of interpenetrating sampling in area frames. Statistical Reporting Service, U.S. Department of Agriculture.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 71-82
Statistique Canada

Problèmes courants sur la désaisonnalisation

ESTELA BEE DAGUM[1]

RÉSUMÉ

Dans cet article, l'auteure analyse trois questions qui, depuis une dizaine d'années, intéressent vivement les spécialistes de la désaisonnalisation qui oeuvrent dans les organismes statistiques. Ces trois questions sont: (1) le choix entre facteurs saisonniers contemporains et facteurs saisonniers prévus pour la désaisonnalisation courante; (2) la définition d'un modèle de révisions optimal pour les séries désaisonnalisées au moyen de facteurs saisonniers contemporains; et (3) le lissage de données désaisonnalisées très irrégulières.

MOTS CLÉS: Facteurs saisonniers contemporains et facteurs saisonniers prévus; révisions; filtres de la tendance-cycle; lissage.

1. INTRODUCTION

Dans les dix dernières années, trois questions importantes touchant la désaisonnalisation ont retenu l'attention des responsables des organismes statistiques: (1) la désaisonnalisation de valeurs courantes; (2) la révision de données désaisonnalisées contemporaines; et (3) le lissage de séries désaisonnalisées très irrégulières.

Cet article vise principalement à analyser chacune de ces questions en fonction de la méthode de désaisonnalisation X-11-ARMMI élaborée par Dagum (1980) et appliquée par Statistique Canada et d'autres organismes statistiques à l'étranger.

Dans la section 2, nous analysons les quatre modes d'application du programme X-11-ARMMI qui vise à produire des valeurs désaisonnalisées courantes. Dans la section 3, nous allons surtout examiner les révisions des données désaisonnalisées contemporaines, qui découlent de l'application des filtres linéaires de la méthode X-11-ARMMI. Enfin, dans la section 4, nous examinons les caractéristiques des filtres de lissage (tendance-cycle) de la X-11-ARMMI.

2. DÉSAISONNALISATION DE VALEURS COURANTES

Une valeur courante peut être désaisonnalisée à l'aide d'un facteur saisonnier contemporain ou d'un facteur saisonnier prévu.

On obtient l'estimation d'un facteur saisonnier contemporain (facteur ou effet qui varie selon qu'il s'agit respectivement, par hypothèse, d'un modèle multiplicatif ou additif) en désaisonnalisant, chaque fois qu'une nouvelle observation est enregistrée, toutes les données antérieures, y compris l'observation en question. Par ailleurs, on détermine un facteur saisonnier prévu au moyen d'une série qui se termine dans l'année précédente. D'une manière générale, on projettera par exemple des facteurs saisonniers pour l'année $T + 1$ au moyen d'une série de données qui se termine en décembre de l'année précédente T.

Il existe quatre modes d'application du programme X-11-ARMMI, qui permettent de produire une valeur désaisonnalisée courante (dernière observation). Ce sont: (i) les extrapolations ARMMI avec facteurs saisonniers contemporains; (ii) les extrapolations ARMMI avec facteurs saisonniers prévus; (iii) les facteurs saisonniers contemporains sans extrapolations ARMMI; et (iv) les facteurs saisonniers prévus sans extrapolations ARMMI.

[1] Estela Bee Dagum, Division des séries chronologiques recherche et analyse, Direction de la méthodologie, Statistique Canada, 13e étage, Immeuble R.H. Coats, Ottawa (Ontario), Canada, K1A 0T6.

Bien que les organismes statistiques se servent des quatre modes pour obtenir des valeurs désaisonnalisées courantes, la fréquence d'application de chacun de ces modes varie selon les organismes. Par exemple, Statistique Canada utilise surtout le mode (i), puis le mode (iii) tandis que le U.S. Bureau of Labor utilise surtout le mode (ii), puis le mode (iv). Les quatre modes ne produisent pas tous la même valeur désaisonnalisée courante et ne renferment pas non plus le même degré d'erreur.

Suivant l'hypothèse d'un modèle de décomposition additif, on désaisonnalisera une valeur courante X_t par l'équation

$$\hat{X}_t^{(\ell)} = X_t - \hat{S}_t^{(\ell)}, \tag{1}$$

où $\hat{S}_t^{(\ell)}$ désigne l'estimation d'un facteur saisonnier prévu; ou

$$\hat{X}_t^{(0)} = X_t - \hat{S}_t^{(0)}, \tag{2}$$

où $\hat{S}_t^{(0)}$ désigne l'estimation d'un facteur saisonnier contemporain.

La valeur courante désaisonnalisée deviendra "finale", c'est-à-dire qu'elle ne sera plus révisée, lorsque m observations auront été ajoutées à la série. Ainsi,

$$\hat{X}_t^{(m)} = X_t - \hat{S}_t^{(m)}, \tag{3}$$

où $\hat{S}_t^{(m)}$ désigne l'estimation finale d'un facteur saisonnier.

On peut donc exprimer la révision totale d'un facteur saisonnier contemporain et d'un facteur saisonnier prévu par les formules suivantes:

$$r_t^{(0,m)} = \hat{S}_t^{(0)} - \hat{S}_t^{(m)}, \quad m > 0; \tag{4}$$

$$r_t^{(\ell,m)} = \hat{S}_t^{(\ell)} - \hat{S}_t^{(m)}, \quad m > 0 > \ell. \tag{5}$$

Suivant l'hypothèse d'un modèle de décomposition additif et du non-remplacement des valeurs extrêmes, $\hat{S}_t^{(m)}$, la valeur finale d'un facteur saisonnier tirée d'une série $X_{t-m}, \ldots, X_t, \ldots, X_{t+m}$, peut être défini comme étant

$$\hat{S}_t^{(m)} = \sum_{j=-m}^{m} h_{m,j} X_{t-j} = h^{(m)}(B) X_t, \tag{6}$$

où $h_{m,j} = h_{m,-j}$ sont les poids des moyennes mobiles symétriques qui doivent être appliquées à la série. Le terme $h^{(m)}(B)$ désigne le filtre linéaire correspondant qui utilise l'opérateur de retard B de telle sorte que $B^n = X_{t-n}$. Young (1968) a montré que l'étendue de ce filtre symétrique est de 145 poids pour les séries mensuelles mais que l'on peut en obtenir une bonne approximation au moyen de 85 poids parce que la valeur des poids attribués aux observations éloignées est très faible; par conséquent, $m = 42$.

Étant donné l'équation (6), nous pouvons exprimer l'estimation d'un facteur saisonnier contemporain $\hat{S}_t^{(0)}$ et l'estimation d'un facteur saisonnier prévu $\hat{S}_t^{(\ell)}$ par

$$\hat{S}_t^{(0)} = \sum_{j=-2m}^{0} h_{0,j} X_{t-j} = h^{(0)}(B) X_t, \quad m = 42, \tag{7}$$

où $h^{(0)}(B)$ désigne le filtre asymétrique saisonnier *contemporain*; et

$$\hat{S}_t^{(\ell)} = \sum_{j=-2m}^{\ell} h_{\ell,j} X_{t-j} = h^{(\ell)}(B) X_t, \; m = 42, \tag{8}$$

où $h^{(\ell)}(B)$ désigne le filtre asymétrique saisonnier *prévu* et $\ell = 1, 2, \ldots, 12$ pour une série mensuelle.

La révision d'une estimation d'un facteur saisonnier contemporain dépend de la distance entre le filtre contemporain et le filtre final, c'est-à-dire $d[h^{(0)}(B), h^{(m)}(B)]$, et des innovations des nouvelles observations $X_{t+1}, X_{t+2}, \ldots, X_{t+m}$.

De même, la révision d'une estimation d'un facteur saisonnier prévu dépend de $d[h^{(\ell)}(B), h^{(m)}(B)]$ et des innovations introduites par $X_{t-\ell}, \ldots, X_t, X_{t+1}, \ldots, X_{t+m}$.

Dagum (1982a et 1982b) a montré dans des études théoriques que

$$d[h^{(0)}(B), h^{(m)}(B)] < d[h^{(\ell)}(B), h^{(m)}(B)] \text{ for } \ell = 1, 2, \ldots 12. \tag{9}$$

La distance entre les deux filtres est définie comme l'écart quadratique moyen entre les fonctions de réponse en fréquence complexes des filtres pour l'ensemble des fréquences saisonnières. Une définition semblable est donnée dans la section suivante (équation 17), avec la racine de l'écart quadratique moyen.

L'inéquation (9) est vraie peu importe que l'on utilise des extrapolations ARMMI ou non. En outre, les deux études précitées ont montré que

$$\begin{aligned} & d[h^{(0)}(B), h^{(m)}(B)] \quad \text{avec extrapolations ARMMI} \\ & < d[h^{(0)}(B), h^{(m)}(B)] \quad \text{sans extrapolations ARMMI,} \end{aligned} \tag{10}$$

de même,

$$\begin{aligned} & d[h^{(\ell)}(B), h^{(m)}(B)] \quad \text{avec extrapolations ARMMI} \\ & < d[h^{(\ell)}(B), h^{(m)}(B)] \quad \text{sans extrapolations ARMMI,} \end{aligned} \tag{11}$$

$$\text{pour } \ell = 1, 2, \ldots, 12.$$

Plusieurs études (Dagum (1978), Bayer et Wilcox (1981), Kenney et Durbin (1982), McKenzie (1984), Dagum et Morry (1984), Pierce (1980) et Pierce et McKenzie (1985)) ont montré que

$$r^{(0,m)} < r^{(\ell,m)}. \tag{12}$$

Mais dans quelques cas,

$$r^{(0,m)} > r^{(\ell,m)}. \tag{13}$$

L'inéquation (13) se vérifie lorsque les observations courantes de la dernière année sont fortement modifiées parce que X_t reçoit le poids le plus élevé dans les estimations de $\hat{S}_t^{(0)}$.

Du point de vue de la révision totale des estimations de facteurs saisonniers, nous pouvons classer les quatre modes de désaisonnalisation en nous fondant sur les résultats des études empiriques mentionnées ci-dessus: le mode (i) (facteurs saisonniers contemporains avec extrapolations ARMMI) produit la révision totale la plus faible; le mode (iii) (facteurs saisonniers contemporains sans extrapolations ARMMI) vient au second rang; le mode (ii) (facteurs saisonniers prévus avec extrapolations ARMMI) vient au troisième rang, suivi du mode (iv) (facteurs saisonniers prévus sans extrapolations ARMMI).

3. RÉVISION DES DONNÉES DÉSAISONNALISÉES CONTEMPORAINES

Statistique Canada a utilisé la désaisonnalisation contemporaine pour la première fois en 1975 dans le cadre de l'Enquête sur la population active. Progressivement, les organismes statistiques d'autres pays ont suivi l'exemple de Statistique Canada et ont adopté cette méthode. L'utilisation de facteurs saisonniers contemporains dans la désaisonnalisation courante nous amène à nous demander à quel rythme faut-il réviser les séries. Kenny et Durbin (1982) ont proposé d'effectuer les révisions après un mois et, par la suite, à chaque année civile. Dagum (1982c) a appuyé ces conclusions et a proposé en plus une révision additionnelle après six mois si la méthode de désaisonnalisation utilisée est la X-11-ARMMI sans extrapolations ARMMI.

Pour deux points quelconques au temps $t + k$, $t + \ell$ $(k < \ell)$, les révisions des estimations de facteurs saisonniers et, partant, des valeurs désaisonnalisées sont définies par

$$r_t^{(\ell,k)} = \hat{X}_t^{(\ell)} - \hat{X}_t^{(k)}, \quad k < \ell. \tag{14}$$

Ces révisions rendent compte: (1) des innovations introduites par les nouvelles observations $X_{t+k+1}, X_{t+k+2}, \ldots, X_{t+k+\ell}$; et (2) des différences entre les deux filtres de désaisonnalisation asymétriques $Y^{(\ell)}(B)$ et $Y^{(k)}(B)$. Si on pose $k = 0$ et qu'on fait varier ℓ de 1 à m, l'équation (14) produit une série de révisions apportées aux valeurs désaisonnalisées contemporaines pour des périodes ou des décalages différents. La *révision totale* de l'estimation du facteur saisonnier contemporain est donnée pour $\ell = m$. Si on pose $\ell = k + 1$ et qu'on fait varier k de 0 à $m - 1$, on obtient par l'équation (14) la série de *révisions apportées par période* à chaque valeur désaisonnalisée estimée et si l'on débute notamment à $k = 0$, on obtient les $m - 1$ révisions consécutives apportées par période à chaque valeur désaisonnalisée estimée avant qu'elle ne devienne finale. Si on pose $\ell = k + 12$ et qu'on fait varier k de 0 à $m - 12$, on obtient par l'équation (14) la série de révisions annuelles.

Les révisions qui nous intéressent en l'occurrence sont celles causées par les différences entre les filtres, et il est possible d'analyser ces différences en examinant les fonctions de réponses en fréquence complexes des filtres correspondants. Comme dans l'équation (6), nous pouvons estimer les valeurs désaisonnalisées pour les années récentes au moyen du programme X-11-ARMMI (avec ou sans extrapolations ARMMI) par

$$\hat{X}_t^{(n)} = \sum_{j=n}^{m} Y_{n,j} X_{t-j} = Y^{(n)}(B)\, X_t. \tag{15}$$

L'équation (15) représente un système linéaire où $\hat{X}_t^{(n)}(n)$ est le produit de convolution de la série initiale X_t et d'une série de poids $Y_{n,j}$ appelée la *fonction de réponse à une impulsion* du filtre. On peut étudier les propriétés de cette réponse à l'aide de sa transformée de Fourier que l'on appelle la *fonction de réponse en fréquence complexe* et qui est définie par

$$\Gamma^{(n)}(\omega) = \sum_{j=-n}^{m} Y_{n,j} e^{-2\pi\omega j}, \quad -\tfrac{1}{2} \leq \omega \leq \tfrac{1}{2}, \tag{16}$$

où ω est la fréquence exprimée en cycles par unité de temps. $\Gamma^{(n)}(\omega)$ décrit entièrement les effets du filtre linéaire sur la série initiale donnée. Dagum (1987) a calculé les révisions mensuelles et annuelles apportées par les filtres contemporains du X-11-ARMMI (avec ou sans extrapolations ARMMI), en se fondant sur la distance mathématique entre les diverses fonctions de réponse en fréquence complexes des filtres. Ces calculs nous permettent de constater une diminution rapide de l'importance des révisions mensuelles pour $\ell = 1, 2$, et 3, puis une diminution lente jusqu'à $\ell = 11$. On observe ensuite une forte augmentation à $\ell = 12$, puis une diminution à $\ell = 13$ et de nouveau une forte augmentation à $\ell = 24$, celle-ci suivie d'une diminution à $\ell = 25$. Dagum (1987) a montré que ces révisions mensuelles suivent la même tendance, qu'on utilise ou non les extrapolations ARMMI.

La forte diminution observée dans les trois premières révisions consécutives s'explique par l'amélioration des poids des filtres de Henderson (tendance-cycle). Le changement de tendance observé consécutivement à $\ell = 12$ et à $\ell = 13$ est dû aux améliorations apportées au filtre saisonnier qui perd de son asymétrie d'une année à l'autre jusqu'à ce que trois années complètes s'ajoutent à la série. On observe les corrections les plus prononcées à $\ell = 1$ et à $\ell = 12$. *Étant donné la non-monotonie des révisions mensuelles, il n'est pas recommandé de réviser les estimations de facteurs saisonniers contemporains à chaque fois qu'une nouvelle observation s'ajoute à la série.*

Pour la révision des séries désaisonnalisées contemporaines, les organismes statistiques procèdent souvent de la façon suivante: ils maintiennent fixe l'estimation du facteur saisonnier contemporain depuis le moment où elle est établie jusqu'à la fin de l'année et révisent ensuite annuellement l'année courante et les années antérieures. Ainsi, les révisions de première année attribuables aux différences entre les filtres sont données par $R^{(0,0)}, R^{(1,0)}, \ldots, R^{(11,0)}$, les révisions de seconde année par $R^{(12,0)}, R^{(13,1)}, \ldots, R^{(23,11)}$, les révisions de troisième année par $R^{(24,12)}, R^{(25,13)}$ et ainsi de suite, où $R^{(\ell,k)}$ est définie par

$$R^{(\ell,k)} = [2\int_0^{1/2} \|\Gamma^{(\ell)}(\omega) - \Gamma^{(k)}(\omega)\|^2 \, d\omega]^{1/2}, \tag{17}$$

$$\ell = 1, 2, \ldots, n, k = 0, 1, 2, \ldots, n - 12,$$

et $n = 42$ pour les filtres du X-11-ARMMI.

Le tableau 1 montre les révisions de première, de seconde et de troisième année apportées par le filtre contemporain du X-11-ARMMI (avec et sans extrapolations ARMMI), les extrapolations étant tirées d'un modèle ARMMI donné et de deux séries de paramètres (d'autres cas sont exposés dans Dagum (1987)). Le modèle ARMMI choisi est le modèle $(0,1,1)\ (0,1,1)_{12}$, c'est-à-dire $(1 - B)\ (1 - B^{12})\ X_t = (1 - \theta B)\ (1 - \Theta B^{12})a_t$ où X_t désigne la série initiale, B est l'opérateur de retard de sorte que $B^n X_t = X_{t-n}$, a_t est un élément purement aléatoire qui représente les résidus, et θ et Θ désignent respectivement les paramètres non saisonnier et saisonnier.

Comme les plus fortes révisions par période se produisent à $\ell = 1$ et $\ell = 12$, on obtiendrait un meilleur modèle de révision en intégrant des révisions mensuelles et annuelles. Ainsi, (1) en soumettant les données de chaque mois (de janvier à novembre par exemple) à une désaisonnalisation contemporaine puis en révisant ces données une seule fois, au moment où les données du mois suivant sont connues, et (2) en appliquant une désaisonnalisation contemporaine aux données de décembre dès qu'elles sont connues, puis en révisant les données de la première année et des années antérieures lorsque les données de janvier sont connues, nous devrions accroître la fiabilité du filtre appliqué dans l'année courante tout en conservant son homogénéité pour les comparaisons d'un mois à l'autre.

Les révisions de première année apportées par le filtre ayant fait l'objet d'une première révision mensuelle seraient donc $R^{(1,1)}, R^{(2,1)}, \ldots, R^{(11,1)}$. La valeur de ces révisions figure dans le tableau 2 et, malgré une tendance très comparable à celle observée dans le tableau 1 (filtre appliqué aux facteurs saisonniers contemporains), *les révisions du tableau 2 sont beaucoup moins marquées s'il n'y a pas d'extrapolation*. Par ailleurs, *l'amélioration est moins notable lorsqu'on utilise des extrapolations ARMMI*. On n'a constaté aucune différence majeure dans le cas des révisions de seconde et de troisième année.

3.1 Estimation des variations liées aux jours ouvrables et modèles ARMMI dans le contexte de la désaisonnalisation contemporaine

Outre le choix du modèle de révision à appliquer, la désaisonnalisation contemporaine pose deux autres problèmes qui ont trait aux modèles ARMMI et aux variations liées aux jours ouvrables.

Tableau 1
Révisions de première, seconde et troisième année apportées par le filtre contemporain du X-11-ARMMI

Révisions $R^{(\ell,k)}$	Sans extrapolations ARMMI	Avec extrapolations ARMMI tirées d'un modèle $(0,1,1)\ (0,1,1)_{12}$			
		$\theta = .40$	$\Theta = .80$	$\theta = .80$	$\Theta = .80$
$R^{(1,0)}$	.12	.12		.06	
$R^{(2,0)}$	.13	.13		.08	
$R^{(3,0)}$	.13	.13		.08	
$R^{(4,0)}$	.13	.13		.09	
$R^{(5,0)}$	.15	.13		.09	
$R^{(6,0)}$	.17	.13		.09	
$R^{(7,0)}$	.16	.13		.09	
$R^{(8,0)}$	.16	.13		.09	
$R^{(9,0)}$	.16	.13		.09	
$R^{(10,0)}$	.16	.14		.09	
$R^{(11,0)}$	.16	.14		.09	
$R^{(12,0)}$	.29	.28		.26	
$R^{(13,1)}$	.27	.27		.26	
$R^{(14,2)}$	.27	.27		.26	
.	.	.		.	
.	.	.		.	
.	.	.		.	
$R^{(23,11)}$	.27	.26		.26	
$R^{(24,12)}$	.20	.16		.16	
$R^{(24,13)}$	.18	.17		.16	
$R^{(36,24)}$	.16	.17		.16	
.	.	.		.	
.	.	.		.	
.	.	.		.	

Tableau 2
Révisions de première année apportées par le filtre de désaisonnalisation ayant fait l'objet d'une première révision mensuelle

Révisions $R^{(\ell,1)}$	Sans extrapolations ARMMI	Avec extrapolations ARMMI tirées d'un modèle $(0,1,1)\ (0,1,1)_{12}$			
		$\theta = .40$	$\Theta = .80$	$\theta = .80$	$\Theta = .80$
$R^{(2,1)}$	.07	.10		.06	
$R^{(3,1)}$	·07	.10		.06	
$R^{(4,1)}$	.07	.10		.07	
$R^{(5,1)}$	.08	.10		.08	
$R^{(6,1)}$	.10	.11		.08	
$R^{(7,1)}$	.11	.11		.08	
$R^{(8,1)}$	.11	.11		.08	
$R^{(9,1)}$	.11	.11		.08	
$R^{(10,1)}$	.12	.11		.08	
$R^{(11,1)}$	.12	.12		.08	

Les séries de flux, c'est-à-dire qui découlent de l'accumulation de données quotidiennes dans les mois civils, sont soumises à un effet systématique attribuable aux variations liées aux jours ouvrables. Ce genre de variations sont imputables principalement au fait que l'activité commerciale fluctue selon les jours de la semaine. Les pratiques comptables et les méthodes de présentation des états financiers peuvent aussi expliquer ce genre de variations. Par exemple, les magasins où la comptabilisation des opérations se fait le vendredi ont tendance à déclarer un chiffre de ventes plus élevé dans les mois qui comptent cinq vendredis que dans ceux qui en comptent quatre. Dans le programme X-11-ARMMI, on estime l'effet des variations liées aux jours ouvrables en appliquant la méthode des moindres carrés ordinaires à un modèle de régression simple déterministe; par conséquent, les poids estimés pour chaque jour varient chaque fois qu'une nouvelle observation s'ajoute à la série. Comme les techniques de régression sont très sensibles aux valeurs aberrantes, ces variations de poids sont parfois inutilement élevées.

Lorsque les séries sont soumises à une désaisonnalisation contemporaine, les estimations des variations liées aux jours ouvrables varient constamment. Pour éviter des révisions inutiles, Statistique Canada utilise habituellement les poids calculés par le programme à la fin de l'année civile précédente ou ceux fournis par les utilisateurs comme facteurs de pondération préalable pour l'année courante. Par la suite, ces poids sont révisés annuellement.

L'effet des variations liées aux jours ouvrables doit être supprimé d'une série avant l'application d'un modèle ARMMI, car ce genre de modèle n'est pas conçu pour tenir compte des variations liées aux jours ouvrables. En d'autres mots, si l'on applique le programme X-11-ARMMI avec extrapolations ARMMI à une série qui est soumise à des variations liées aux jours ouvrables, on estime *a priori* ces variations et, si elles sont appréciables, on les extrait de la série initiale avant l'application du modèle ARMMI.

Un autre problème que soulève la désaisonnalisation contemporaine a trait au nombre de fois que des modèles ARMMI devraient être définis. En règle générale, Statistique Canada utilise l'option automatisée ARMMI une fois par année; si le modèle est accepté, il est maintenu fixe toute une année. Seuls les paramètres varient lorsque des observations s'ajoutent. Pour maintenir fixe le modèle, il convient d'appliquer l'option du modèle fourni par l'utilisateur. En maintenant le modèle ARMMI fixe, on évite les révisions inutiles qui peuvent découler de changements de modèle attribuables uniquement à l'existence de valeurs aberrantes.

4. LISSAGE DE SÉRIES DÉSAISONNALISÉES VOLATILES

Un des principaux objectifs de la désaisonnalisation des séries chronologiques économiques est de produire de l'information sur la conjoncture économique actuelle et, plus particulièrement, de déterminer la phase du cycle dans laquelle se trouve l'économie. Comme la désaisonnalisation consiste à éliminer les variations saisonnières, laissant par le fait même les variations de la tendance-cycle et les variations irrégulières, il est souvent difficile de déterminer la tendance à court terme ou les points de retournement cycliques des séries qui sont très irrégulières. Dans ce cas, il peut être préférable de lisser les séries désaisonnalisées au moyen d'estimateurs de la tendance-cycle, qui éliminent le plus possible la composante irrégulière sans toucher à la composante cyclique.

L'utilisation de valeurs de la tendance-cycle a été analysée par plusieurs auteurs, dont récemment Moore et coll. (1981), Kenny et Durbin (1982), Maravall (1986) et Dagum et Laniel (1987). Bien que cette pratique soit encore peu répandue, certains organismes statistiques tels que Statistique Canada et l'Australian Bureau of Statistics lissent certaines de leurs séries désaisonnalisées, notamment celles qui sont fortement touchées par les irrégulières.

Young (1968) a calculé, pour la variante X-11 de la Census Method II, les filtres linéaires combinés qui sont appliqués à la série initiale pour produire une estimation centrale (symétrique) de la tendance-cycle. Ces filtres sont semblables à celui du X-11-ARMMI avec ou sans extrapolations ARMMI. Dagum et Laniel (1987) ont poussé plus loin les travaux de Young (1968) en calculant l'estimation des filtres asymétriques de la tendance-cycle du X-11-ARMMI avec ou sans extrapolations ARMMI.

La figure 1 montre la fonction de gain des filtres de désaisonnalisation centraux (symétriques) et celle des filtres des données désaisonnalisées lissées (tendance-cycle). Il est clair que les filtres de la tendance-cycle suppriment tout bruit présent dans la série, le bruit étant défini comme la puissance présente dans toutes les fréquences $\omega \leq .166$. Cette dernière fréquence correspond à la première harmonique de la fréquence saisonnière fondamentale d'une série mensuelle. Cette structure découle de la convolution des filtres de désaisonnalisation avec le filtre de Henderson de 13 termes (tendance-cycle).

La figure 2a montre la fonction de gain des filtres appliqués aux facteurs saisonniers contemporains et des filtres de la tendance-cycle ayant fait l'objet d'une première révision mensuelle selon le X-11-ARMMI *sans* extrapolations ARMMI. La figure 2b montre les déphasages correspondants exprimés en mois plutôt qu'en radians. Nous constatons que le gain pour tout $\omega \leq .166$ est beaucoup plus grand dans le cas de ces deux filtres asymétriques que dans le cas du filtre central. Nous observons, en outre, de fortes amplifications pour les fréquences qui sont près de la fréquence saisonnière fondamentale. Toutes ces observations signifient que les valeurs désaisonnalisées contemporaines et les valeurs désaisonnalisées lissées ayant fait l'objet d'une première révision renfermeront plus de bruit que les estimations finales. Par ailleurs, on note que les déphasages sont très peu prononcés, soit moins d'un mois pour les fréquences cycliques les plus importantes, $0 < \omega < .055$ (c.-à-d., cycles d'une périodicité égale ou supérieure à 18 mois).

Les figures 3a et 3b montrent le gain et le déphasage des filtres appliqués aux facteurs saisonniers contemporains et des filtres de la tendance-cycle ayant fait l'objet d'une première révision mensuelle selon la méthode X-11-ARMMI avec extrapolations ARMMI. Les extrapolations sont tirées d'un modèle MMI $(0,1,1)(0,1,1)_{12}$ avec $\theta = .40$ et $\Theta = .60$. Les fonctions de gain de la figure 3a s'apparentent plus à la fonction de gain du filtre symétrique (central) que celles observées pour la méthode X-11-ARMMI sans les extrapolations ARMMI. Il n'y a aucune amplification autour de la fréquence saisonnière fondamentale mais on constate, comme dans le premier cas, une diminution de puissance aux fréquences élevées. En revanche, le déphasage est plus prononcé (près d'un mois) aux fréquences basses et moins prononcé à toutes les fréquences élevées.

Dagum et Laniel (1987) ont étudié l'évolution chronologique des révisions apportées par les filtres de la tendance-cycle et l'ont comparée à celle des révisions apportées par les filtres de désaisonnalisation. Les résultats de leur analyse, qui sont résumés dans le tableau 3, montrent que les révisions totales apportées par les filtres asymétriques de la tendance-cycle tendent vers zero beaucoup plus rapidement que celles apportées par les filtres de désaisonnalisation correspondants. De fait, la révision totale apportée par le filtre de la tendance-cycle trois mois après l'application du filtre contemporain n'est que de 0.1, alors que le filtre de désaisonnalisation ne produit pas de valeur comparable avant que 24 mois n'aient été ajoutés à la série. Sauf en ce qui concerne les premières révisions totales (filtre appliqué aux facteurs saisonniers courants), les révisions apportées par les filtres de la tendance-cycle sont inférieures à celles apportées par les filtres de désaisonnalisation correspondants. En outre, les premières tendent vers zero beaucoup plus rapidement que les secondes.

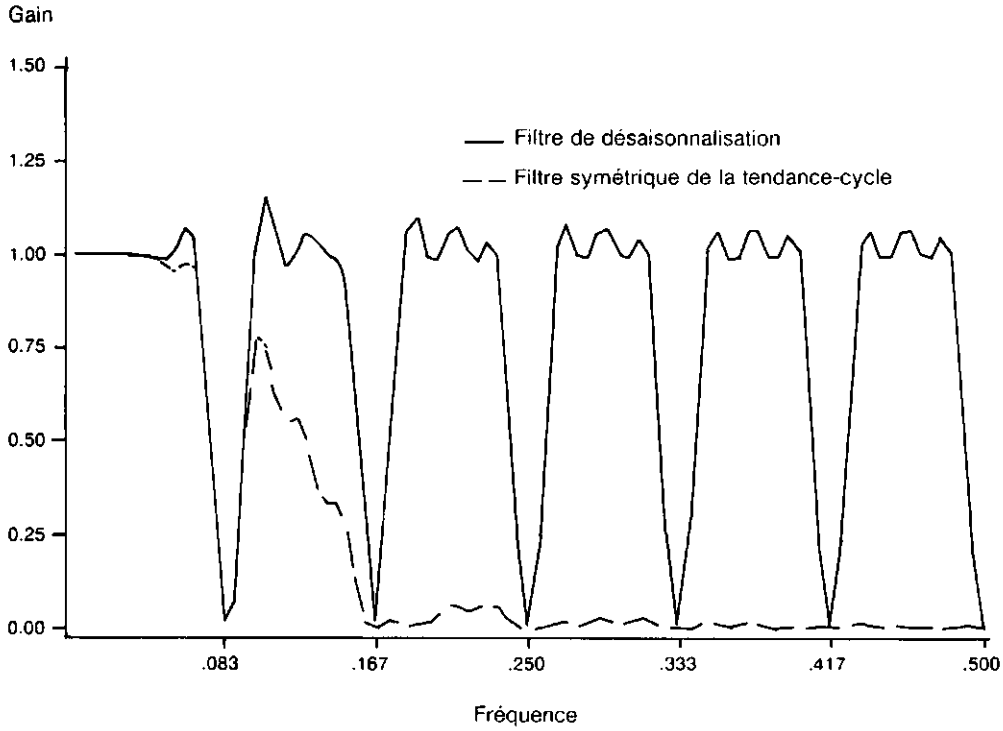

Figure 1. Fonctions de gain du filtre symétrique de la tendance-cycle et du filtre du désaisonnalisation du X-11-ARMMI.

Tableau 3
Évolution chronologique des révisions totales apportées par les filtres de la tendance-cycle et les filtres asymétriques de désaisonnalisation selon la méthode X-11-ARMMI

Révisions $R^{(\ell,k)*}$	Sans extrapolations		Avec extrapolations tirées d'un modèle $(0,1,1)\ (0,1,1)_{12}$ $\theta = .40$ $\Theta = .60$	
	Filtres de la tendance-cycle	Filtres de désaisonnalisation	Filtres de la tendance-cycle	Filtres de désaisonnalisation
$R^{(48,0)}$	.45	.36	.41	.32
$R^{(48,1)}$	.27	.33	.26	.32
$R^{(48,2)}$	.15	.32	.15	.32
$R^{(48,3)}$	.11	.32	.11	.31
$R^{(48,4)}$	.12	.32	.11	.31
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$R^{(48,12)}$	.10	.23	.09	.20
$R^{(48,24)}$	.07	.13	.05	.10
$R^{(48,36)}$	.03	.05	.02	.04
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$R^{(48,47)}$	.01	.01	.01	.01

* $\ell = 48$ pour le filtre de la tendance-cycle "final" et $\ell = 42$ pour le filtre de désaisonnalisation final. Toutefois, les valeurs des révisions apportées par les filtres de désaisonnalisation sont calculées aussi pour $\ell = 48$ puisqu'après $\ell = 42$, ces valeurs sont finales et, de ce fait, ne varient pas.

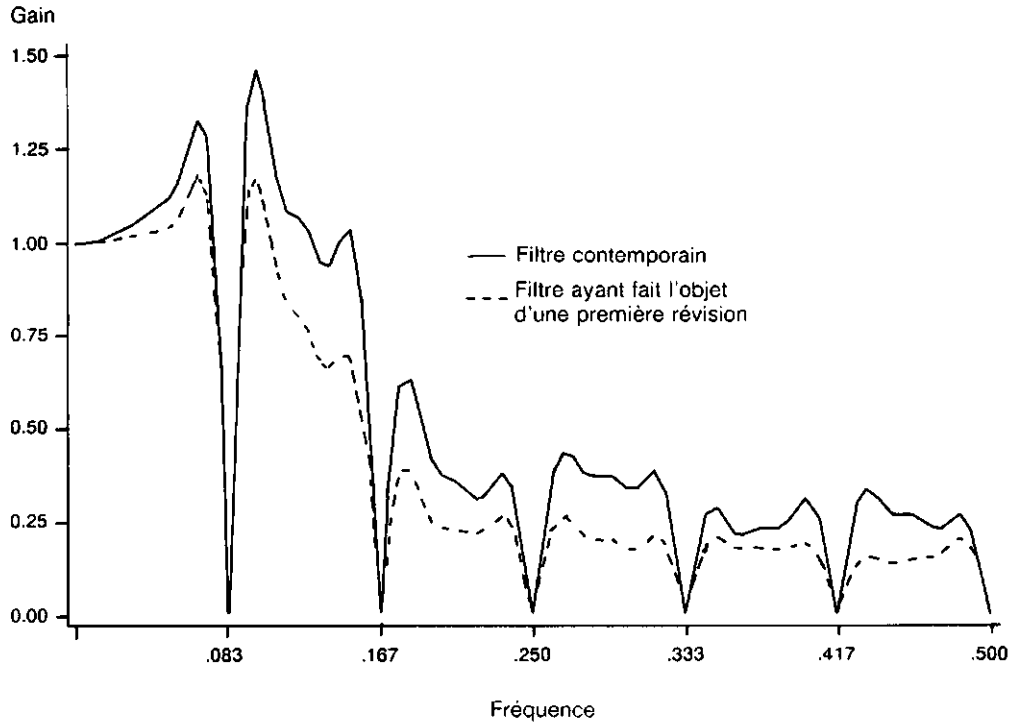

Figure 2a. Fonctions de gain du filtre contemporain et du filtre ayant fait l'objet d'une première révision du X-11-ARMMI sans extrapolations ARMMI.

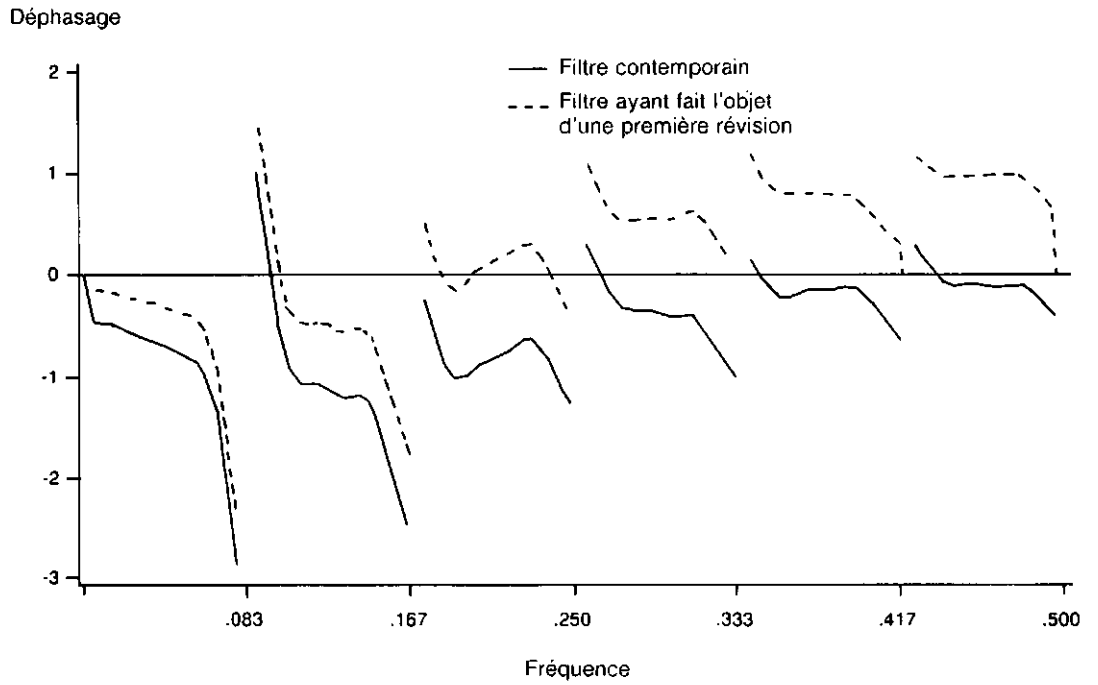

Figure 2b. Fonctions de déphasage du filtre contemporain et du filtre ayant fait l'objet d'une première révision du X-11-ARMMI sans extrapolations ARMMI.

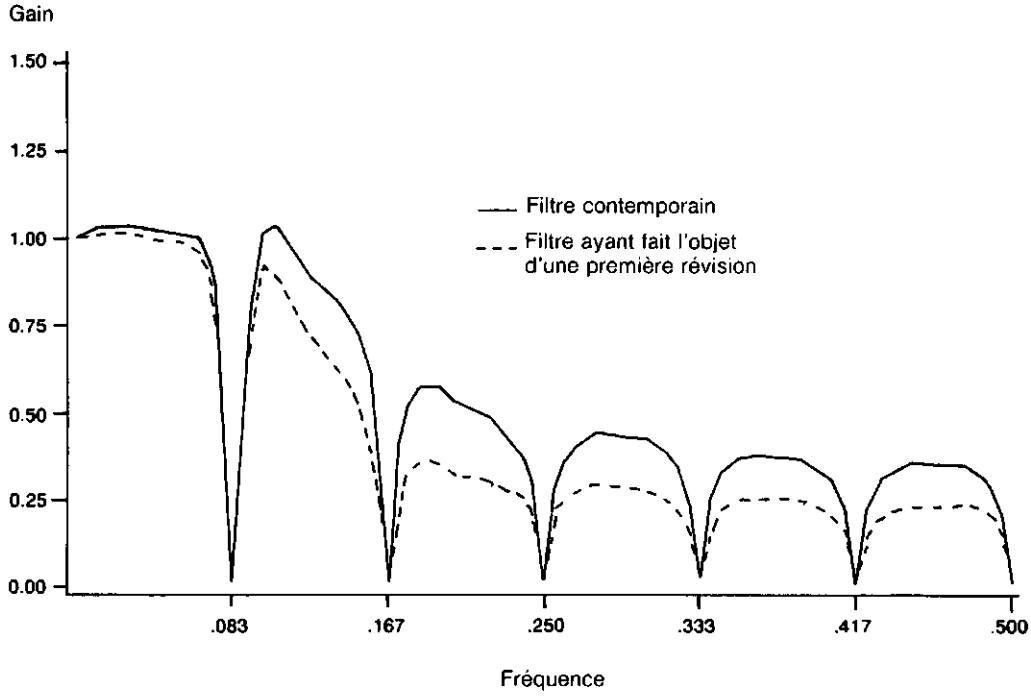

Figure 3a. Fonctions de gain du filtre contemporain et du filtre ayant fait l'objet d'une première révision du X-11-ARMMI avec extrapolations ARMMI (θ = .40, Θ = .60).

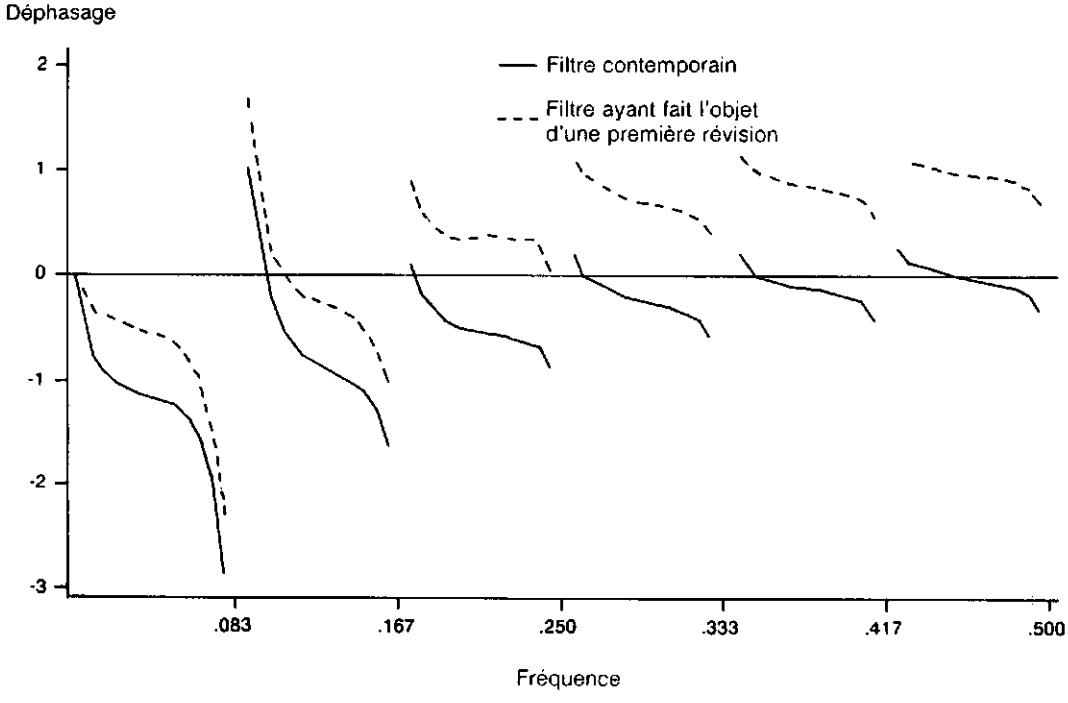

Figure 3b. Fonctions de déphasage du filtre contemporain et du filtre ayant fait l'objet d'une première révision du X-11-ARMMI avec extrapolations ARMMI (θ = .40, Θ = .60).

BIBLIOGRAPHIE

BAYER, A., et WILCOX, D. (1981). An evaluation of concurrent seasonal adjustment. Document technique, Board of Governors of the Federal Reserve System.

DAGUM, E.B. (1978). *Comparison and Assessment of Seasonal Adjustment Methods for Labour Force Series.* Stock No. 052-003-00603-1, U.S. Government Printing Office.

DAGUM, E.B. (1980). *La méthode de désaisonnalisation X-11-ARMMI.* N° 12-564F au répertoire, Statistique Canada.

DAGUM, E.B. (1982a). Revisions of time varying seasonal filters. *Journal of Forecasting*, 1, 173-187.

DAGUM, E.B. (1982b). The effects of asymmetric filters on seasonal factor revisions. *Journal of the American Statistical Association*, 77, 732-738.

DAGUM, E.B. (1982c). Revisions of seasonally adjusted data due to filter changes. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 39-45.

DAGUM, E.B., et MORRY, M. (1984). Basic issues on the seasonal adjustment of the Canadian Consumer Price Index. *Journal of Business and Economic Statistics*, 2, 250-259.

DAGUM, E.B. (1987). Monthly versus annual revisions of concurrent seasonally adjusted series. Dans *Time Series and Econometric Modelling*, (éds. I.B. MacNeill et G.J. Umphrey), New York: D. Reidel, 131-196.

DAGUM E.B., et LANIEL, N. (1987). Revisions of trend-cycle estimators of moving average seasonal adjustment method. *Journal of Business and Economic Statistics*, (en voie de rédaction).

KENNY, P., et DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic time series. *Journal of the Royal Statistical Society,* Sér. A, 145, 1-41.

MARAVALL, A. (1986). An application of model-based estimation of unobserved components. *International Journal of Forecasting*, 2, 305-318.

MOORE, G.H., BOX, G.E.P., KAITZ, H.B., STEPHENSON, J.A., et ZELLNER, A. (1981). Seasonal adjustment of the monetary aggregates. Dans *Report of the Committee of Experts on Seasonal Adjustment Techniques*, Washington: Board of Governors of the Federal Reserve System.

McKENZIE, S. (1984). Concurrent seasonal adjustment with Census X-11. *Journal of Business and Economic Statistics*, 2, 235-249.

PIERCE, D.A. (1980). Data revisions with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.

PIERCE, D., et McKENZIE, S. (1985). On concurrent seasonal adjustment. Special Studies Paper 164, Federal Reserve Board.

Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 83–91
Statistique Canada

Sur l'estimation efficace des taux de chômage à l'aide de données de l'enquête sur la population active

S. KUMAR et A.C. SINGH[1]

RÉSUMÉ

Tout comme avec la méthode d'estimation de X^2 min de Neyman (1949) pour les échantillons aléatoires simples, la méthode de $Q^{(T)}$ minimum proposée par Singh (1985) pour les plans de sondage complexes produit des estimateurs de paramètres de modèle asymptotiquement efficaces. La variable $Q^{(T)}$ peut être considérée comme une variable X^2 pour des données d'enquête qualitatives. Les estimateurs $Q^{(T)}$ min peuvent remplacer avantageusement les estimateurs obtenus par les moindres carrés pondérés qui ont souvent un comportement instable dans le cas des échantillons complexes. Les auteurs exposent tout d'abord la méthode de $Q^{(T)}$ min puis l'appliquent à l'estimation des paramètres d'un modèle logit pour des estimations de taux de chômage calculées à partir des données de l'EPA d'octobre 1980 classées selon deux critères: l'âge et le niveau d'instruction. Il appert que l'efficacité de trace des estimations lissées calculées par Kumar et Rao (1986) à l'aide de la méthode des pseudo-EMV peut être légèrement améliorée à l'aide de la méthode de $Q^{(T)}$ min. Détail digne de mention, les pseudo-EMV pour les cellules prises individuellement ont un comportement très comparable à celui des estimateurs efficaces $Q^{(T)}$ min utilisés pour l'EPA.

MOTS CLÉS: Pseudo-EMV; estimateur par les moindres carrés pondérés (MCP); estimateur $Q^{(T)}$ min; efficacité asymptotique; fonction de vraisemblance approximative; cote généralisée.

1. INTRODUCTION

En se basant sur les données de l'enquête sur la population active (EPA) d'octobre 1980, Kumar et Rao (1984, 1986) ont proposé et analysé un modèle de régression logistique (modèle logit) pour les taux de chômage. Ils ont utilisé la théorie élaborée par Roberts (1985) et Roberts, Rao et Kumar (1987). Ces derniers ont généralisé la méthode proposée par Rao et Scott (1981, 1984) pour corriger le X^2 en fonction des effets du plan de sondage et ont vérifié la validité de l'ajustement du modèle logit. Kumar et Rao ont considéré des taux de chômage pour diverses cellules (ou domaines) obtenues en classant simultanément la population visée en fonction de deux critères: l'âge et le niveau d'instruction. Le modèle logit renfermait des effets linéaires et quadratiques pour l'âge tandis qu'il ne renfermait que des effets linéaires pour le niveau d'instruction. Singh et Kumar (1986) ont analysé les mêmes données d'enquête au moyen d'une autre méthode, le test $Q^{(T)}$, qui a été proposée par Singh (1985). Le test $Q^{(T)}$ est un test du chi-carré fondé sur une cote généralisée de composantes principales. Ce test a produit des résultats comparables à ceux obtenus par la méthode du X^2 corrigé.

Dès qu'on a construit un modèle convenable, il faut trouver de bons estimateurs de ces paramètres. Ces estimateurs permettent ensuite d'estimer assez fidèlement les taux réels des domaines. Les estimations ainsi obtenues, souvent appelées "estimations lissées", sont particulièrement utiles dans les cas où les estimations d'enquête ne sont pas assez précises à cause d'un nombre insuffisant d'observations. Comme les estimations lissées sont calculées après qu'on a vérifié la validité de l'ajustement du modèle, il faut s'attendre que le biais des estimations soit négligeable. Kumar et Rao (1986) ont utilisé la méthode des pseudo-EMV (pseudo-estimateurs du maximum de vraisemblance) dans une situation fictive d'échantillons

[1] S. Kumar, méthodologiste principal, Division des méthodes d'enquêtes sociales, Statistique Canada, 4-D2, Immeuble Jean Talon, Parc Tunney, Ottawa (Ontario), K1A 0T6. A.C. Singh, Département de mathématique et de statistique, Memorial University of Newfoundland, St. John's (Terre-Neuve), Canada, A1C 5S7.

binomiaux indépendants pour estimer les paramètres d'un modèle logit après avoir vérifié la validité de l'ajustement de ce modèle pour les données de l'EPA d'octobre 1980. Ils ont constaté que les estimations lissées de taux de chômage étaient beaucoup plus efficaces que les estimations d'enquête dans l'exemple de l'EPA.

Les pseudo-EMV sont particulièrement utiles lorsqu'il est impossible ou difficile de calculer la fonction de vraisemblance à cause de la complexité du plan de sondage. Dans des conditions de régularité acceptables, la méthode des pseudo-EMV produit des estimations convergentes et asymptotiquement normales (Imrey, Koch et Stokes 1982). Dans cet article, nous tentons de produire des estimations asymptotiquement efficaces (au sens défini à la section 3) des paramètres de modèle et, partant, des estimations de domaines. Nous décrivons aussi l'estimateur $Q^{(T)}$ min, qui a été défini dans Singh (1985), à l'aide de la méthode des cotes généralisées; on peut considérer cet estimateur comme analogue à l'estimateur X^2 min de Neyman pour les échantillons aléatoires simples. Il convient de souligner que la méthode des MCP (moindres carrés pondérés) produit également des estimateurs asymptotiquement efficaces dans le cas des plans de sondage complexes (Koch, Freeman et Freeman 1975). Toutefois, ces estimateurs sont habituellement instables avec des échantillons de taille moyenne à cause de la quasi-singularité de la matrice des covariances estimées des estimations de cellules (voir Imrey, Koch et Stokes 1982; Fay 1985). Par contre, les estimateurs $Q^{(T)}$ min sont à l'abri de cette instabilité. Nous verrons que la méthode de $Q^{(T)}$ min peut permettre de contourner le problème de l'instabilité en utilisant une version modifiée de la matrice des covariances estimées, où l'on élimine les valeurs propres relativement minimes de la décomposition spectrale de la matrice.

La section 2 présente la notation des termes utilisés dans l'étude ainsi qu'une brève analyse du test $Q^{(T)}$. Dans la section 3, nous décrivons l'estimateur $Q^{(T)}$ min et son comportement asymptotique. La section 4 contient l'exemple d'application aux données de l'EPA. Cet exemple nous permet de constater avec intérêt que les pseudo-EMV sont presque aussi efficaces que les estimations efficaces $Q^{(T)}$ min au niveau de la cellule. Si nous prenons une mesure globale, comme l'efficacité de trace, nous constatons que les pseudo-EMV ne sont que légèrement inférieures aux estimations $Q^{(T)}$ min. Enfin, la section 5 sert de conclusion.

2. LE TEST $Q^{(T)}$: BRÈVE ANALYSE

Nous allons décrire brièvement le test $Q^{(T)}$ afin de motiver l'utilisation de la méthode d'estimation de $Q^{(T)}$ min (pour plus de détails, voir Singh 1985, Singh et Kumar 1986). Soit I le nombre de domaines disjoints et v_i le paramètre d'intérêt pour le i-ième domaine. Considérons un modèle pour $v = (v_1, v_2, \ldots, v_I)'$ tel que

$$H_0\colon h(v) = X\theta \tag{2.1}$$

où X est une matrice connue $I \times r$ de plein rang r, θ est un r-vecteur de paramètres inconnus et h est une fonction univalente continûment différentiable, par exemple log ou logit.

Posons $\hat{v}$ comme I-vecteur des estimations d'enquête et supposons qu'en vertu du théorème de la limite centrale approprié,

$$\hat{v} \mathrel{\dot{\sim}} MVN(v, \Gamma/n) \tag{2.2}$$

où " $\dot{\sim}$ " signifie "distribué asymptotiquement comme", n est la taille de l'échantillon et Γ est la matrice des covariances asymptotiques de $\sqrt{n}(\hat{v} - v)$.

Choisissons maintenant un faible facteur de réduction de dimension $\epsilon(>0)$ (.01 ou .005 peuvent être des valeurs pratiques d'ϵ). Trouvons un nombre T de telle sorte qu'avec les valeurs propres $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_I$ de la matrice des covariances estimées $\hat{\Gamma}$, nous ayons

$$T = \max \left\{ t : t > r \text{ et } \sum_{i=t}^{I} \hat{\lambda}_i / \sum_{i=1}^{I} \hat{\lambda}_i \geq \epsilon \right\}. \tag{2.3}$$

Bien qu'aléatoire, la variable T peut être considérée comme indépendante pour nos fonctions asymptotiques. Il convient de souligner qu'en l'absence de valeurs propres relativement minimes (c.-à-d. que $\hat{\Gamma}$ n'est pas mal conditionnée), une faible valeur de ϵ n'aura normalement pas d'effet de réduction de dimension et T coïncidera alors avec I.

Examinons le problème qui consiste à tester l'hypothèse nulle H_0 par rapport à l'hypothèse alternative K_0: $h(v) \neq X\theta$ dans la catégorie de tests fondés sur les T premières composantes principales W de $\hat{v}$. Soit P_i le vecteur propre normé correspondant à $\hat{\lambda}_i$ (ce vecteur n'est pas nécessairement unique), et soit M_T la matrice $(I \times T)$ des vecteurs propres P_i correspondant aux T premières valeurs propres les plus élevées. Alors,

$$W = M_T'\hat{v} \dot{\sim} MVN(\mu, D_T/n), \tag{2.4}$$

où

$$\mu = M_T'v, \; D_T = diag(\lambda_1, \ldots, \lambda_T).$$

Lorsqu'il est fondé sur les composantes principales W, le test d'hypothèse qui portait à l'origine sur un vecteur v de dimension I se ramène à un test d'hypothèse sur un paramètre μ de dimension T, où l'hypothèse est définie comme suit:

$$H_0': \mu = M_T'h^{-1}(X\theta) \text{ vs } K_0': \mu \neq M_T'h^{-1}(X\theta). \tag{2.5}$$

On peut calculer la statistique $Q^{(T)}$ comme une cote des composantes principales utilisant la fonction de vraisemblance approximative de θ, qui est définie par la distribution limite de W (2.4) pour le calcul des cotes efficaces (voir Cox et Hinkley 1974, p. 321-324). Nous allons désigner $Q^{(T)}$ comme un test de cote généralisée en vertu duquel H_0 serait rejetée pour les valeurs fortes de la formule quadratique

$$Q^{(T)}(\theta^o) = Y(\theta^o)'\Delta_T Y(\theta^o) - Z_T(\theta^o)'\Lambda_T Z_T(\theta^o) \tag{2.6}$$

$$\dot{\sim} \chi^2_{T-r}$$

où

$$Y(\theta^o) = \hat{v} - v(\theta^o), \; \Delta_T = n \sum_{i=1}^{T} (P_i P_i'/\hat{\lambda}_i),$$

$$Z_T(\theta^o) = B'\Delta_T Y(\theta^o), \; B = (\partial v/\partial\theta), \; \Lambda_T = (B'\Delta_T B)^{-1},$$

et θ^o est un point fixe dans l'espace des paramètres nuls. Dans le calcul de $Q^{(T)}$, n'importe quelle estimation racine n-convergente de θ, sous H_0, peut être substituée à θ^o, comme peut l'être la pseudo-EMV de θ. Il est à noter que $Q^{(T)}$ dans (2.6) est en réalité définie par une formule quadratique dans W mais que, pour des raisons de commodité, elle est exprimée en fonction de $\hat{v}$.

Pour tester H_0 par rapport à K_0 dans la catégorie de tests fondés sur W, l'optimalité asymptotique du test $Q^{(T)}$ provient de celle de la cote. Pour les faibles valeurs de $\epsilon > 0$, $\hat{v}$ et W seront comparables en ce sens que les composantes principales, dans ce cas, assureront

la meilleure forme de réduction de dimension qui soit, avec une perte minimum d'information. Le test $Q^{(T)}$ (pour de faibles valeurs de ϵ) devrait donc être robuste par rapport au test Q correspondant où il n'y a aucune réduction de dimension. En revanche, Q peut être instable (au sens d'une probabilité d'erreur de première espèce gonflée) dans le cas des échantillons finis à cause de la quasi-singularité possible de $\hat{\Gamma}$. Le test $Q^{(T)}$ est censé atténuer ce problème d'instabilité mais cela ne peut se faire sans qu'il y ait une perte d'information; Q peut donc renfermer des composantes peu fiables dans la direction des vecteurs propres correspondant aux valeurs propres relativement minimes. Cette perte d'information entraîne une diminution de la puissance du test $Q^{(T)}$ pour des alternatives en direction de la (quasi-) singularité. Cependant, cette perte de puissance est compensée par un gain au niveau du contrôle de l'erreur de première espèce. Comme H_0 est un sous-ensemble de H_0', $Q^{(T)}$ sera un test conservatif pour H_0; cela sert à assurer le contrôle de l'instabilité.

Une version spéciale de $Q^{(T)}$ (θ°), asymptotiquement équivalente et décrite par une formule plus simple et semblable à celle du critère X^2 standard de Pearson-Fisher, peut être obtenue en remplaçant θ° par un estimateur $\tilde{\theta}$ qui minimise l'expression $(\hat{v} - v(\theta))'\Delta_T(\hat{v} - v(\theta))$. Nous avons donc,

$$\begin{aligned} Q^{(T)}(\tilde{\theta}) &= Y(\tilde{\theta})'\Delta_T Y(\tilde{\theta}) \\ &= \sum_{i=1}^{T} [P_i'(\hat{v} - v(\tilde{\theta}))]^2/\hat{\lambda}_i \qquad (2.7) \\ &\sim \chi^2_{T-r} \end{aligned}$$

Supposons maintenant que le test $Q^{(T)}$ ou un autre test, comme le test X^2 corrigé, nous a amené à croire qu'un modèle H_0 convient à un vecteur de données $\hat{v}$. Dans la section suivante, nous définissons une méthode asymptotiquement efficace pour estimer les paramètres θ sous H_0, utilisant la statistique $Q^{(T)}$. Les estimations ainsi obtenues permettent d'établir une série d'estimations lissées de v qui correspondent aux estimations d'enquête $\hat{v}$.

3. L'ESTIMATEUR $Q^{(T)}$ MIN

Considérons la fonction de vraisemblance approximative pour la moyenne μ des T premières composantes principales W de $\hat{v}$, qui sont définies par l'équation (2.4). Supposons que le modèle H_0: $h(v) = X\theta$ est accepté. Alors, la fonction du noyau $K(\theta)$ de la fonction de vraisemblance approximative pour $\mu(\theta)$ est donnée par

$$\begin{aligned} K(\theta) &= (W - \mu(\theta))'D_T^{-1}(W - \mu(\theta)) \\ &= (\hat{v} - v(\theta))'\Delta_T(\hat{v} - v(\theta)) \qquad (3.1) \end{aligned}$$

La valeur $\tilde{\theta}$ qui minimise $K(\theta)$ correspond à l'EMV de θ pour la fonction de vraisemblance approximative de μ sous H_0. Cet estimateur $\tilde{\theta}$ sera asymptotiquement efficace (ou le meilleur estimateur asymptotiquement normal (MEAN) au sens de Neyman, 1949) dans une catégorie restreinte, notamment dans la catégorie de tests fondés sur W. D'après l'estimateur de X^2 min de Neyman (1949), l'estimateur $\tilde{\theta}$ est donc devenu dans Singh (1985) l'estimateur de $Q^{(T)}$ min. Il convient de souligner que l'estimateur $\tilde{\theta}$ dépend du facteur de réduction de dimension ϵ par Δ_T. Par conséquent, $\tilde{\theta}$ varie avec ϵ.

Étant donné les composantes principales W et sous H_0, nous pouvons calculer les estimations lissées de v de la façon suivante. Nous devons trouver tout d'abord la valeur $\tilde{\theta}$ qui minimise $K(\theta)$; en d'autres termes, la valeur recherchée est la solution de r équations

$$B'\Delta_T(\hat{v} - v(\theta)) = 0 \tag{3.2}$$

où $B(= \partial v / \partial \theta)$ et v ont tous deux rapport avec θ. Une méthode itérative comme celle de Newton-Raphson peut servir à résoudre l'équation (3.2). La valeur θ peut être estimée par la méthode des moindres carrés pondérés (MCP) ou la méthode des pseudo-EMV. Il est alors possible de calculer l'estimateur $Q^{(T)}$ min de v à l'aide de la formule suivante

$$\tilde{v} = h^{-1}(X\tilde{\theta}). \tag{3.3}$$

Les comportements asymptotiques de $\tilde{\theta}$ et de $\tilde{v}$ sont définis par la proposition suivante.

Proposition 3.1 Désignons l'expression $(B'\Delta_T B)^{-1}$ par le symbole $\wedge_T$. Nous avons

$$\begin{aligned} &\text{(a)} \quad \tilde{\theta} - \theta \approx \wedge_T B'\Delta_T(\hat{v} - v(\theta)) \dot{\sim} MVN(0, \wedge_T) \\ &\text{(b)} \quad \tilde{v} - v \approx B\wedge_T B'\Delta_T(\hat{v} - v(\theta)) \dot{\sim} MVN(0, B\wedge_T B') \end{aligned} \tag{3.4}$$

où " $\approx$ " signifie qu'il y a très peu de possibilité que la différence entre les deux membres de l'expression soient importante.

Cette proposition se démontre par l'application de la méthode-δ aux fonctions $B'\Delta_T(\hat{v} - v(\theta))$ et $\tilde{v} - v(\theta)$, ce qui donne

$$B'\Delta_T(\hat{v} - v(\theta)) - (B'\Delta_T B)(\tilde{\theta} - \theta) = o_p(1),$$

$$\tilde{v} - v(\theta) - B(\tilde{\theta} - \theta) = o_p(1).$$

La proposition ci-dessus nous amène à conclure que la matrice des covariances asymptotiques de l'estimateur $Q^{(T)}$ min $\tilde{\theta}$ est l'inverse de la matrice d'information $B'\Delta_T B$ pour θ, qui a été déterminée à l'aide de la fonction de vraisemblance approximative de θ définie en (2.4). Par conséquent, lorsqu'il n'y a pas de réduction de dimension, l'estimateur $\tilde{\theta}$ sera asymptotiquement équivalent à l'estimateur par les MCP de Koch, Freeman et Freeman (1975). Comme nous l'avons dit dans l'introduction, l'estimateur par les MCP a habituellement un comportement instable dans le cas des échantillons finis à cause d'une estimation inefficace de Γ. Par contre, pour un $\tilde{\theta}$ donné ($\epsilon > 0$), l'estimateur $\tilde{\theta}$ a normalement un comportement stable avec les échantillons finis, en ce sens que son comportement asymptotique peut fournir une bonne approximation de son comportement en général. Cette condition ne peut être réalisée sans compromettre l'optimalité asymptotique de $\tilde{\theta}$ car, pour que cette condition se réalise, cet estimateur doit être limité à une catégorie moindre, c'est-à-dire à la catégorie des estimateurs fondés sur les T premières composantes principales W. En revanche, l'estimateur par les MCP conserve son optimalité asymptotique dans une catégorie plus large, notamment celle des estimateurs fondés sur le vecteur complet de données $\hat{v}$. Si, pour une valeur faible de ϵ, la statistique $Q^{(T)}$ donne un résultat non significatif pour H_0, l'estimateur $Q^{(T)}$ min correspondant ($\hat{v}$) est susceptible de se comparer avantageusement à l'estimateur par les MCP au point de vue de la robustesse.

4. ESTIMATIONS DES TAUX DE CHÔMAGE PAR LA MÉTHODE DE $Q^{(T)}$ MIN

Considérons les données de l'enquête sur la population active (EPA) d'octobre 1980 que Kumar et Rao (1984, 1986) et Roberts, Rao et Kumar (1987) ont analysées en appliquant l'extension de la méthode du X^2 corrigé de Rao-Scott à la régression logistique. Ces auteurs ont montré que le modèle logit défini ci-dessous s'ajustait bien aux estimations d'enquête de taux d'emploi (v_{jt}) pour le tableau de 60 cellules formées simultanément en fonction de l'âge (10 catégories) et du niveau d'instruction (6 catégories). Le modèle est

$$log \frac{v_{j\ell}}{1 - v_{j\ell}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_\ell \tag{4.1}$$

où A_j représente le point médian $12 + 5j$ pour le j-ième groupe d'âge ($j = 1, \ldots, 10$), et $E_\ell (\ell = 1, \ldots, 6)$ représente le nombre médian d'années de scolarité pour chaque catégorie, soit 7, 10, 12, 13, 14 et 16.

On peut exprimer le modèle (4.1) au moyen de la notation définie à la section 2 en numérotant les soixante cellules par ordre lexicographique. Ainsi, (4.1) peut être récrite $h(v) = X\theta$, où v est le vecteur des taux d'emploi, h est la fonction logit, X est une matrice 60×4 dont la i-ième ligne est $(1, A_i, A_i^2, E_i)$, et θ est $(\beta_0, \beta_1, \beta_2, \beta_3)'$. Nous avons aussi

$$H = (\partial h / \partial v) = D_v^{-1} D_{1-v}^{-1}, B = H^{-1} X, \tag{4.2}$$

où D_v et D_{1-v} sont des matrices diagonales dont les éléments de la diagonale sont définis par les indices inférieurs.

À l'aide de la pseudo-fonction de vraisemblance appliquée à la distribution binomiale mixte, Kumar et Rao (1984) ont calculé les pseudo-EMV de θ pour le modèle (4.1) et ont obtenu les valeurs suivantes:

$$\bar{\theta} = (-3.10, 0.211, -0.00218, 0.1509)'. \tag{4.3}$$

Ils ont également calculé la valeur du X^2 corrigé de premier ordre de Rao-Scott (qu'ils désignent par G_c^2); la valeur obtenue, 55.3, entraîne l'acceptation du modèle (2.1) quand on le compare à la distribution X_{56}^2.

Le modèle (4.1) a aussi été testé à l'aide de la méthode $Q^{(T)}$ (voir Singh 1985, Singh et Kumar 1986); là encore, le test s'est soldé par l'acceptation du modèle. Pour $\epsilon = .01$, nous avons $T = 51$ selon la matrice des covariances estimées $\hat{\Gamma}$, qui a été calculée par Kumar et Rao (1984). Si nous prenons maintenant les pseudo-EMV $\bar{\theta}$, nous avons

$$Q^{(51)}(\bar{\theta}) = 58.665 - 4.454 = 54.211 \tag{4.4}$$

Lorsque $\epsilon = .005$, T est égal à 54 et

$$Q^{(54)}(\bar{\theta}) = 67.774 - 2.343 = 65.431 \tag{4.5}$$

Lorsque $\epsilon = 0$, $T = 58$ parce que deux cellules renferment des taux de chômage nuls. Dans ce cas,

$$Q^{(58)}(\bar{\theta}) = 87.302 - 0.812 = 86.49 \tag{4.6}$$

En comparant les valeurs de $Q^{(51)}$, $Q^{(54)}$ et $Q^{(58)}$ aux distributions χ_{47}^2, χ_{50}^2 et χ_{54}^2 respectivement, nous constatons que les deux premiers tests entraînent l'acceptation de (4.1) tandis que le dernier entraîne son rejet. Il est possible de vérifier le degré d'instabilité des estimateurs en examinant l'écart qui existe entre $Q^{(58)}$ et $Q^{(T)}$ (pour $T = 51, 54$); cet écart peut être jugé très significatif en se référant à la distribution χ_{58-T}^2. L'analyse révèle une certaine instabilité pour le test Q où il n'y a aucune réduction de dimension. Il est clair que le test des MCP renfermerait aussi des problèmes d'instabilité à cause de la difficulté que l'on aurait à inverser la matrice $\hat{\Gamma}$ qui est singulière. La méthode de $Q^{(T)}$ min serait donc préférable à la méthode de Q min ou à celle des MCP. Dans le but de minimiser la perte d'information, il est recommandé d'utiliser la méthode qui a la valeur de T la plus élevée à la condition, bien sûr, que la valeur de $Q^{(T)}$ correspondante entraîne l'acceptation du modèle.

Nous passons maintenant au calcul d'estimations asymptotiquement efficaces. Les estimations de Q min et les estimations par les MCP n'ont pas été calculées parce que $\hat{\Gamma}$ est singulière. Les estimations $\tilde{\theta}$ de $Q^{(T)}$ min ont été calculées pour $\epsilon = .005$ et $\epsilon = .01$ en appliquant la méthode itérative de Newton-Raphson et en utilisant $\bar{\theta}$ comme l'estimateur initial de θ dans la résolution de l'équation (3.2). Les valeurs de $\tilde{\theta}_T$ et de $Q^{(T)}(\tilde{\theta})$ (dans ce dernier cas, le terme négatif de (2.6) disparaît) pour $\epsilon = .005$ et $T = 54$ sont les suivantes

$$\tilde{\theta}_{54} = (-2.7112,\ 0.1944,\ -0.00196,\ 0.1432)', \text{ et}$$

$$Q^{(54)}(\tilde{\theta}_{54}) = 63.4737. \tag{4.7}$$

Pour $\epsilon = .01$, $T = 51$, nous avons

$$\tilde{\theta}_{51} = (-2.6739,\ 0.19702,\ -0.00202,\ 0.1364)', \text{ et}$$

$$Q^{(51)}(\tilde{\theta}_{51}) = 55.2518. \tag{4.8}$$

Tableau 1
Efficacité des estimations lissées des taux de chômage par rapport aux estimations d'enquête[a]

N° de cellule	$Q^{(51)}$ Min	$Q^{(54)}$ Min	Pseudo-EMV	N° de celule	$Q^{(51)}$ Min	$Q^{(54)}$ Min	Pseudo-EMV
1	5.87	5.74	5.44	31	9.01	9.32	8.65
2	3.62	3.62	3.28	32	8.76	9.46	10.68
3	3.45	3.55	3.12	33	36.93	42.93	51.59
4	52.45	51.65	43.46	34	51.55	60.23	81.12
5	104.77	114.30	96.21	35	69.76	79.93	98.37
7	5.33	5.14	4.38	36	9.17	11.01	15.07
8	9.36	9.53	8.09	37	3.48	3.01	3.45
9	6.85	7.16	6.70	38	13.74	15.91	18.00
10	25.65	28.40	26.31	39	66.87	80.98	97.30
11	13.34	14.13	17.73	40	154.81	187.73	221.50
12	27.74	30.85	30.85	41	49.14	67.56	80.61
13	8.64	8.84	7.15	42	17.32	21.73	24.98
14	13.84	13.84	12.37	43	8.57	9.28	8.49
15	8.20	8.49	9.47	44	27.42	31.65	30.74
16	23.14	24.09	27.75	45	58.55	70.67	75.72
17	18.20	18.20	21.49	46	94.11	114.13	121.49
18	9.87	11.14	12.51	47	82.12	112.65	108.52
19	15.87	16.03	13.66	48	26.54	39.41	41.22
20	11.44	11.98	12.56	49	4.95	5.37	4.41
21	12.39	12.39	15.53	50	12.11	14.10	11.17
22	24.83	24.83	32.02	51	6.75	8.61	7.50
23	16.43	18.16	21.55	52	8.83	11.45	9.90
24	6.98	7.83	10.06	53	52.64	71.49	61.14
25	7.49	7.74	6.99	55	3.59	3.93	3.03
26	10.33	11.33	12.32	56	7.33	8.96	8.23
27	6.47	7.18	8.69	57	23.50	29.83	22.11
28	125.81	140.57	172.91	58	221.23	294.59	208.77
29	33.88	38.13	52.00	59	6.45	8.82	6.62
30	14.89	15.24	20.43	60	38.90	52.84	41.96

[a] Les cellules 6 et 54 ne sont pas inclues parce que les taux de chômage observés correspondants étaient zéro.

Les valeurs de $Q^{(T)}(\tilde{\theta})$ pour $T = 54$ et 51 et celles de $Q^{(T)}(\bar{\theta})$ pour les mêmes valeurs de T permettent de tirer des conclusions similaires.

Le tableau 1 donne les efficacités relatives aux estimations du taux de chômage $1 - v$ pour les 60 cellules correspondant aux trois estimations lissées. Les trois sortes d'estimations lissées sont: les pseudo-EMV, les estimations de $Q^{(51)}$ min et celles de $Q^{(54)}$ min. Les variances des pseudo-EMV sont tirés de Kumar et Rao (1986) tandis que celles des estimations de $Q^{(T)}$ min ont été établis à l'aide des éléments diagonaux de $B \wedge_T B'$ dans (3.4). Comme le soulignent Kumar et Rao (1986) à propos des pseudo-EMV, les estimations lissées obtenues par la méthode de $Q^{(T)}$ min produisent des gains d'efficacité considérables par rapport aux estimations d'enquête. Le rapport entre l'efficacité de trace des estimations lissées et l'efficacité de trace des estimations d'enquête est de 17.9, 18.95 et 19.88 respectivement pour la méthode des pseudo-EMV, celle de $Q^{(51)}$ min et celle de $Q^{(54)}$ min. Ainsi, la méthode de $Q^{(T)}$ min produit des estimations lissées légèrement plus efficaces que la méthode des pseudo-EMV. En ce qui a trait à l'efficacité au niveau de la cellule, le tableau 1 indique que les pseudo-EMV se comparent très avantageusement aux estimations efficaces de $Q^{(T)}$ min.

5. CONCLUSION

En ce qui a trait au calcul des pseudo-EMV, la forme usuelle de la fonction de vraisemblance se rapporte à des échantillons aléatoires simples (c.-à-d., échantillonnage multinomial ou échantillonnage multinomial mixte). Les pseudo-EMV produisent effectivement des estimations de paramètres de modèle convergentes sans que l'on ait à estimer la matrice des covariances Γ. Toutefois, ce genre d'estimations ne sont pas asymptotiquement efficaces dans le cas de données d'échantillons complexes. En revanche, les estimations de $Q^{(T)}$ min sont asymptotiquement efficaces en ce qui concerne la catégorie d'estimateurs fondés sur W (les T premières composantes principales du vecteur $\hat{v}$ des estimations d'enquête). Pour examiner la performance relative obtenue à l'aide du pseudo-EMV et du $Q^{(T)}$ min, il serait souhaitable d'effectuer une étude de simulation pour comparer les efficacités. On adapte les estimations aux plans de sondage complexes en utilisant une matrice $\hat{\Gamma}$. Si $\hat{\Gamma}$ n'est pas mal conditionnée, c'est-à-dire qu'elle n'a aucune valeur propre relativement minime, les estimateurs par les MCP sont parfaitement stables et, par conséquent, asymptotiquement efficaces. Dans ce genre de situation, on constatera souvent qu'il n'y a pas de réduction de dimension pour de faibles valeurs de ϵ et que T coïncide avec I. De plus, les estimations de $Q^{(T)}$ min ne perdront pas de leur efficacité par rapport aux estimations par les MCP. Toutefois, en raison des problèmes d'instabilité qui sont si souvent liés aux données qualitatives recoupées, les estimations de $Q^{(T)}$ min devraient bien valoir en robustesse les estimations par les MCP.

REMERCIEMENTS

Les recherches du premier auteur ont été appuyées par Statistique Canada et le Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

COX, D.R., et HINKLEY, D.W. (1974). *Theoretical Statistics*. London: Chapman and Hall.

FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

IMREY, P.B., KOCH, G.G., et STOKES, M.E. (1982). Categorical data analysis: Some reflections on the log-linear model and logistic regression. Part II: Data analysis. *International Statistical Review*, 50, 35-63.

KOCH, G.G., FREEMAN, D.H. Jr., et FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.

KUMAR, S., et RAO, J.N.K. (1984). Régression logistique et analyse de données de l'enquête sur la population active. *Techniques d'enquête*, 10, 62-81.

KUMAR, S., et RAO, J.N.K. (1986). On smoothed estimates of unemployment rates from labour force survey data. Dans *Small Area Statistics: An International Symposium '85* (éds. R. Platek et M.P. Singh), Ottawa: Carleton University.

NEYMAN, J. (1949). Contribution to the Theory of the X^2 test. Dans *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability* (éd. J. Neyman), Berkeley: University of California Press, 230-273.

RAO, J.N.K., et SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two way tables. *Journal of the American Statistical Association*, 76, 221-230.

RAO, J.N.K., et SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

ROBERTS, G.R. (1985). *Contributions to chi-squared tests with survey data*. Thèse de doctorat, Carleton University, Ottawa.

ROBERTS, G., RAO, J.N.K., et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

SINGH, A.C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Document de travail, Division des méthodes d'enquêtes sociales, Statistique Canada.

SINGH, A.C., et KUMAR, S. (1986). Categorical data analysis for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (à paraître).

Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 93-100
Statistique Canada

Méthode d'échantillonnage avec probabilités de sélection proportionnelles à la taille

A. DEY et A.K. SRIVASTAVA[1]

RÉSUMÉ

On propose ici un nouveau plan d'échantillonnage sans remise avec probabilités inégales de sélection de n unités ($n > 2$) dans une population finie. Ce plan assure que les probabilités de sélection sont proportionnelles à la taille. Il offre l'avantage de simplifier le processus de sélection et d'estimation et de produire un estimateur de la variance non négatif. On montre que la variance de l'estimateur de Horvitz-Thompson obtenu à l'aide de ce nouveau plan est plus petite que celle des estimateurs produits habituellement selon un plan d'échantillonnage avec probabilités proportionnelles à la taille et avec remise. Le plan proposé donne également de bons résultats par rapport au plan d'échantillonnage sans remise élaboré par Sampford (1967).

MOTS CLÉS: Échantillonnage avec probabilités inégales; estimateur de Horvitz-Thompson.

1. INTRODUCTION

Dans un plan d'échantillonnage sans remise avec probabilités inégales de sélection de n unités à partir d'une population finie comprenant N unités, si π_i désigne la probabilité d'inclusion de la i-ième unité dans l'échantillon, $i = 1, 2, \ldots, N$, l'estimateur de Horvitz-Thompson (1952) (l'estimateur H-T) de Y, qui est la valeur totale de la variable d'intérêt, y, dans la population étudiée, s'exprime

$$\hat{Y} = \sum_{i \in s} (y_i/\pi_i), \tag{1.1}$$

où y_i est la valeur de y chez la i-ième unité et où la sommation porte sur les unités incluses dans l'échantillon. La variance de $\hat{Y}$ est

$$Var\ (\hat{Y}) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij})\ (y_i/\pi_i - y_j/\pi_j)^2 \tag{1.2}$$

où π_{ij} désigne la probabilité d'inclusion du couple d'unités i et j dans l'échantillon ($i \neq j$; $i, j = 1, 2, \ldots, N$).

On peut s'attendre que la variance de $\hat{Y}$ sera beaucoup moins grande si on utilise un plan d'échantillonnage qui assure que les π_i sont proportionnels à une mesure donnée de la taille, disons, x_i, pour $i = 1, 2, \ldots, N$, où on suppose que les x_i sont presque proportionnels aux y_i. Les plans d'échantillonnage dans lesquels les π_i sont proportionnels aux x_i sont appelés des plans d'échantillonnage avec probabilités de sélection proportionnelles à la taille (PSPT). Pour une présentation détaillée des techniques d'échantillonnage avec probabilités inégales, y compris les plans d'échantillonage avec PSPT, voir la monographie de Brewer et Hanif (1983).

Les plans d'échantillonnage avec probabilités inégales et sans remise, en général, et les plans d'échantillonnage avec PSPT, en particulier, devraient posséder entre autres caractéristi-

[1] A. Dey et A.K. Srivastava, Indian Agricultural Statistics Research Institute, Library Avenue, Nouvelle-Delhi 110012, Inde.

ques souhaitables l'avantage de simplifier le processus de sélection et d'estimation, de permettre de calculer un estimateur de la variance non négatif et d'être plus efficaces que les techniques d'échantillonnage avec probabilités proportionnelles à la taille (PPT) et avec remise. Malheureusement, pour des échantillons de taille plus grande que deux, il n'y a pas encore beaucoup de techniques qui satisfont pleinement à toutes ces exigences.

Dans le présent article, on propose un plan d'échantillonnage pour des échantillons de taille arbitraire n où $n > 2$. La technique est plutôt simple tant pour la sélection de l'échantillon que pour l'estimation, étant donné qu'on peut disposer d'expressions compactes pour les π_i. Elle donne également la possibilité de calculer un estimateur positif de la variance de l'estimateur H-T de Y. La performance réalisée par l'estimateur H-T au moyen du plan proposé est comparée à celle de l'estimateur résultant de l'utilisation d'une technique d'échantillonnage avec PPT et avec remise. À partir de cette comparaison, on trouve une condition suffisante simple qui, si elle est satisfaite, assure une performance du nouveau plan supérieure à celle de l'autre technique. Les résultats d'une étude empirique faite avec quelques populations naturelles indiquent que le plan proposé se compare avantageusement à celui élaboré par Sampford (1967).

2. MÉTHODE D'ÉCHANTILLONNAGE

Soient une population de N unités, y la variable d'intérêt et x une variable auxiliaire prise comme mesure de la taille. On suppose que les valeurs de x sont connues pour toutes les unités de la population. On veut maintenant tirer un échantillon de taille n ($n > 2$). Pour commencer, on suppose que n est *pair*.

On divise la population en m groupes m ($> n/2$) de telle sorte que le i-ième groupe contienne N_i unités ($N_i > 2$) et $i = 1, 2, \ldots, m$) et que, pour chaque groupe,

$$X_i/X > (n - 2)/[n(m - 1)], \tag{2.1}$$

où

$$X_i = \sum_{u=1}^{N_i} x_{i_u},$$

x_{i_u} est la valeur de x chez la u-ième unité du i-ième groupe et $X = X_1 + X_2 + \ldots + X_m$.

La condition (2.1) est satisfaite si les X_i ($i = 1, 2, \ldots, m$) sont presque égaux. Il a été montré que, dans des populations réelles, telles que celles considérées par Rao et Bayless (1969) et d'autres auteurs, cette condition est satisfaite pour plusieurs valeurs de m si les groupes sont formés de manière que leurs tailles X_i respectives sont presque égales entre elles. Rao et Lanke (1984) ont proposé une méthode de groupement des unités dans laquelle N unités sont groupées en R groupes de telle sorte que la valeur totale de chaque groupe, X_i, est presque égale d'un groupe à un autre et que la taille des groupes est soit $[N/R]$ ou $[N/R] + 1$, où $[x]$ est le plus grand nombre entier contenu dans x. On peut également appliquer la méthode de Rao-Lanke pour former les groupes.

Une fois les m groupes formés, la méthode d'échantillonnage proposée comprend les étapes suivantes:

Étape 1. Sélection de $n/2$ groupes parmi les m groupes, au moyen de la méthode d'échantillonnage de Midzuno (1951) avec probabilités $\{P_i'\}$, c'est-à-dire sélection d'un groupe avec probabilité

$$P_i' = [n(m-1)P_i - (n-2)]/(2m-n), \text{ où } P_i = X_i/X,$$

et des $(n/2) - 1$ autres groupes avec probabilités égales et sans remise.

Étape 2. Sélection de deux unités, dans chacun des groupes choisis, suivant l'une quelconque des méthodes avec PSPT, par exemple à l'aide de la méthode de Durbin (1967), c'est-à-dire sélection dans le i-ième groupe choisi ($i = 1, 2, \ldots, n/2$) d'une unité avec probabilité

$$p_{i_u|i} = x_{i_u}/X_i$$

et de la deuxième unité avec probabilité révisée

$$p_{i_u|i_v} = x_{i_v}[1/(X_i - 2x_{i_v}) + 1/(X_i - 2x_{i_u})]/D_i$$

où
$$D_i = [1 + \sum_{u=l}^{N_i} x_{i_u}/(X_i - 2x_{i_u})].$$

Avec cette méthode d'échantillonnage, la probabilité de sélection de la i_u-ième unité est évidemment

$$\pi_{i_u} = n\, p_{i_u} \tag{2.2}$$

où

$$p_{i_u} = x_{i_u}/X.$$

De plus, les probabilités de sélection d'une paire d'unités sont

$$\pi_{i_u i_v} = \frac{n\, p_{i_u} p_{i_v} (P_i - p_{i_u} - p_{i_v})}{D_i (P_i - 2\, p_{i_u})\ (P_i - 2\, p_{i_v})} \tag{2.3}$$

et

$$\pi_{i_u j_v} = \frac{n\,(n-2)\, p_{i_u} p_{j_v}}{(m-1)(m-2)\, P_i P_j}\ [(m-1)(P_i + P_j) - 1], \tag{2.4}$$

$$i \neq j;\ i, j = 1, 2, \ldots, m.$$

Ainsi, nous voyons que le plan proposé est bel et bien un plan d'échantillonnage avec PSPT.

Tel qu'il a été mentionné précédemment, on peut utiliser à l'étape 2 de la méthode proposée n'importe quel plan avec PSPT pour sélectionner deux unités. Comme la méthode

de Durbin (1967), qui est équivalente à celle de Rao (1963) et de Brewer (1963), donne en général de bons résultats, c'est elle qu'on a décidé d'appliquer à l'étape 2.

3. ESTIMATEUR DE LA VARIANCE

Deux estimateurs sans biais bien connus de la variance de $\hat{Y}$, $Var(\hat{Y})$, ont été élaborés par Horvitz et Thompson (1952) et par Yates et Grundy (1953). Ces deux estimateurs présentent l'inconvénient de prendre parfois des valeurs négatives. Dans la présente section, on examine un estimateur positif de la variance qui met à profit le fait que le plan d'échantillonnage proposé est à deux degrés.

À l'aide d'un résultat Des Raj (1966), un estimateur non biaisé de $Var(\hat{Y})$ prend la forme suivante:

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n/2} \pi_i^{-1} \sum_{u<v}\sum \left[\frac{\pi_{i_u|i}\,\pi_{i_v|i}}{\pi_{i_u i_v|i}} - 1\right]\left[\frac{y_{i_u}}{\pi_{i_u|i}} - \frac{y_{i_v}}{\pi_{i_v|i}}\right]^2$$

$$+ \sum_{i<j}^{n/2}\sum^{n/2} \left(\frac{\pi_i\pi_j}{\pi_{ij}} - 1\right)\left[\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j}\right]^2, \tag{3.1}$$

où

$$\pi_i = nP_i/2,$$

$$\pi_{ij} = \frac{n(n-2)}{4(m-2)}\{(P_i + P_j) - 1/(m-1)\},$$

$$\pi_{i_u|i} = 2p_{i_u}/P_i,$$

$$\pi_{i_u i_v|i} = \frac{2p_{i_u}p_{i_v}(P_i - p_{i_u} - p_{i_v})}{D_i P_i (P_i - 2p_{i_u})(P_i - 2p_{i_v})},$$

et

$$\hat{Y}_i = \sum_{u=1}^{2} y_{i_u}/\pi_{i_u|i}, \tag{3.2}$$

y_{i_u} étant la valeur de y chez la u-ième unité dans le i-ième groupe.

Les deux termes du membre de droite de l'équation (3.1) correspondent à l'estimateur de la variance de Yates-Grundy calculés respectivement pour la méthode de Durbin (étape 2) et pour la méthode de Midzuno (étape 1). Comme l'estimateur de la variance de Yates-Grundy est toujours positif pour chacune de ces deux méthodes d'échantillonnage, il s'ensuit que l'estimateur de la variance de l'équation (3.1) est également positif. Cependant, l'estimateur en (3.1) n'est ni l'estimateur de la variance de Horvitz-Thompson ni l'estimateur de la variance de Yates-Grundy.

4. COMPARAISON DU PLAN PROPOSÉ ET DE LA STRATÉGIE AVEC PPT AVEC REMISE

Dans la présente section, nous comparons l'efficacité des deux stratégies suivantes:

Stratégie 1. Plan d'échantillonnage proposé utilisé avec l'estimateur de Horvitz-Thompson.

Stratégie 2. Plan d'échantillonnage avec PPT et avec remise utilisé avec l'estimateur habituel.

La stratégie 1 est plus efficace (variance plus petite) que la stratégie 2 si et seulement si

$$\begin{aligned}&\sum_{i=1}^{m}\sum_{u\neq v}^{N_i}\pi_{i_u i_v}(y_{i_u}/p_{i_u}-Y)(y_{i_v}/p_{i_v}-Y)\\&+\sum_{i\neq j}^{m}\sum_{u}^{N_i}\sum_{v}^{N_j}\pi_{i_u j_v}(y_{i_u}/p_{i_u}-Y)(y_{j_v}/p_{j_v}-Y)<0.\end{aligned}\tag{4.1}$$

Après un certain nombre de manipulations algébriques longues mais élémentaires, l'inégalité (4.1) se ramène à ceci

$$\begin{aligned}&-\sum_{i=1}^{n}(n/D_i)\sum_{u=1}^{N_i}(y_{i_u}-Y_i p_{i_u}/P_i)^2/(P_i-2p_{i_u})\\&-n(n-2)\left[\sum_{i=1}^{m}(Y_i/P_i-Y)\right]^2/[(m-2)(m-1)]\\&-n(m-2)^{-1}\sum_{i=1}^{m}[\{(2n-m-2)P_i-(n-2)(m-1)^{-1}\}(Y_i/P_i-Y)^2]<0,\end{aligned}\tag{4.2}$$

où

$$Y_i=\sum_u y_{i_u}$$

De toute évidence, (4.2) se vérifie si

(i) $(2n-m-2)>0$, et si

(ii) $P_i>(n-2)/[(m-1)(2n-m-2)]$. (4.3)

De plus, comme pour la première étape du plan d'échantillonnage nous utilisons la méthode de Midzuno avec des probabilités révisées $\{P_i'\}$, chacun des P_i doit satisfaire la condition (2.1), c'est-à-dire que chaque P_i doit satisfaire

$$P_i>(n-2)/[n(m-1)].$$

Par conséquent, (4.2) se vérifie si

$$m\leq(n-2).\tag{4.4}$$

Tableau 1

Description des populations

Numéro de population	Source	N	y	x
1.	Des Raj (1965)	20	Nombre de ménages	Nombre de ménages estimé à l'oeil
2.	Rao (1963)	14	Nombre d'acres de maïs en 1960	Nombre d'acres de maïs en 1958
3.	Cochran (1963, p. 204)	10	Poids des pêches	Poids des pêches estimé à l'oeil
4.	Hanurav (1967)	20	Population en 1967	Population en 1957
5.	Hanurav (1967)	19	Population en 1967	Population en 1957
6.	Hanurav (1967)	16	Population en 1967	Population en 1957
7.	Hanurav (1967)	17	Population en 1967	Population en 1957
8.	Cochran (1963, p. 325)	10	Nombre de personnes par îlot de logements	Nombre de pièces par îlot de logements
9.	Cochran (1963, p. 156, villes 1-16)	16	Population en 1930	Population en 1920
10.	Cochran (1963, p. 156, villes) 33-49)	17	Population en 1930	Population en 1920
11.	Sampford (1962, p. 61)	35	Nombre d'acres d'avoine en 1957	Nombre d'acres d'avoine en 1947
12.	Sukhatme et Sukhatme (1970, p. 256, cercles 1-20)	20	Nombre d'acres de blé	Nombre de villages
13.	Sukhatme et Sukhatme (1970, p. 256, cercles 21-40)	20	Nombre d'acres de blé	Nombre de villages
14.	Yates (1960, p. 163)	20	Volume de bois d'oeuvre	Volume de bois d'oeuvre estimé à l'oeil

Tableau 2

Efficacité relative en pourcentage des stratégies 1 et 3 par rapport à la stratégie 2, avec les populations décrites au tableau 1 ($n = 4$)

Population Numéro	Stratégie 1				Stratégie 3
	$m = 3$	4	5	6	
1.	130.1	118.7	120.8	124.5	127.8
2.	132.6	130.2	–	–	127.1
3.	149.1	–	–	–	147.9
4.	120.7	120.6	122.7	129.7	117.8
5.	129.1	138.7	158.7	–	125.1
6.	158.0	173.1	–	–	139.5
7.	151.9	144.8	169.2	–	131.9
8.	168.5	–	–	–	145.5
9.	118.3	116.3	–	–	109.5
10.	126.6	–	–	–	112.2
11.	113.8	116.2	135.6	129.9	113.8
12.	117.4	128.0	119.0	–	119.3
13.	122.2	120.6	–	–	119.7
14.	124.8	123.1	115.4	113.2	116.3

Il semble donc que, pour que la stratégie 1 soit supérieure à la stratégie 2, il faille choisir m de manière que

$$n/2 < m \leq (n - 2). \tag{4.5}$$

Il est toutefois clair que (4.4) n'est qu'une condition suffisante mais non nécessaire. Dans le cas où $n > 6$, la condition (4.5) donne plusieurs choix pour la valeur de m, tandis que si $n = 6$, (4.5) implique que $m = 3$. Pour $n = 4$, aucune valuer de m ne peut satisfaire (4.5). On s'est donc employé à étudier l'efficacité de la stratégie 1, lorsque $n = 4$, avec un certain nombre de populations naturelles pour diverses valeurs de m qui ne satisfont pas (4.5). Une description des populations est donnée au tableau 1, tandis que le tableau 2 donne l'efficacité relative de la stratégie 1 par rapport à celle de la stratégie 2 avec les populations décrites dans le tableau 1. Le tableau 2 compare également la performance de l'estimateur H-T utilisé de concert avec le plan d'échantillonnage de Sampford (1967), ce qu'on désigne ici par la stratégie 3, et la performance de la stratégie 2.

Il ressort du tableau 2 que la performance de la stratégie proposée (stratégie 1) se compare avantageusement à celle de la stratégie 3 dans le cas de la plupart des populations étudiées. Bien entendu, les deux stratégies sont toutes deux supérieures à la stratégie 2.

Pour obtenir l'efficacité relative de la stratégie 1, les unités ont été groupées en ne veillant uniquement qu'à ce que la condition (2.1) soit satisfaite. On a également tenté d'utiliser la méthode de Rao et Lanke (1984) pour former les groupes. Cette méthode n'a toutefois pas conduit à une grande efficacité dans tous les cas. D'autres recherches sont nécessaires pour décider du «meilleur» choix des groupes. Pour certaines populations, il a été impossible de former des groupes satisfaisant à la condition (2.1) avec des valeurs élevées de m; c'est pour cette raison que l'efficacité relative dans ces cas-là n'apparaît pas dans le tableau 2.

En conclusion, il semble indiquer de faire un bref commentaire au sujet des cas où la taille de l'échantillon voulue, n, est *impaire*. On peut tirer un échantillon avec PSPT pour n impair en sélectionnant $(n + 1)$ unités à l'aide de la méthode proposée, puis en rejetant au hasard une unité. Les expressions pour les π_i et π_{ij} sont également assez simples dans ce cas-là. De

toute évidence, quand une des $(n + 1)$ unités de l'échantillon est rejetée au hasard, l'échantillon définitif comprend deux unités de chacun de $(n - 1)/2$ groupes et seulement une unité d'un des groupes. Un estimateur non biaisé et positif de $Var(\hat{Y})$, semblable à celui de l'expression (3.1), peut ensuite être calculé à partir des $(n - 1)/2$ groupes qui contiennent deux unités de l'échantillon.

REMERCIEMENTS

Les auteurs remercient l'arbitre pour ses conseils pertinents concernant la première version.

BIBLIOGRAPHIE

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

BREWER, K.R.W. et HANIF, M. (1983). *Sampling with Unequal Probabilities*. Lecture Notes in Statistics, No. 15, New York: Springer-Verlag.

COCHRAN, W.G. (1963). *Sampling Techniques*, (2e éd.). New York: John Wiley.

DES RAJ (1965). Variance estimation in randomized systematic sampling with probability proportional to size, *Journal of the American Statistical Association*, 60, 278-284.

DES RAJ (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.

DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Journal of the Royal Statistical Society*, Sér. C, 16, 152-164.

HANURAV, T. (1967). Optimum utilization of auxiliary information: πps sampling of two units from a stratum. *Journal of the Royal Statistical Society*, Sér. B, 29, 379-391.

HORVITZ, D.G. et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

MIDZUNO, H. (1951). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 2, 99-108.

RAO, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 58, 202-215.

RAO, J.N.K. et BAYLESS, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units by stratum. *Journal of the American Statistical Association*, 64, 540-559.

RAO, J.N.K. et LANKE, J. (1984). Simplified unbiased variance estimation for multistage designs. *Biometrika*, 71, 387-395.

SAMPFORD, M.R. (1962). *An Introduction to Sampling Theory*. Édimbourg: Oliver and Boyd.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

SUKHATME, P.V. et SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*, (2e éd.). Ames, Iowa: Iowa State University Press.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3e éd.). Londres: Griffin.

YATES, F. et GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Sér. B, 15, 253-261.

Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 101-117
Statistique Canada

Plan d'échantillonnage pour l'enquête sur la santé et les limitations d'activités

D. DOLSON, K. McCLEAN, J.-P. MORIN, et A. THÉBERGE[1]

RÉSUMÉ

L'Enquête sur la santé et les limitations d'activités s'inscrit dans le cadre du programme visant à établir une base de données relative à la population ayant une incapacité ou un handicap au Canada. On décrit le plan d'échantillonnage utilisé pour la partie de l'enquête couvrant la population vivant hors des institutions. Les méthodes employées pour déterminer les tailles des échantillons et pour la sélection de ceux-ci sont aussi exposées.

MOTS CLÉS: Incapacité; échantillonnage stratifié; échantillonnage à deux degrés; allocation optimale; échantillonnage sans remise.

1. INTRODUCTION

Dans le cadre du programme visant à obtenir plus d'information sur la population des personnes handicapées au Canada, l'Enquête sur la santé et les limitations d'activités (ESLA) a été réalisée au cours de l'automne 1986. Elle vise à recueillir des informations sur la nature des problèmes éprouvés par cette population et, d'une facon générale, sur leurs activités quotidiennes (à la maison, au travail, à l'école, dans les déplacements, etc.). L'enquête se divise en deux parties: l'une couvre la population vivant en institution et l'autre, qui fait l'objet du présent article, couvre la population hors des institutions.

On a partitionné le Canada en 238 régions infra-provinciales (RIP). Ces régions comprennent toutes les municipalités du Québec et de l'Ontario qui regroupent plus de 125,000 habitants et toutes celles des autres provinces de plus de 75,000 habitants. Les autres régions sont composées d'agglomérations de sous-divisions de recensement respectant la contiguïté géographique et les frontières entre les provinces. Le nombre de ces régions dans chaque province est proportionnel à la racine carrée de la population moins les municipalités précédemment définies. Un des principaux objectifs de l'enquête est de produire au niveau des RIP des statistiques sur la population handicapée afin de permettre une analyse détaillée des divers besoins. En outre, on fournira des estimations pour trois groupes d'âge: enfants (moins de 15 ans), adultes (de 15 à 64 ans) et personnes âgées (65 ans et plus).

La cueillette des informations s'est effectuée en deux étapes. La première consiste en une question à plusieurs volets, incluse dans le formulaire 2B du recensement de la population du Canada de 1986, qui porte sur les limitations du répondant dans divers secteurs d'activités ainsi que sur sa propre évaluation de sa condition. Une copie de cette question (question n° 20 du formulaire du recensement de 1986) est donnée en appendice. La seconde étape a eu lieu quelque temps après le recensement. Elle est constituée d'un questionnaire de sélection et d'un suivi qui recueille des informations sur les problèmes et les activités des répondants handicapés.

La première étape vise essentiellement à répartir les répondants en deux groupes: ceux qui ont répondu "oui" à au moins une des parties de la question 20 et ceux qui ont répondu "non" à toutes les parties. Le but est d'identifier à l'avance une grande partie de la population potentiellement handicapée afin de concentrer le maximum des ressources de l'enquête sur le groupe cible. Notons toutefois que les enquêtes précédentes montrent que le groupe cible ne sera pas complètement identifié par cette question. (Voir Dolson et coll. (1984) et Dolson et coll. (1986)).

[1] D. Dolson, K. McClean, J.-P. Morin, et A. Théberge, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6.

La seconde étape est l'ESLA. On procède à des interviews individuelles pour la strate des "oui" et à des interviews téléphoniques pour la strate des "non". Du point de vue opérationnel, l'interview est formée de deux parties: le questionnaire de sélection et le suivi.

Le rôle du questionnaire de sélection est de déterminer pour quel répondant le questionnaire de suivi est pertinent. Le questionnaire destiné aux adultes comprend les 17 activités de la vie quotidiennes (AVQ) utilisées lors de l'Enquête sur la santé et les incapacités au Canada qui a eu lieu en 1983 et 1984 ainsi que la partie (a) de la question 20 du recensement et quelques questions sur les maladies et les handicaps mentaux (voir l'appendice B). Si une réponse positive est obtenue à au moins une de ces questions, l'interviewer procède au suivi, sinon l'interview prend fin. La partie (a) de la question du recensement est posée de nouveau pour tenir compte d'un éventuel changement de situation causé soit par une réponse d'un intermédiaire lors du recensement, soit par une réévaluation du répondant de son propre état.

La section de sélection sur le questionnaire destiné aux enfants comprend des questions sur les appareils spéciaux, les limitations aux activités, la fréquentation d'une école spéciale et les affections ou problèmes de santé. Une réponse positive à au moins une de ces questions entraîne l'interview de suivi. La question du recensement n'est pas posée de nouveau parce que tous les interviews relatives aux enfants nécessitent un intermédiaire et que la question sur les limitations d'activités est équivalente à la partie (a) de la question 20 du recensement.

La deuxième section du présent article décrit comment la population canadienne a été subdivisée en différentes sous-populations pour fins d'estimation. La section trois est consacrée au plan de sondage de l'ESLA. La section suivante traite du fichier des données géographiques et démographiques projetées pour 1986 qui a servi à former les bases de sondage. Enfin la dernière section décrit comment on a procédé à l'échantillonnage.

2. POPULATIONS COUVERTES

Les personnes qui résident en permanence dans les hôpitaux généraux, les hôpitaux psychiatriques, les centres de soins spéciaux ou les établissements pour personnes âgées et malades chroniques, les centres de traitement ou établissements pour handicapés physiques et les orphelinats ou foyers pour enfants font l'objet d'une partie spéciale de l'enquête: l'ESLA-institutions. Dans le présent article, il est question de la partie de l'enquête qui couvre la population canadienne à l'exclusion des personnes couvertes par l'ESLA-institutions, des personnes qui résident dans les prisons, les camps militaires, les foyers pour les jeunes délinquants, les navires de guerre, les établissements pénitentiaires, les établissements de correction et des résidents des logements collectifs de la catégorie "autres" (par exemple, cirques, communes non-religieuses).

Chaque secteur de dénombrement (SD) dont la population n'est pas totalement exclue de l'enquête est classé dans une des cinq bases de sondage suivantes:

1. Réserves indiennes où le dénombrement avait été fait au moyen d'interviews en 1981;
2. Autres réserves indiennes;
3. SD où le dénombrement se fait habituellement par interviews;
4. SD dans les RIP de Whitehorse, Yellowknife, Pine Point, Hay River et Fort Smith;
5. Tous les autres SD.

L'ordre de priorité pour l'appartenance à une base est 1-2-4-3-5. Ainsi, un SD formé par une réserve indienne situé dans la RIP de Whitehorse est classé comme étant une réserve indienne.

Chaque SD est divisé en deux: le SD "oui", formé des personnes qui donneraient une réponse positive à la question du recensement, et le SD "non" formé des personnes qui y répondraient négativement. Un plan d'échantillonnage différent est utilisé pour chacune des cinq bases de sondage: dans la première base, tous les SD "oui" sont sélectionnés et aucun des SD "non"; dans la base deux, tous les SD "oui" et un échantillon des SD "non" sont sélectionnés; dans la troisième base, aucun des SD "non" n'est choisi et un échantillon des SD "oui" sont sélectionnés; tous les SD de la base quatre sont sélectionnés; et enfin un échantillon des SD "oui" et un échantillon des SD "non" sont choisis dans la base cinq.

3. PLAN DE SONDAGE

La méthode d'échantillonnage dont il est question dans cette section a été utilisée pour les bases de sondage trois et cinq. Pour des raisons d'espace, le plan de sondage relatif à la deuxième base de sondage n'est pas exposé dans le présent article. (Pour plus de renseignements sur la méthodologie de l'ESLA, voir Dolson et coll. (1986)).

3.1 Plan d'échantillonnage

Chaque province est partitionnée en régions infra-provinciales (RIP) lesquelles sont elles-mêmes partitionnées en secteurs de dénombrement (SD).

Les habitants de chaque SD sont divisées en un SD "oui" et un SD "non", selon qu'ils répondraient positivement ou négativement à la question 20 du formulaire du recensement. Dans chaque RIP, on stratifie les SD "oui" en grands et petits SD suivant le critère expliqué dans la quatrième section du présent article. Les personnes appartenant à un SD "oui" sont associés à une strate et à une RIP alors que les personnes appartenant à un SD "non" ne sont associés qu'à leur SD. Dans chaque province, la population est subdivisée en trois groupes d'âges: enfants (moins de 15 ans), adultes (15 à 64 ans) et personnes âgées (65 ans et plus).

La méthode d'échantillonnage est un plan de sondage stratifié à deux degrés pour les SD "oui" de chaque RIP et un plan de sondage à deux degrés pour les SD "non" de la province. Les SD constituaient les unités primaires et les répondants les unités secondaires.

Toutes les personnes ayant répondu au formulaire 2B du recensement dans un SD "oui" sélectionné dans l'échantillon sont interviewées et un tiers d'entre elles le sont dans les SD "non" sélectionnés dans l'échantillon.

3.2 Allocation de l'échantillon

Ce plan de sondage doit permettre de minimiser les coûts d'échantillonnage étant donné un coefficient de variation maximum des estimations et une variance fixée de l'estimateur $\hat{B}$ du biais relatif B. On définit B comme étant le rapport du nombre T_0 de personnes "non" qui ont une caractéristique d'intérêt dans la province, sur le nombre T_1 de personnes "oui" qui ont une caractéristique d'intérêt dans la province. Par personne "non" il faut entendre une personne qui répondrait négativement à tous les volets de la question 20 du formulaire du recensement, et par personne "oui", une personne qui répondrait positivement à au moins un des volets.

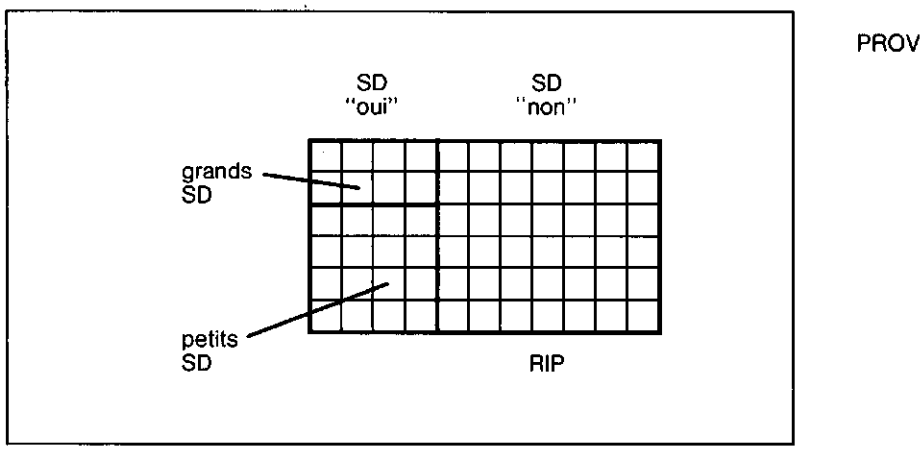

Figure 1. Illustration du plan de sondage.

Notons N_0 le nombre de SD "non" dans la province, N_{jk}, le nombre de SD "oui" dans la strate j et la RIP k de la province, n_0 et n_{jk} les tailles d'échantillons correspondantes et c_0 et c_{jk} les coûts unitaires d'échantillonnage correspondants. Si on a N_p RIP dans la province, on veut donc minimiser

$$\sum_{k=1}^{N_p} (c_{1k} n_{1k} + c_{2k} n_{2k}) + c_0 n_0$$

étant donné

$$CV^2(y_k) \leq CV_*^2; \ \mathrm{Var}(\hat{B}) = \mathrm{Var}_*(\hat{B});$$

$$n_{jk} \leq N_{jk}; \ n_{2k} = \lambda_k n_{1k}; \ n_0 \leq N_0$$

$$(j = 1, 2; \ k = 1, \ldots, N_p)$$

où λ_k est le rapport du nombre attendu de personnes handicapées dans les petits SD sur le nombre attendu de personnes handicapées dans les grand SD de la RIP k, y_k est le nombre estimé de personnes "oui" qui ont une caractéristique d'intérêt dans la RIP k et les valeurs notées avec des * sont des constantes.

Si la fraction de sondage dans les SD "oui" est f_1, M_{ijk} est le nombre de personnes "oui" dans le SD i de la strate j de la RIP k de la province, et p_{ijk} est la probabilité d'une caractéristique d'intérêt pour une personne "oui" du SD i de la strate j de la RIP k, alors on a

$$E(y_k) = Y_k = \sum_{j=1}^{2} \sum_{i=1}^{N_{jk}} M_{ijk} p_{ijk},$$

$$\mathrm{Var}\ (y_k) = \sum_{j=1}^{2} \left\{ \frac{N_{jk}^2}{n_{jk}} \left(1 - \frac{n_{jk}}{N_{jk}}\right) S_{jk}^2 + \frac{N_{jk}}{n_{jk}} \left(\frac{1 - f_1}{f_1}\right) \sum_{i=1}^{N_{jk}} M_{ijk} S_{ijk}^2 \right\},$$

où

$$y_k = \sum_{j=1}^{2} \sum_{i=1}^{N_{jk}} \frac{N_{jk}}{n_{jk}} M_{ijk} p_{ijk},$$

$$S_{jk}^2 = \frac{1}{N_{jk} - 1} \sum_{i=1}^{N_{jk}} \left(M_{ijk} p_{ijk} - \left(\sum_{i=1}^{N_{jk}} \frac{M_{ijk} p_{ijk}}{N_{jk}} \right) \right)^2,$$

$$S_{ijk}^2 = \frac{M_{ijk}}{M_{ijk} - 1} p_{ijk} \left(1 - p_{ijk}\right).$$

Après quelques opérations algébriques, on obtient

$$\mathrm{Var}\ (y_k) = \frac{1}{n_{1k}} \left\{ N_{1k}^2 S_{1k}^2 + \left(\frac{1 - f_1}{f_1}\right) N_{1k} \sum_{i=1}^{N_{1k}} M_{i1k} S_{i1k}^2 + \frac{N_{2k}^2 S_{2k}^2}{\lambda_k} \right.$$

$$\left. + \left(\frac{1 - f_1}{f_1}\right) \frac{N_{2k}}{\lambda_k} \sum_{i=1}^{N_{2k}} M_{i2k} S_{i2k}^2 \right\} - \sum_{j=1}^{2} N_{jk} S_{jk}^2.$$

On peut donc représenter $CV^2(y_k)$ sous la forme suivante:

$$CV^2(y_k) = \frac{\text{Var}(y_k)}{Y_k^2} = \frac{A_k}{n_{1k}} - B_k. \tag{3.1}$$

Par ailleurs, B, le biais relatif, et $\hat{B}$, son estimateur, sont donnés par

$$B = \frac{T_0}{T_1} = \frac{\sum_{i=1}^{N_0} M_{i0}p_{i0}}{\sum_{k=1}^{N_p} Y_k},$$

$$\hat{B} = \frac{t_0}{t_1} = \frac{\sum_{i=1}^{n_0} \frac{N_0}{n_0} M_{i0}p_{i0}}{\sum_{k=1}^{N_p} y_k},$$

où M_{i0} est le nombre de personnes "non" dans le SD i de la province, et p_{i0} est la probabilité d'une caractéristique d'intérêt pour une personne "non" du SD i.

En supposant que t_0 et t_1 sont indépendants, on a

$$\text{Var}(\hat{B}) = B^2\left(\frac{\text{Var}(t_0)}{T_0^2} + \frac{\text{Var}(t_1)}{T_1^2}\right). \tag{3.2}$$

Après quelques opérations algébriques, si f_0 est la fraction de sondage dans les SD "non", on obtient

$$\text{Var}(t_0) = \frac{1}{n_0}\left\{N_0^2S^2 + N_0\left(\frac{1-f_0}{f_0}\right)\sum_{i=1}^{N_0} M_{i0}S_i^2\right\} - N_0S^2$$

où

$$S^2 = \frac{1}{N_0 - 1}\sum_{i=1}^{N_0}\left(M_{i0}p_{i0} - \left(\sum_{i=1}^{N_0}\frac{M_{i0}p_{i0}}{N_0}\right)\right)^2, \quad S_i^2 = \frac{M_{i0}}{M_{i0}-1}p_{i0}(1 - p_{i0})$$

ce qui peut être représenté sous la forme suivante:

$$\text{Var}(t_0) = \frac{A_0}{n_0} - B_0. \tag{3.3}$$

D'autre part, en supposant les y_k indépendants, on a

$$\text{Var}(t_1) = \sum_{k=1}^{N_p}\text{Var}(y_k).$$

En considérant l'équation (3.1), on a

$$\mathrm{Var}\,(t_1) = \sum_{k=1}^{N_p} \left(\frac{A_k Y_k^2}{n_{1k}} - B_k Y_k^2\right). \tag{3.4}$$

De (3.2), (3.3) et (3.4), on obtient

$$\begin{aligned}
\mathrm{Var}\,(\hat{B}) &= B^2\left\{\left(\frac{A_0}{n_0 T_0^2} - \frac{B_0}{T_0^2}\right) + \sum_{k=1}^{N_p}\left(\frac{A_k Y_k^2}{n_{1k} T_1^2} - \frac{B_k Y_k^2}{T_1^2}\right)\right\} \\
&= \frac{1}{n_0}\left(\frac{B^2 A_0}{T_0^2}\right) + \sum_{k=1}^{N_p} \frac{1}{n_{1k}}\left(\frac{B^2 A_k Y_k^2}{T_1^2}\right) - B^2\left(\frac{B_0}{T_0^2} + \sum_{k=1}^{N_p} \frac{B_k Y_k^2}{T_1^2}\right) \\
&= \frac{X}{n_0} + \sum_{k=1}^{N_p} \frac{W_k}{n_{1k}} - Z.
\end{aligned}$$

On peut réexprimer le problème d'optimisation comme étant celui de minimiser

$$\sum_{k=0}^{N_p} c_k n_k$$

en tenant compte des contraintes

$$0 < a_k \leq n_k \leq b_k \quad (k = 0, 1, 2, \ldots, N_p) \tag{3.5}$$

et

$$\sum_{k=0}^{N_p} d_k / n_k = e \tag{3.6}$$

où, pour $k = 1, 2, \ldots, N_p$,

$$n_k = n_{1k},\ c_k = c_{1k} + c_{2k}\lambda_k,\ a_k = \frac{A_k}{CV_*^2 + B_k},\ b_k = \min\,(N_{1k}, N_{2k}/\lambda_k).$$

En pratique, plutôt que de prendre $b_k = \min\,(N_{1k}, N_{2k}/\lambda_k)$ on prend

$$b_k = \frac{N_{1k} N_{2k}(1 + \lambda_k)}{\lambda_k^2 N_{1k} + N_{2k}},$$

puis si $n_k > N_{1k}$, alors on prend $n_{1k} = N_{1k}$ et

$$n_{2k} = \lambda_k n_k + \frac{(n_k - N_{1k}) N_{2k}}{N_{1k}\lambda_k},$$

tandis que si $\lambda_k n_k > N_{2k}$ alors on prend $n_{2k} = N_{2k}$ et

$$n_{1k} = n_k + \frac{(\lambda_k n_k - N_{2k}) N_{1k} \lambda_k}{N_{2k}}.$$

On considère donc que $N_{2k} / (N_{1k} \lambda_k)$ petits SD équivalent à un grand SD. Il y a, en moyenne, autant de personnes handicapées dans un grand SD que dans $N_{2k} / (N_{1k} \lambda_k)$ petits.

En procédant ainsi, on ne respecte pas toujours la relation $n_{2k} = \lambda_k n_{1k}$. Toutefois, on évite de se contenter d'un CV supérieur à celui visé, lorsque, par exemple, il reste des petits SD à observer (même si tous les grands SD ont été sélectionnés).

Il se peut que pour certaines valeurs de k, on ait $a_k \geq b_k$. Dans un tel cas, on prend $n_k = b_k$. Soient

$$E_1 = \{k = 0, 1, 2, \ldots, N_p \mid n_k = a_k\},$$

$$E_2 = \{k = 0, 1, 2, \ldots, N_p \mid n_k = b_k > a_k\},$$

$$E_3 = \{k = 0, 1, 2, \ldots, N_p \mid a_k < n_k < b_k\},$$

$$E_4 = \{k = 0, 1, 2, \ldots, N_p \mid n_k = b_k \leq a_k\},$$

la solution existe si

$$\sum_{k=0}^{N_p} d_k / b_k \leq e,$$

et elle est de la forme

$$n_k = \begin{cases} a_k & (k \in E_1) \\ b_k & (k \in E_2 \cup E_4) \\ K (d_k / c_k)^{1/2} & (k \in E_3) \end{cases} \tag{3.7}$$

où

$$K = \frac{\sum_{k \in E_3} (d_k / c_k)^{1/2}}{e - \sum_{k \in E_1} d_k / a_k - \sum_{k \in E_2 \cup E_4} d_k / b_k}, \tag{3.8}$$

puisque les n_k $(k \in E_3)$ minimisent $\sum_{k \in E_3} c_k n_k$, sous la contrainte

$$\sum_{k \in E_3} d_k / n_k = e - \sum_{k \in E_1} d_k / a_k - \sum_{k \in E_2 \cup E_4} d_k / b_k.$$

Quels sont les ensembles E_1, E_2, E_3 et E_4 qui correspondent à la solution? L'ensemble E_4 est facile à déterminer. Par ailleurs, on doit avoir

$$a_k < (d_k/c_k)^{1/2}K < b_k \ (k \epsilon E_3), \quad (d_k/c_k)^{1/2}K \geq b_k \ (k \epsilon E_2),$$

$$(d_k/c_k)^{1/2}K \leq a_k \ (k \epsilon E_1). \tag{3.9}$$

Il s'agirait donc d'essayer chacune des possibilités pour E_1, E_2 et E_3 jusqu'à ce qu'on obtienne une valeur de k qui satisfasse (3.9). Pour réduire le nombre de possibilités à examiner, remarquons que si pour $k' \geq k$,

$$b'_k (c'_k/d'_k)^{1/2} \geq b_k (c_k/d_k)^{1/2} \quad (k, k' \epsilon \{0, 1, \ldots, N_p\}), \tag{3.10}$$

alors il existe k^* tel que $E_2 = \{0, 1, 2, \ldots, k^*\}$, ou bien $E_2 = \{\ \}$. Tandis que si pour $k' \geq k$,

$$a'_k (c'_k/d'_k)^{1/2} \geq a_k (c_k/d_k)^{1/2} \quad (k, k' \epsilon \{0, 1, \ldots, N_p\}), \tag{3.11}$$

alors il existe k^{**} tel que $E_1 = \{k^{**}, k^{**} + 1, \ldots N_p\}$ ou bien $E_1 = \{\ \}$.

3.3 Estimation des paramètres

Afin d'effectuer le calcul de l'allocation optimale de l'échantillon, les quantités suivantes doivent être évaluées:

P_1 = proportion des individus ayant été sélectionnés lors de l'ESLA et qui ont répondu "oui" à la question 20 du recensement;

P_2 = proportion des individus n'ayant pas été sélectionnés lors de l'ESLA et qui ont répondu "oui" à la question 20 du recensement;

P_3 = proportion des individus ayant été sélectionnés lors de l'ESLA et qui ont répondu "non" à la question 20 du recensement.

Puisque ces paramètres ne peuvent être calculés directement à partir des données de l'Enquête sur la santé et les incapacités au Canada, un test qu'on a appelé "l'étude de calibration" a été effectué en septembre et octobre 1985.

La question 20 du recensement fut posée intégralement comme question supplémentaire lors de l'Enquête sur la population active (EPA) de septembre. Elle s'adressait à un échantillon d'environ 36,000 individus. Les questions sur les 17 AVQ et une question sur les handicaps mentaux constituaient un supplément à l'EPA d'octobre et elles étaient demandées aux mêmes individus.

Pour chacun des groupes d'âges de cinq ans, on s'est servi des valeurs pondérées de l'étude de calibration afin d'estimer la probabilité d'une réponse positive P (*oui*) à la question 20 du recensement. Le questionnaire de sélection de l'ESLA diffère de celui de l'étude de calibration. Dans l'ESLA, il y a plus de questions sur les troubles mentaux ou psychologiques et la partie (a) de la question 20 du recensement est posée de nouveau. Donc, on ne s'est pas appuyé seulement sur l'étude de calibration pour calculer les paramètres.

4. FICHIER GÉOGRAPHIQUE ET DÉMOGRAPHIQUE POUR 1986

4.1 Description des informations disponibles

Lors de l'allocation de l'échantillon qui a eu lieu au printemps de 1986, on disposait des données suivantes pour établir les projections de la population par groupe d'âge et par SD:

1. projection de la population par groupe d'âge et par province en 1986;
2. estimé de la population par groupe d'âge et par DR en 1984;
3. population par groupe d'âge et par SD en 1981;
4. un fichier de conversion pour établir la correspondance entre les SD de 1981 et ceux de 1986;
5. estimé du nombre de logements par SD en 1986.

La facon dont est construit le fichier de conversion repose sur le concept d'ensembles correspondants. Chaque ensemble correspondant est la plus petite région formée de SD, dont les frontières sont demeurées inchangées. Par exemple, si trois SD de 1981 ont servi à former deux SD de 1986, ce groupe de trois SD de 1981 (ou deux SD de 1986) forme un ensemble correspondant.

Les quatre méthodes décrites dans la sous-section qui suit visent à établir des projections de la population par groupe d'âge et par ensemble correspondant en 1986. Si un ensemble correspondant est formé de plusieurs SD de 1986, la population projetée pour cet ensemble correspondant peut être répartie entre ces SD de facon proportionnelle en utilisant l'estimé du nombre de logements par SD en 1986.

4.2 Méthodes d'estimation

Pour la province p, notons

$EC_{l,k}$ = le l-ième ensemble correspondant de la k-ième DR ($l = 1, 2, \ldots, N_k$; $k = 1, 2, \ldots, N_p$),

$EC_{l,k;81}(j)$ = population de $EC_{l,k}$ dans le j-ième groupe d'âge en 1981 ($j = 1, 2, \ldots, 16$),

$DR_{k;84}(j)$ = estimé de la population de la k-ième DR dans le j-ième groupe d'âge en 1984,

$\hat{P}_{86}(j)$ = projection de la population dans le j-ième groupe d'âge dans la province en 1986.

Pour les trois méthodes qui suivent, la première étape consiste à calculer $\widehat{DR}_{k;86}(j)$, c'est-à-dire la projection de la population de la k-ième DR dans le j-ième groupe d'âge en 1986. On suppose qu'il existe des K_j' ($j = 1, 2, \ldots, 16$) tels que

$$\widehat{DR}_{k;86}(j) = K_j'(\widehat{DR}_{k;84}(j)) \quad (k = 1, 2, \ldots, N_p; j = 1, 2, \ldots, 16),$$

$$\sum_{k=1}^{N_p} \widehat{DR}_{k;86}(j) = \hat{P}_{86}(j) \quad (j = 1, 2, \ldots, 16).$$

Cela implique que

$$\widehat{DR}_{k;86} = \frac{\hat{P}_{86}(j)\, DR_{k;84}(j)}{\sum_{k=1}^{N_p} DR_{k;84}(j)}.$$

La première méthode pour estimer $EC_{l,k;86}(j)$ consiste à supposer l'existence de K_j ($j = 1, \ldots, 16$) tels que

$$\widehat{EC}_{l,k;86}(j) = K_j EC_{l,k;81}(j) \quad (l = 1, \ldots, N_k; j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} \widehat{EC}_{l,k;86}(j) = \widehat{DR}_{k;86}(j) \quad (j = 1, 2, \ldots, 16).$$

On dira que cette méthode utilise le modèle simple. On obtiendra

$$\widehat{EC}_{l,k;86}(j) = \frac{\widehat{DR}_{k;86}(j)\, EC_{l,k;81}(j)}{\sum_{l=1}^{N_k} EC_{l,k;81}(j)} \quad (l = 1, \ldots, N_k; j = 1, \ldots, 16).$$

Avec ce modèle simple, l'estimation de la population totale de $EC_{l,k}$ en 1986 est

$$\sum_{j=1}^{16} \frac{\widehat{DR}_{k;86}(j)\, EC_{l,k;81}(j)}{\sum_{l=1}^{N_k} EC_{l,k;81}(j)}.$$

Si on croit être en mesure de fournir une meilleure estimation, $\widehat{EC}_{l,k;86}(tot)$ de cette quantité par des moyens indépendants (par exemple, à partir de l'estimé du nombre de logements dans $EC_{l,k}$ en 1986), alors on peut utiliser des modèles plus élaborés pour estimer $EC_{l,k;86}(j)$. Le modèle multiplicatif est précisé par les équations suivantes:

$$\widehat{EC}_{l,k;86}(j) = K_j(EC_{l,k;81}(j)) + e_l' \quad (l = 1, \ldots, N_k; j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} \widehat{EC}_{l,k;86}(j) = K(\widehat{DR}_{k;86}(j)) \quad (j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} e_l' = 0,$$

$$\sum_{j=1}^{16} \widehat{EC}_{l,k;86}(j) = \widehat{EC}_{l,k;86}(tot) \quad (l = 1, \ldots, N_k).$$

On peut interpréter e_l comme étant la migration intra-DR nette du l-ième ensemble correspondant.

Le troisième modèle utilisé, appelé modèle additif, est donné par les équations suivantes:

$$\hat{EC}_{l,k;86}(j) = EC_{l,k;81}(j) + e_l + f_j \quad (l = 1, \ldots, N_k; j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} \hat{EC}_{l,k;86}(j) = \hat{DR}_{k;86}(j) + D \quad (j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} e_l = D,$$

$$\sum_{j=1}^{16} \hat{EC}_{l,k;86}(j) = \hat{EC}_{l,k;86}(tot) \quad (l = 1, \ldots, N_k).$$

On suppose donc que les accroissements (ou diminutions) de population pour chaque groupe d'âge des ensembles correspondants de la DR peuvent se décomposer en deux termes: un qui ne dépend que de l'ensemble correspondant et non de l'âge (e_l), et un qui ne dépend que de l'âge et pas de l'ensemble correspondant (f_j).

Un dernier modèle, trivial celui-là, serait simplement de poser

$$\hat{EC}_{l,k;86}(j) = EC_{l,k;81}(j) \quad (l = 1, 2, \ldots, N_k; j = 1, \ldots, 16).$$

4.3 Évaluation des méthodes d'estimation

Les quatre méthodes ont été évaluées en utilisant des données pour la période allant de 1976 à 1981. On a utilisé la projection établie en 1976 de la population par groupe d'âge et par province en 1981 ($\hat{P}_{81}(j)$), la population par groupe d'âge et par SD en 1976, un fichier de conversion 1976-1981 et l'estimé pré- censitaire du nombre de logements par SD en 1981. Des estimations pour la population par groupe d'âge et par DR en 1979 (l'équivalent de $DR_{k;84}(j)$) n'existant pas, on a posé

$$\hat{DR}_{k;81} = \frac{\hat{P}_{81}(j)\,\hat{DR}_{k;84}(j)}{\sum_{k=1}^{N_p} \hat{DR}_{k;84}(j)}.$$

Pour $\hat{EC}_{l,k;81}(tot)$, nécessaire pour les modèles multiplicatif et additif, on a utilisé

$$\hat{EC}_{l,k;81}(tot) = \frac{\sum_{j=1}^{16} EC_{l,k;76}(j) \quad \sum_{j=1}^{16} \hat{DR}_{k;81}(j)}{\sum_{l=1}^{N_k} \sum_{j=1}^{16} EC_{l,k;76}(j)}.$$

Tableau 1
Comparaison des quatres modèles

Prov.	EFF_S	EFF_M	EFF_A
T.-N.	0.890	0.891	0.887
Î.P.-É.	0.903	0.914	0.919
N.-É.	0.960	0.972	0.912
N.-B.	0.869	0.868	0.884
Qué.	0.778	0.764	0.818
Ont.	0.932	0.930	0.916
Man.	0.892	0.904	0.912
Sask.	0.732	0.749	0.801
Alb.	0.818	0.827	0.860
C.-B.	0.713	0.716	0.775
Yukon	0.770	0.768	0.840
T.N.-O.	1.252	1.246	1.157

On a ensuite calculé pour chaque province p, une mesure d'efficacité des modèles simple, multiplicatif et additif par rapport au modèle trivial:

$$EFF_m = \frac{\sum_{k=1}^{N_p} \sum_{l=1}^{N_k} \sum_{j=1}^{16} \left((\widehat{EC}^{(m)}_{l,k;81}(j) - EC_{l,k;81}(j) \right)^2}{\sum_{k=1}^{N_p} \sum_{l=1}^{N_k} \sum_{j=1}^{16} \left((\widehat{EC}^{(T)}_{l,k;81}(j) - EC_{l,k;81}(j) \right)^2} \quad (m = S, M, A),$$

où $\widehat{EC}^{(m)}_{l,k;81}(j)$ avec $m = S, M, A$ et T sont les projections obtenues à l'aide des modèles simple, multiplicatif, additif et trivial respectivement. Le tableau 1 présente des valeurs obtenues.

Le modèle simple donne les pires résultats pour une province et une territoire, le modèle multiplicatif pour deux provinces, et le modèle additif pour sept provinces et un territoire.

Le modèle simple se révèle le meilleur pour cinq provinces,le modèle multiplicatif pour deux provinces et un territoire, et le modèle additif pour trois provinces et un territoire.

Le modèle simple ayant aussi, comme son nom l'indique, l'avantage d'être simple, c'est le modèle qui a été retenu.

4.4 Méthode de stratification selon la taille des secteurs de dénombrement

Si on faisait un tirage aléatoire simple pour sélectionner les SD à l'intérieur de chaque région infra-provinciale (RIP), alors les personnes handicapées faisant partie d'un SD avec de nombreuses personnes handicapées auraient moins de chance d'être sélectionnées que celles faisant partie d'un petit SD, c'est-à-dire un SD avec peu de personnes handicapées. Pour éviter de trop grandes différences au niveau des probabilités de sélection, dans chaque RIP on stratifie la population des SD selon le nombre de personnes handicapées qui s'y trouve, puis on utilise l'allocation proportionnelle, c'est-à-dire que dans chaque strate, le nombre de SD choisis est proportionnel au nombre de personnes handicapées.

À partir des résultats d'enquêtes précédentes, on aétabli un lien entre la pyramide des âges de la population d'un SD et le nombre de personnes handicapées qu'on peut s'attendre à trouver dans ce SD. Le nombre de personnes handicapées étant inconnu, la variable utilisée pour la stratification et l'allocation de l'échantillon est le nombre attendu de personnes handicapées.

Dans le cas présent, on n'a que deux strates: celle des grands SD et celle des petits SD. Puisqu'on utilise l'allocation proportionnelle, pour déterminer d'une manière optimale la taille frontière au-delà de laquelle un SD est grand et en deçà de laquelle il est petit, on s'est servi d'un critère qu'on peut trouver dans Raj (1968). La taille frontière doit être égale à la moyenne de la taille moyenne des petits SD et de la taille moyenne des grands SD.

5. SÉLECTION DE L'ÉCHANTILLON

Il fallait choisir des échantillons pour les trois populations (enfants, adultes et personnes âgées) parmi les grands et les petits SD "oui" de chaque RIP, tant pour la base trois que pour la base cinq, et parmi les SD "non" de chaque province pour la base cinq. Lorsque, dans une RIP, il y avait moins de deux grands SD ou moins de deux petits SD, on sélectionnait tous les SD de cette RIP pour les trois populations. Les échantillons "oui" et "non" étaient créés indépendamment en utilisant l'algorithme à un passage décrit par Bebbington (1975). Les échantillons des trois populations, pour la composante "oui" et pour la composante "non", étaient emboîtés afin de minimiser le nombre total de SD choisis.

Au tableau suivant figurent les tailles des échantillons par province pour chaque groupe d'âge.

Tableau 2
Tailles d'échantillons par province et par groupe d'âge

Province	Enfants		Adultes		Personnes âgées	
	Nombre de SD "oui" sélectionnés	Nombre de SD "non" sélectionnés	Nombre de SD "oui" sélectionnés	Nombre de SD "non" sélectionnés	Nombre de SD "oui" sélectionnés	Nombre de SD "non" sélectionnés
T.-N.	880	136	405	154	476	173
Î.P.-É.	242	242	111	217	82	166
N.-É.	1257	157	434	130	438	115
N.B.	1142	162	459	146	453	138
Qué.	4749	153	1070	114	1488	133
Ont.	6085	158	1304	116	1542	120
Man.	1082	203	457	169	367	144
Sask.	2291	265	942	241	921	193
Alb.	2762	190	909	176	1389	222
C.-B.	3117	170	752	125	948	119

6. DISCUSSION

L'enquête post-censitaire est une méthode de sondage relativement nouvelle qui est appelée à connaître de nombreux développements au cours des prochaines années. Ce type d'enquête permet beaucoup de flexibilité dans la collecte des données et l'utilisation d'échantillons de grande taille dispersés partout à travers le pays avec des coûts et des délais convenables. L'Enquête sur la santé et les limitations d'activités constitue une première expérience dans ce domaine au Canada pour une enquête de cette importance.

Le plan d'échantillonnage exposé dans le présent article représente une tentative pour maximiser l'emploi des possibilités offertes par l'approche post-censitaire avec une utilisation optimale des ressources disponibles. Le contrôle de la taille de l'échantillon demeure une des difficultés majeures inhérentes à la méthode proposée. La détermination de l'allocation de l'échantillon étant effectuée avant le recensement, tous les calculs doivent être faits à partir de projections basées sur le recensement antérieur. Dans ce contexte, la taille d'un échantillon constitué d'un ensemble de petites régions choisies suivant les résultats de ces projections risque de connaître des variations considérables une fois concrétisée dans le véritable recensement.

On se trouve alors placé devant, d'une part, la possibilité d'obtenir un échantillon insuffisant pour les exigences de qualité des estimations et, d'autre part, l'éventualité de dépasser les ressources allouées à la collecte des données. Afin d'obvier à ces difficultés, la stratégie suivante a été mise en oeuvre. Un nombre cible d'interviews pour chaque population a été calculé pour l'échantillon "oui". Ce nombre était basé sur la taille d'échantillon requise pour produire des estimés répondant à nos critères de qualité. Cependant, un nombre de SD supérieur à celui nécessaire pour obtenir le nombre cible d'interviews a été sélectionné, et cela pour les raisons mentionnées précédemment. Si le nombre réel d'interviews à faire, tel que calculé sur le terrain, est supérieur au nombre cible, alors, pour des raisons de coût, un sous-échantillon de SD est exclu de l'enquête. Ce n'est que pour le bureau régional de Halifax (recouvrant les provinces du Nouveau-Brunswick, de la Nouvelle-Écosse et de l'Île-du-Prince-Édouard) que le nombre d'interviews de l'échantillon "oui" était passablement plus élevé que le nombre cible. Il fut donc décidé d'exclure certains SD de cette partie de l'échantillon. Pour savoir quels SD seraient exclus, il fallait connaître le nombre cible et le nombre réel d'interviews pour chaque SD. Pour 40% des SD, on a dû imputer le nombre réel d'interviews car il n'était pas disponible à temps.

Pour ce faire, on connaissait le total, par district de commissaire au recensement, des nombres réels d'interviews; la portion de ce total qui n'avait pas déjà été utilisée a été distribuée parmi les SD qui requéraient une imputation, proportionnellement au nombre cible d'interviews.

On a ensuite calculé la différence, pour chaque population, entre le nombre réel et le nombre cible d'interviews pour chacune des deux strates de chaque RIP. Une différence positive (réel-cible) indiquait une population pour laquelle des SD pouvaient être exclus de l'enquête. Dans chaque strate, les SD ont été classifiés en trois groupes (1, 2 et 3) selon qu'ils étaient sélectionnés pour trois, deux ou une seule des populations respectivement. Le fichier des SD a ensuite été trié par strate et par groupe de façon ascendante, l'ordre des SD à l'intérieur d'un groupe étant aléatoire. Chaque SD a été considéré successivement et était supprimé pour les trois populations si:

1) une différence positive demeurait non négative après suppression du SD;
2) une différence négative n'était pas réduite davantage.

De cette façon, chaque différence positive a été réduite à un nombre aussi près que possible de zéro compte tenu de l'ordre aléatoire des SD.

REMERCIEMENTS

Les auteurs remercient D.A. Binder pour sa contribution à l'élaboration du plan de sondage et P. Reed pour la partie informatique de l'allocation de l'échantillon. Les auteurs aimeraient également remercier les arbitres pour leurs observations utiles.

APPENDICE

Question n° 20 du formulaire 2B du recensement

20. a) Êtes-vous limité(e) dans vos activités à cause d'une incapacité physique, d'une incapacité mentale ou d'un problème de santé chronique: (consultez le guide)

À la maison?

☐ non, je ne suis pas limité(e)

☐ oui, je suis limité(e)

À l'école ou au travail?

☐ non, je ne suis pas limité(e)

☐ oui, je suis limité(e)

☐ sans objet

Dans d'autres activités, par ex., dans vos trajets entre la maison et votre lieu de travail ou dans vos loisirs?

☐ non, je ne suis pas limité(e)

☐ oui, je suis limité(e)

b) Avez-vous des incapacités ou handicaps à long terme?

☐ non

☐ oui

Questions de sélection pour l'ESLA (questionnaire des adultes)

1. Éprouvez-vous des difficultés à entendre ce qui se dit au cours d'une conversation normale avec une autre personne?
2. Éprouvez-vous des difficultés à entendre ce qui se dit au cours d'une conversation en groupe avec au moins trois autres personnes?
4. Éprouvez-vous des difficultés à lire les caractères ordinaires d'un journal (avec des verres si vous en portez habituellement)?
5. Éprouvez-vous des difficultés à voir clairement la figure de quelqu'un à 12 pieds/4 mètres (par ex., d'un bout à l'autre d'une pièce) avec des verres si vous en portez habituellement?
7. Éprouvez-vous des difficultés à parler et être compris(e)?
8. Éprouvez-vous des difficultés à marcher sur une distance de 400 verges/mètres sans vous reposer (environ trois pâtés de maisons)?
9. Éprouvez-vous des difficultés à monter et descendre un escalier (environ 12 marches)?
10. Éprouvez-vous des difficultés à transporter un objet de 10 livres sur une distance de 30 pieds/5 kilogrammes sur 10 mètres (par ex., un sac d'épicerie)?
11. Éprouvez-vous des difficultés à vous déplacer d'une pièce à une autre?
12. Éprouvez-vous des difficultés à vous tenir debout pendant de longues périodes, c'est-à-dire pendant plus de 20 minutes? Rappelez-vous qu'il s'agit de problèmes qui devraient durer 6 mois ou plus.
13. Lorsque vous êtes debout, éprouvez-vous des difficultés à vous pencher et à ramasser un objet sur le plancher (par ex., un soulier)?

14. Éprouvez-vous des difficultés à vous habiller et vous déshabiller?
15. Éprouvez-vous des difficultés à vous mettre au lit et à sortir du lit?
16. Éprouvez-vous des difficultés à vous couper les ongles d'orteils?
17. Éprouvez-vous des difficultés à vous servir de vos doigts pour saisir ou manier un objet?
18. Éprouvez-vous des difficultés à étendre le bras dans n'importe quelle direction pour prendre quelque chose (par ex., au dessus de votre tête)?
19. Éprouvez-vous des difficultés à couper vos aliments?
20. À cause d'une affection ou un problème de santé chronique qui devrait durer 6 mois ou plus, êtes-vous limité(e) dans le genre ou la quantité d'activités que vous pouvez faire ...

    (i) à la maison? (ii) à l'école ou au travail? (iii) dans vos autres occupations comme les déplacements, les sports ou les loisirs?
21. Un professionnel de l'enseignement ou de la santé vous a-t-il déjà dit que vous aviez des difficultés d'apprentissage?
22. De temps à autre, chacun éprouve des difficultés à se souvenir du nom d'une personne familière ou à apprendre quelque chose de nouveau ou il nous arrive d'être confus pendant quelques instants. Toutefois, avez-vous en permanence des problèmes de mémoire ou d'apprentissage?
23. À cause d'une affection ou d'un problème chronique d'ordre émotif, psychologique, nerveux ou mental, êtes-vous limité(e) dans le genre ou la quantité d'activités que vous pouvez faire ...

    (i) à la maison? (ii) à l'école ou au travail? (iii) dans vos autres occupations comme les déplacements, les sports ou les loisirs?

BIBLIOGRAPHIE

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

CARTER, R.G., GILES, P.D., et SHERIDAN, M.J. (1982). Description and rationale for the screen tests for the January 1983 Disability Survey. Projet d'établissement d'une base de données sur les invalides, Division de la santé, Statistique Canada.

CHAMBRE DES COMMUNES (1981). Obstacles, Rapport du Comité spécial concernant les invalides et les handicapés. Ottawa.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3e éd.). New York: John Wiley.

DOLSON, D., GILES, P., et MORIN, J.-P. (1984). Méthode d'enquête sur les personnes souffrant d'une incapacité à l'aide de questions supplémentaires de l'Enquête sur la population active. *Techniques d'enquête*, 10, 203-214.

DOLSON, D., McCLEAN, K., MORIN, J.-P., et THÉBERGE, A. (1986). Rapport de méthodologie de l'ESLA. Document de travail, Statistique Canada.

GRABOWIECKI, F. (1982). Discussion of the target population for the Disability Survey. Projet d'établissement d'une base de données sur les invalides, Division de la santé, Statistique Canada.

GRABOWIECKI, F. (1983). Contenu des données de l'enquête de Statistique Canada auprès des handicapés. Rapport technique, Division de la santé, Statistique Canada.

LAZARUS, G., et NESICH, R. (1985). A report on the methodology of the Canadian Health and Disability Survey. Document de travail, Statistique Canada.

McDOWELL, I. (1981). Un examen des questions proposées par l'OECD dans le cadre de l'enquête canadienne. *Revue d'épidémiologie et de santé publique*, 29, 412-429.

MORIN, J.-P., et DOWLER, L. (1986). Proposition d'une méthodologie pour l'ESLA-institutions. Document de travail, Statistique Canada.

MORIN, J.-P. (1986). Comparaison initiale de l'ESIC et de l'ESG. Document de travail, Statistique Canada.

ORGANISATION MONDIALE DE LA SANTÉ, (1980). Classification internationale des déficiences, incapacités et handicaps. Genève, Suisse.

RAJ, D. (1968). *Sampling Theory*. New York: McGraw-Hill.

WILSON, R.W., et McNEIL, J.M. (1981). Analyse préliminaire de l'incapacité au sein de l'OCDE, d'après le test initial de l'enquête post-censitaire. *Revue d'épidémiologie et de santé publique*, 29, 469-475.

Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 119-131
Statistique Canada

Comparaison d'estimateurs de totaux de population obtenus par sondages successifs à deux degrés à l'aide de l'information auxiliaire

F.C. OKAFOR[1]

RÉSUMÉ

Singh et Srivastava (1973) ont élaboré un estimateur linéaire non biaisé de moyennes de population qui pourrait être utilisé dans des sondages successifs à l'aide de plusieurs variables auxiliaires dont les moyennes de population connues ne changent pas d'une période à l'autre. Dans le présent document, trois estimateurs composites T_1, T_2 et T_3, utilisant chacun une variable auxiliaire dont la moyenne de population connue change d'une période à l'autre, sont présentés pour l'estimation du total de population de la période donnée. Les estimateurs proposés sont comparés à l'estimateur habituel, T_o, et à l'estimateur habituel de sondages successifs, T', du total de population de la période donnée sans l'aide de l'information auxiliaire. Nous observons que l'utilisation conjuguée de l'information auxiliaire et d'une méthode par sondages successifs ne produit pas toujours uniformément un gain d'efficacité par rapport à T_o ou T'. Toutefois, quand ils ont été appliqués à une enquête visant à estimer la taille moyenne des arbres dans des plantations de teck, les estimateurs T_1, T_2 et T_3 se sont avérés plus efficaces que T_o ou que T'.

MOTS CLÉS: Périodes successives; appariement partiel; variable auxiliaire.

1. INTRODUCTION

La théorie et la pratique du sondage d'une même population à des moments différents - qu'on appelle sondages échelonnés ou sondages successifs - ont été beaucoup étudiées par certains statisticiens d'enquête. Les principaux objectifs des sondages successifs sont d'estimer des paramètres de population (par exemple des totaux, des moyennes, des ratios de population, etc.) pendant la période la plus récente et d'estimer les variations dans ces paramètres d'une période à l'autre.

La théorie des sondages successifs a été proposée pour la première fois par Jessen (1942). Beaucoup d'autres auteurs ont depuis étudié la question, en particulier en ce qui concerne l'estimation de moyennes de population; notamment Singh (1968), Abraham et coll. (1969), Kathuria et Singh (1971) et Kathuria (1976), pour n'en nommer que quelques-uns.

Singh (1968) a été le premier à étendre la théorie des sondages à un seul degré aux sondages échelonnés à deux degrés. Il a utilisé un plan de sondage dans lequel, à la deuxième période, une fraction λ des unités de sondage du premier degré (USPD) choisies à la période précédente est retenue, en plus des unités de sondage correspondants du second degré (USSD) et d'une fraction μ ($\lambda + \mu = 1$) choisie de nouveau. Il a ensuite obtenu un estimateur non biaisé à variance minimum de la moyenne de population de la période donnée.

Abraham et coll. (1969) ont considéré le cas où un appariement partiel des unités était effectué aux deux degrés. Les unités étaient choisies à l'aide d'une méthode de sondage aléatoire simple et sans remise (SASSR). Kathuria (1975) a modifié cette façon de procéder en utilisant une méthode de sélection des USPD avec probabilités proportionnelles à la taille et avec remise (PPTAR) et proposé un estimateur linéaire composite pour estimer la moyenne de population de la période donnée.

[1] F.C. Okafor, Département de statistique, Université d'Ibadan, Ibadan, Nigeria.

Quand une variable auxiliaire est très corrélée à la caractéristique étudiée, on peut améliorer l'estimation de la moyenne (ou du total) de population de cette caractéristique en utilisant la variable auxiliaire. Shivtar Singh et Srivastava (1973) ont utilisé de l'information auxiliaire pour améliorer l'estimateur de Singh (1968). Ils ont obtenu un estimateur linéaire non biaisé de la moyenne de population de la période la plus récente en utilisant des variables auxiliaires dont les moyennes de population étaient connues et ne changeaient pas d'une période à l'autre. Kathuria (1978) a poussé davantage dans cette voie en supposant que la moyenne de population de la variable auxiliaire n'est pas connue. Il a utilisé une technique d'échantillonnage double (ou sondage à deux phases) pour estimer d'abord la moyenne de population de la variable auxiliaire et ensuite la moyenne de population de la caractéristique à l'étude.

Dans leurs ouvrages, Shivtar Singh et coll. (1973) et Kathuria (1978) ont supposé qu'il était possible d'obtenir l'information nécessaire sur les variables auxiliaires des répondants ou des unités déclarantes (UD). Cela n'est généralement pas le cas. Il peut arriver que le caractère délicat de la question ou le refus pur et simple des répondants de fournir toute information faussent l'information sur la variable auxiliaire au point de la rendre inutile. Il se peut aussi que l'information sur la variable auxiliaire ne puisse être recueillie parce que la question qui aurait permis de l'obtenir n'est pas incluse dans le questionnaire.

Shivtar Singh et coll. ont également supposé que le total de population connu de la variable auxiliaire est le même à toutes les périodes. Il se peut que cela ne soit pas vrai en pratique. Si le total de population de la caractéristique principale varie d'une période à l'autre, il y a tout lieu de penser que le total de population de toute autre variable qui serait corrélée à la caractéristique principale variera également.

Dans le présent document, trois estimateurs composites de total de population utilisant de l'information auxiliaire et un plan de sondage comportant des sondages successifs à deux degrés sont proposés. Les performances des trois estimateurs sont comparées empiriquement entre elles; les trois estimateurs ont également été appliqués à une enquête visant à estimer la taille moyenne des arbres dans des plantations de teck.

2. SONDAGE POUR DEUX PÉRIODES

Pour les trois estimateurs proposés, nous supposons que le total de population de la variable auxiliaire change à la deuxième période.

Les estimateurs du total (ou de la moyenne) de population fondés sur un plan d'appariement partiel sont meilleurs que les estimateurs habituels du total (ou de la moyenne) de population sans appariement partiel. On peut donc s'attendre que les estimateurs proposés, T_1, T_2 et T_3, donneront de meilleurs résultats que l'estimateur habituel du total de population, T_0, et que l'estimateur fondé sur le même plan d'appariement partiel mais n'utilisant pas d'information auxiliaire, T'.

Dans le calcul de ces estimateurs, nous supposons que:

(i) la taille de l'échantillon est constante à chaque période,
(ii) la mesure normalisée P_i de la taille de la i-ième unité de sondage du premier degré (USPD) est fixée pour chaque période,
(iii) N et M_i, les tailles de population respectives des USPD et des unités de sondage du second degré (USSD) prélevées à partir de la i-ième USPD, sont constantes dans les deux périodes,
(iv) le total (ou la moyenne) de population de la variable auxiliaire est connu.

Les hypothèses (i) à (iii) s'appliquent à T', T_1, T_2 et T_3, tandis que l'hypothèse (iv) s'applique à T_1, T_2 et T_3, mais non à T' et T_0.

À la première période, un échantillon S_1 de n USPD est tiré avec probabilités proportionnelles à la taille et avec remise (PPTAR) à l'aide d'une mesure normalisée de la taille de la i-ième unité ($i = 1, 2, \ldots, N$). Pour choisir les USSD, nous adoptons la méthode de

Cochran (1977, p. 306), qui stipule que si la i-ième USPD de S_1 est choisie θ_i fois ($i = 1, 2, \ldots, n$), on tire θ_i sous-échantillons indépendants de taille m_i à partir des M_i USSD.

À la deuxième période, nous prélevons un échantillon de λn USPD ($0 < \lambda > 1$) à partir de S_1 selon un plan de sondage aléatoire simple et sans remise (SASSR). Les USSD choisies à la première période sont retenues pour chacune de ces λn USPD appariées. Ensuite, un nouvel échantillon de μn ($\mu = 1 - \lambda$) USPD est tiré indépendamment des n USPD par sondage avec PPTSR, avec P_i comme mesure normalisée de la taille de la i-ième USPD. Dans chacune des μn USPD, les USSD sont choisies de la même façon qu'à la première période.

3. NOTATION

Nous définissons y_{ij} (x_{ij}) comme la valeur de la variable à l'étude pour la j-ième USSD dans la i-ième pendant la période donnée (ou la période précédente). De plus, z_{hij} est définie comme la valeur de la variable auxiliaire pour la j-ième USSD dans la i-ième USPD à la h-ième période ($h = 1, 2$). Les moyennes de l'échantillon des USSD dans la i-ième USPD sont

$$\bar{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}, \quad \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \quad \text{et} \quad \bar{z}_{hi} = \frac{1}{m_i} \sum_{j=1}^{m_i} z_{hij}.$$

Le total de population pour la i-ième USPD et le total de population pour l'ensemble des USPD qui correspondent à la variable auxiliaire sont

$$Z_{hi} = \sum_{j=1}^{M_i} z_{hij} \quad \text{et} \quad Z_h = \sum_{j=1}^{N} Z_{hi}.$$

Nous définissons aussi quelques autres notations de la façon suivante:

$$S_b^2(y) = \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y\right)^2 \text{ est la variance entre les USPD;}$$

$$S_w^2(y) = \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{wi}^2(y) \text{ est la variance entre les USSD de l'ensemble des USPD;}$$

$$S_{wi}^2(y) = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_i)^2 \text{ est la variance entre les USSD de la } i\text{-ième USPD;}$$

$$S^2(y) = S_b^2(y) + S_w^2(y);$$

$$C_b(x,y) = \rho_b S_b(x)\, S_b(y) \text{ est la covariance de } x \text{ et } y \text{ entre les USPD;}$$

$$C_w(x,y) = \rho_w S_w(x)\, S_w(y) \text{ est la covariance de } x \text{ et } y \text{ entre les USSD de l'ensemble des USPD;}$$

$$C(x,y) = C_b(x,y) + C_w(x,y).$$

Les coefficients de corrélation entre x et y calculés entre les USPD et à l'intérieur des USPD sont respectivement ρ_b et ρ_w.

4. ESTIMATEURS DE TOTAUX DE POPULATION ET VARIANCES OPTIMUMS

4.1 Cas (i)

Le premier estimateur du total de population Y, de la seconde période est utilisé lorsqu'on n'a pas d'information sur la variable auxiliaire, mais qu'on connaît le total de population de la variable auxiliaire pour les USPD choisies. Il s'exprime ainsi:

$$T_1 = \theta(1)\, T_m(1) + (1 - \theta(1))\, T_u(1) \tag{4.1}$$

où $\theta(1)$ est une constante choisie e telle sorte que la variance de T_1, $V(T_1)$ est minimum, tandis que

$$\begin{aligned} T_m(1) = {} & \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\} \\ & - b(1) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} \right. \\ & \left. - \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} \right] \end{aligned}$$

est l'estimateur par différence de Y fondé sur l'échantillon apparié, que

$$T_u(1) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\}$$

est l'estimateur de Y fondé sur l'échantillon non apparié et que $k(1)$ et $b(1)$ sont des constantes connues.

Pour cet estimateur, on suppose qu'on connaît le total de population de la variable auxiliaire, Z_i, pour chaque USPD choisie à la première période. On connaît également le total de population pour l'ensemble de USPD, Z, à chaque période. Aucune autre information sur la variable auxiliaire n'est obtenue des répondants ou des unités déclarantes (UD).

Maintenant, en minimisant $V(T_1)$ par rapport à $\theta(1)$ et en résolvant l'équation, on obtient la valeur optimum suivante de $\theta(1)$

$$\theta_0(1) = \lambda A_2(1) / \Delta(1)$$

où

$$A_2(1) = S^2(y) + k^2(1)\, S_b^2(z_2) - 2k(1)\, C_b(z_2, y),$$

$$\Delta(1) = A_2(1) + \mu^2 \{ b^2(1)\, A_1(1) - 2b(1)\beta(1) \}.$$

La valeur optimum de $k(1)$ est obtenue en minimisant $V(T_u(1))$ par rapport à $k(1)$. Cela donne $k_0(1) = C_b(z_2, y) / S_b^2(z_2)$.

On peut montrer qu'en utilisant la méthode proposée par Jessen (1942), la valeur optimum de $V(T_1)$ pour une valeur donnée de λ est

$$V_0(T_1) = \frac{1}{n}\,[A_2(1) + \mu\{b^2(1)A_1(1) - 2b(1)\,\beta(1)\}]\,A_2(1)/\Delta(1) \tag{4.2}$$

où

$$A_1(1) = S^2(x) + k^2(1)\,S_b^2(z_1) - 2k(1)\,C_b(z_1,x),$$

$$\beta(1) = C(x,y) + k^2(1)C_b(z_1,z_2) - k(1)\,[C_b(x,z_2) + C_b(z_1,y)],$$

$$\Delta(1) = A_2(1) + \mu^2\{b^2(1)\,A_1(1) - 2b(1)\,\beta(1)\}.$$

Lorsqu'on minimise la variance de $T_m(1)$, la valeur optimum de $b(1)$ est

$$b_0(1) = \beta(1)/A_1(1).$$

Si on substitue $b_0(1)$ dans (4.2), la variance optimum devient

$$V_0(T_1) = \frac{1}{n}\left[\frac{A_1(1)\,A_2(1) - \mu\beta^2(1)}{A_1(1)\,A_2(1) - \mu^2\beta^2(1)}\right]A_2(1). \tag{4.3}$$

Lorsqu'on minimise $V_o(T_1)$ dans (4.2) par rapport à μ, la fraction d'appariement optimum se ramène à $\lambda_0 = 1 - \mu_0$ où

$$\mu_0 = A_2(1)\,[A_2(1) + \{A_2^2(1) + A_2(1)\,(b^2(1)A_1(1) - 2b(1)\beta(1))\}^{1/2}]^{-1}. \tag{4.4}$$

Si $A_2(1) = A_1(1)$, c'est-à-dire si la variation de la population est la même dans les deux périodes, l'expression (4.3) donne

$$V_0(T_1) = \frac{1}{n}\left[\frac{A^2(1) - \mu\beta^2(1)}{A^2(1) - \mu^2\beta^2(1)}\right]A(1) \tag{4.5}$$

et si on substitue $b_0(1)$ à $b(1)$, la fraction d'appariement optimum donnée dans l'équation (4.4), μ_0, devient

$$\mu_0 = A(1)\,[A(1) + [A^2(1) - \beta^2(1)\}^{1/2}]^{-1}. \tag{4.6}$$

Quand on substitue μ_0 dans (4.5), la variance se ramène à

$$V_0(T_1) = \frac{1}{2n}\,[A(1) + \{A^2(1) - \beta^2(1)\}^{1/2}]. \tag{4.7}$$

4.2 Cas (ii)

Le deuxième estimateur est l'estimateur habituel dans lequel l'information relative aussi bien à la caractéristique principale qu'à la caractéristique auxiliaire a été obtenue des unités déclarantes et dans lequel le total de population de la caractéristique auxiliaire est connu. Il s'exprime ainsi:

$$T_2 = \theta(2)\, T_m(2) + (1 - \theta(2))\, T_u(2), \tag{4.8}$$

où

$$\begin{aligned} T_m(2) = \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} & \left\{ \frac{M_i \bar{y}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{2i}}{P_i} - Z_2 \right) \right. \\ & - b(2) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{1i}}{P_i} - Z_1 \right) \right\} \right. \\ & \left. \left. - \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{1i}}{P_i} - Z_1 \right) \right\} \right] , \end{aligned}$$

et

$$T_u(2) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{2i}}{P_i} - Z_2 \right) \right\} .$$

Ici, le total de population global de la variable auxiliaire est connu dans les deux périodes. En outre, l'information sur la variable auxiliaire, z_{ij}, est obtenue pour chaque USSD de l'échantillon. Il s'agit de la façon habituelle d'utiliser l'information auxiliaire dans les plans de sondage décrits dans les ouvrages portant sur le sujet. On peut montrer que la variance optimum de T_2 est

$$V_0(T_2) = \frac{1}{n} [A_2(2) + \mu\{b^2(2) A_1(2) - 2b(2)\beta(2)\}] A_2(2)/\Delta(2) \tag{4.9}$$

et que le poids optimum est

$$\theta_0(2) = \lambda A_2(2)/\Delta(2)$$

où

$$A_2(2) = S^2(y) + k^2(2)\, S^2(z_2) - 2k(2)\, C(z_2, y),$$

$$A_1(2) = S^2(x) + k^2(2)\, S^2(z_1) - 2k(2)\, C(z_1, x),$$

$$\beta(2) = C(x,y) + k^2(2)\, C(z_1, z_2) - k(2)\, \{C(z_1, y) + C(x, z_2)\},$$

$$\Delta(2) = A_2(2) + \mu^2 \{b^2(2)\, A_1(2) - 2b(2)\, \beta(2)\}.$$

La valeur optimum de $k(2)$ est $k_0(2) = C(z_2,y)/S_2(z_2)$.

En substituant le coefficient optimum de régression, $b_0(2) = \beta(2)/A_1(2)$, obtenue en minimisant la variance de $T_m(2)$, dans (4.9) et en supposant que $A_2(2) = A_1(2) = A(2)$, on obtient

$$V_0(T_2) = \frac{1}{n}\left[\frac{A^2(2) - \mu\beta^2(2)}{A^2(2) - \mu^2\beta^2(2)}\right] A(2). \tag{4.10}$$

Si la valeur optimum de μ est substituée dans (4.10), la variance devient

$$V_0(T_2) = \frac{1}{2n}\,[A(2) + \{A^2(2) - \beta^2(2)\}^{1/2}]. \tag{4.11}$$

4.3 Cas (iii)

La troisième façon d'utiliser l'information auxiliaire connue pour améliorer l'estimation du total de population de la période donnée, Y, dans le cadre d'un plan de sondage donné ressemble beaucoup à la deuxième façon. La seule différence est qu'on ne connaît pas le total de population de la caractéristique auxilière; par contre, on connaît la moyenne de population pour les USPD choisies.

L'estimateur s'exprime ainsi:

$$T_3 = \theta(3)\,T_m(3) + (1 - \theta(3))\,T_u(3), \tag{4.12}$$

où

$$\begin{aligned} T_m(3) = {} & \frac{1}{\lambda n}\sum_{i=1}^{\lambda n}\frac{M_i}{P_i}\{\bar{y}_i - k(3)\,(\bar{z}_{2i} - \bar{Z}_{2i})\} \\ & - b(3)\left[\frac{1}{\lambda n}\sum_{i=1}^{\lambda n}\frac{M_i}{P_i}\{\bar{x}_i - k(3)\,(\bar{z}_{1i} - \bar{Z}_{1i})\}\right. \\ & \left. - \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{P_i}\{\bar{x}_i - k(3)\,(\bar{z}_{1i} - \bar{Z}_{1i})\}\right], \end{aligned}$$

et

$$T_u(3) = \frac{1}{n\mu}\sum_{i=1}^{n\mu}\frac{M_i}{P_i}\{\bar{y}_i - k(3)\,(\bar{z}_{2i} - \bar{Z}_{2i})\}.$$

Pour cet estimateur, on suppose que les valeurs tant de la variable principale que de la variable auxiliaire sont obtenues pour chaque USSD de l'échantillon dans les deux périodes. On suppose également que la moyenne de population, $\bar{Z}_i$, de la variable auxiliaire est connue pour les USPD choisies.

La variance optimum de T_3 pour une valeur donnée de λ s'exprime ainsi:

$$V_0(T_3) = \frac{1}{n}\,[A_2(3) + \mu\{b^2(3)A_1(3) - 2b(3)\beta(3)\}]\,A_2(3)/\Delta(3) \tag{4.13}$$

tandis que le poids optimum est donné, comme d'habitude, par l'expression suivante:

$$\theta_0(3) = \lambda A_2(3)/\Delta(3),$$

où

$$A_2(3) = S^2(y) + k^2(3)\, S_w^2(z_2) - 2k(3)\, C_w(z_2,y),$$

$$A_1(3) = S^2(x) + k^2(3)\, S_w^2(z_1) - 2k(3)\, C_w(z_1,x),$$

$$\beta(3) = C(x,y) + k^2(3)\, C_w(z_1,z_2) - k(3)\, \{C_w(z_1,y) + C_w(z_2,x)\},$$

$$\Delta(3) = A_2(3) + \mu^2\, \{b^2(3) A_1(3) - 2b(3)\, \beta(3)\}.$$

La valeur optimum de $k(3)$ est $k_0(3) = C_w(z_2,y)/S_w^2(z_2)$.

Si on substitue le coefficient optimum de régression dans (4.13) et si on suppose que la variance de population est la même dans les deux périodes, (4.13) se ramène alors à

$$V_0(T_3) = \frac{1}{n}\left[\frac{A^2(3) - \mu\beta^2(3)}{A^2(3) - \mu^2\beta^2(3)}\right] A(3). \tag{4.14}$$

Quand la valeur optimum de μ est substituée dans (4.14), la variance devient

$$V_0(T_3) = \frac{1}{2n}\,[A(3) + [A^2(3) - \beta^2(3)]^{1/2}]. \tag{4.15}$$

4.4 Efficacité des estimateurs proposés

Nous utiliserons les variances données en (4.7), (4.11) et (4.15) pour comparer l'efficacité des trois estimateurs, T_1, T_2 et T_3 par rapport à

$$T_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{M_i \bar{y}_i}{P_i}.$$

T_0 est l'estimateur de y quand il n'y a pas d'appariement partiel des unités et qu'aucune information auxiliaire n'est utilisée. Nous avons aussi comparé l'efficacité de T_0 par rapport à l'estimateur habituel avec appariement partiel, T', qui n'utilise pas d'information auxiliaire, pour mieux faire comprendre la performance des estimateurs proposés.

L'estimation habituelle avec appariement partiel est définie comme suit:

$$T' = \theta' T'_m + (1 - \theta')\, T'_u, \tag{4.16}$$

où

$$T'_m = \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i \bar{y}_i}{P_i} - b' \left\{ \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i \bar{x}_i}{P_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{M_i \bar{x}_i}{P_i} \right\},$$

et

$$T'_u = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \frac{M_i \bar{y}_i}{P_i}.$$

La variance optimum de T', obtenue en utilisant la valeur optimum de b', $b'_0 = C(x,y)/S^2(x)$, et en supposant que $S^2(y) = S^2(x)$ est

$$V_0(T') = \frac{1}{n} \left[\frac{S^2(y) - \mu C(x,y)}{S^2(y) - \mu^2 C(x,y)} \right] S^2(y). \tag{4.17}$$

Si on substitue la valeur optimum de μ dans (4.17), la variance de T' devient

$$V_0(T') = \frac{1}{2n} [S^2(y) + \{S^4(y) - C^2(x,y)\}^{1/2}]. \tag{4.18}$$

Pour calculer l'efficacité des divers estimateurs, les hypothèses suivantes ont été faites au sujet des coefficients de corrélation et de la constante k:

$$\rho_b(x,z_2) = \rho_b(z_1,y) = \rho_b(z_1,z_2) = \rho_b;$$

$$\rho_w(x,z_2) = \rho_w(z_1,y) = \rho_w(z_1,z_2) = \rho_w;$$

$$k(1) = k(2) = k(3) = 1.$$

Les valeurs de l'efficacité n'ont été présentées que pour des valeurs positives de ρ_b et ρ_w et une série de valeurs de

$$\delta = S_w^2(y)/S_b^2(y),\ R_b = S_b^2(z)/S_b^2(y) \text{ et } R_w = S_w^2(z)/S_b^2(y).$$

Si on regarde le tableau 2, on constate qu'aucune des stratégies T_1, T_2 ou T_3 (plan de sondage et estimateur) n'est uniformément plus efficace que la stratégie T_0. C'est le contraire pour T', qui est toujours plus efficace que T_0, et qui, au pire, n'offre qu'un faible gain par rapport à T_0 (voir tableau 1).

D'après les résultats des tableaux 1 et 2, il faut préférer T_1 à T' seulement lorsque $R_b = 0.05$, $\rho_b = 0.8$ et $R_b = 0.5$.

Tableau 1

Efficacité de T' par rapport à T_0

ρ_b	δ	$\rho_w = 0.2$	$\rho_w = 0.8$
0.2	0.05	1.01	1.01
	0.5	1.01	1.04
	5.0	1.01	1.17
	0.05	1.22	1.25
0.8	0.5	1.11	1.25
	5.0	1.02	1.25

T_2 est meilleur que T' quand

(i) $\rho_w = 0.2,\ R_b = R_w = 0.05;$

(ii) $\rho_b = \rho_w = 0.8,\ R_b = R_w = 0.05,\ 0.5;$

(iii) $\delta = 0.5,\ 5.0,\ R_w = R_b = 0.05,\ \rho 0.5,\ \rho_b = 0.2$ et $\rho_w = 0.8.$

T_3 est généralement plus efficace que T' quand

(i) $\delta = 5.0,\ \rho_w = 0.8;$

(ii) $\delta = 0.5,\ \rho_w = 0.8$ and $R_w = 0.05,\ 0.5.$

Le gain maximum d'efficacité de T' par rapport à T_0 est de 25% (voir tableau 1). D'après les chiffres du tableau 2, le gain maximum de T_1 par rapport à T_0 est de 155%; il est obtenu quand $\rho_b = \rho_w = 0.8, \delta = 0.05$ et $R_b = 0.5$. Le gain maximum d'efficacité de T_2 par rapport à T_0 est de 172%; il est obtenu quand $\rho_b = \rho_w = 0.8$ et $\delta = R_w = 0.05$. Nous constatons également que lorsque $\rho_b = \rho_w = 0.8$, et $\delta = R_w = 5.0$, le gain maximum de T_3 par rapport à T_0 est de 104%. Il est donc évident que l'utilisation d'une variable auxiliaire a beaucoup amélioré l'efficacité de l'appariement partiel des unités.

Si maintenant nous comparons entre elles les trois stratégies T_1, T_2 et T_3, nous pouvons conclure qu'aucune n'est uniformément meilleure qu'une autre, même si le gain maximum d'efficacité de T_2 est supérieur au gain maximum d'efficacité de T_1, lui-même plus élevé que le gain maximum de T_3 par rapport à T_0. En général, T_1 est meilleur que T_2 quand $\rho_w = 0.2$, tandis que T_2 est meilleur que T_1 quand $\rho_w = 0.8$. T_1 est préférable à T_3 quand $\rho_b = 0.8$, $\rho_w = 0.2$ et $R_b = 0.05,\ 0.5$ ou quand $\rho_b = \rho_w = 0.8$ et $\delta = R_b = 0.05$. Enfin, T_3 est meilleur que T_2 quand $\rho_w = 0.8$ et $R_b = 5.0$ ou quand $\rho_b = \rho_w = 0.2$ et $R_b = 0.5,\ 5.0$.

5. APPLICATION

Les estimateurs proposés ont été appliqués à une enquête sur la hauteur des arbres dans des plantations de teck. L'objectif était d'estimer la hauteur moyenne des arbres en utilisant la circonférence des troncs comme information auxiliaire.

Tableau 2

Efficacité de T_1, T_2, et T_3 par rapport à T_0

		$\rho_w = 0.2$								
		$R_b = 0.05$			$R_b = 0.5$			$R_b = 5.0$		
			R_w			R_w			R_w	
ρ_b	δ	0.05	0.5	5.0	0.05	0.5	5.0	0.05	0.5	Stratégie
0.2	0.05	1.04	1.04	1.04	0.83	0.83	0.83	0.20	0.20	T_1
		1.01	0.73	0.18	0.81	0.62	0.17	0.20	0.19	T_2
		0.98	0.71	0.18	0.98	0.71	0.18	0.98	0.71	T_3
	0.5	1.03	1.03	1.03	0.87	0.87	0.87	0.27	0.27	T_1
		1.04	0.85	0.26	0.88	0.74	0.25	0.27	0.25	T_2
		1.02	0.84	0.26	1.02	0.84	0.26	1.02	0.84	T_3
	5.0	1.02	1.02	1.02	0.97	0.97	0.97	0.60	0.60	T_1
		1.04	1.03	0.67	0.99	0.99	0.65	0.60	0.60	T_2
		1.03	1.03	0.67	1.03	1.03	0.67	1.03	1.03	T_3
0.8	0.05	1.62	1.62	1.62	2.53	2.53	2.53	0.45	0.45	T_1
		1.53	0.94	0.19	2.35	1.23	0.20	0.45	0.38	T_2
		1.16	0.77	0.18	1.16	0.77	0.18	1.16	0.77	T_3
	0.5	1.34	1.34	1.34	1.74	1.74	1.74	0.45	0.45	T_1
		1.34	1.03	0.27	1.76	1.28	0.29	0.54	0.48	T_2
		1.11	0.88	0.26	1.11	0.88	0.26	1.11	0.88	T_3
	5.0	1.07	1.07	1.07	1.13	1.13	1.13	0.83	0.83	T_1
		1.10	1.09	0.69	1.16	1.15	0.72	0.84	0.83	T_2
		1.05	1.03	0.67	1.05	1.03	0.67	1.05	1.03	T_3

		$\rho_w = 0.8$										
		$R_b = 0.05$				$R_b = 0.5$			$R_b = 5.0$			
				R_w			R_w			R_w		
ρ_b	δ	5.0	0.05	0.5	5.0	0.05	0.5	5.0	0.05	0.5	5.0	Stratégie
0.2	0.05	0.20	1.05	1.05	1.05	0.83	0.83	0.83	0.20	0.20	0.20	T_1
		0.11	1.07	0.85	0.23	0.85	0.70	0.21	0.19	0.19	0.12	T_2
		0.18	1.04	0.83	0.23	1.04	0.83	0.23	1.04	0.83	0.23	T_3
	0.5	0.27	1.06	1.06	0.89	0.89	0.89	0.89	0.27	0.27	0.27	T_1
		0.15	1.21	1.30	0.41	1.00	1.06	0.38	0.28	0.28	0.19	T_2
		0.26	1.18	1.26	0.41	1.18	1.26	0.41	1.18	1.26	0.41	T_3
	5.0	0.60	1.17	1.17	1.17	1.09	1.09	1.09	0.62	0.62	0.62	T_1
		0.46	1.31	1.64	2.03	1.22	1.51	1.87	0.67	0.76	0.84	T_2
		0.67	1.30	1.63	2.00	1.30	1.63	2.00	1.30	1.63	2.00	T_3
0.8	0.05	0.45	1.65	1.65	1.65	2.55	2.55	2.55	0.46	0.46	0.46	T_1
		0.15	1.70	1.22	0.25	2.72	1.64	0.27	0.46	0.42	0.18	T_2
		0.18	1.27	0.98	0.24	1.26	0.98	0.24	1.27	0.98	0.24	T_3
	0.5	0.45	1.50	1.50	1.50	1.88	1.88	1.88	0.56	0.56	0.56	T_1
		0.21	1.75	1.83	0.46	2.34	2.65	0.50	0.59	0.61	0.31	T_2
		0.26	1.40	1.43	0.43	1.40	1.43	0.43	1.40	1.43	0.43	T_3
	5.0	0.83	1.30	1.30	1.30	1.35	1.35	1.35	0.95	0.95	0.95	T_1
		0.85	1.46	1.85	2.25	1.53	1.98	2.53	1.03	1.22	1.38	T_2
		0.67	1.39	1.74	2.04	1.39	1.74	2.04	1.39	1.74	2.04	T_3

Tableau 3

Efficacité estimée des estimateurs proposés par rapport à T_0 dans l'estimation de la hauteur moyenne des arbres dans des plantations de teck

Estimateurs	Hauteur moyenne (m)	Variance (m^2)	Efficacité estimée en %
T_0 (sans appariement)	20.04	6.3118	100
T' (appariement partiel)	18.06	4.0680	155
T_1	17.86	0.0718	8791
T_2	17.31	0.0651	9635
T_3	17.99	4.0183	157

Les arbres utilisés dans l'enquête ont été plantés en 1965 suivant différents espacements, ce qui a produit des plantations ayant le nombre d'arbres suivant par hectare: 2,000, 800, 400 et 250. Pour mesurer la hauteur des arbres, un périmètre de 40 mètres sur 40 a été tracé dans chacune des 8 plantations (USPD) prélevées parmi 16 plantations à l'aide d'un plan de sondage avec PPTAR. Le nombre d'arbres dans chaque plantation a été utilisé comme mesure de la taille. Tous les arbres à l'intérieur du périmètre de 40 m sur 40 formaient les unités de sondage du second degré et leur circonférence à hauteur de poitrine a été mesurée. Pour le calcul de la hauteur, un sous-échantillon d'arbres a été sélectionné des arbres de périmètre 40m sur 40 dans chaque USPD choisie. Une première série de calculs a été faite en 1981 et une seconde en 1983. Le plan de sondage utilisé était le même que celui qui a été décrit dans la section 2 et comportait un appariement partiel des USPD dans une proportion de 50 %.

Les valeurs estimées de l'efficacité sont présentées dans le tableau 3. Les estimations de la variance et de la covariance de l'échantillon ont été utilisées pour calculer les variances optimums de T', T_1, T_2 et T_3 parce que les valeurs de ces variances et covariances pour l'ensemble de la population n'étaient pas connues. Par conséquent, le fait que les valeurs obtenues pour les variances estimées optimums de T_1 et T_2 sont faibles est attribuable d'une part, à l'utilisation de données d'échantillon et, d'autre part, à la nature même des estimateurs.

Nous constatons que l'estimateur T_2 est plus efficace que T_1 et que T_3, tandis que T_1 est plus efficace que T_3 dans l'estimation de la hauteur moyenne des arbres à l'aide de la circonférence comme information auxiliaire.

REMERCIEMENTS

Je tiens à remercier l'arbitre et le rédacteur associé, dont les précieux commentaires m'ont aidé à améliorer le présent document. Je remercie également Dr. O. Abe du Département de statistique de l'Université d'Ibadan, pour les retouches apportées à la version préliminaire du document révisé.

BIBLIOGRAPHIE

ABRAHAM, T.P., KHOSLA, R.K., et KATHURIA, O.P. (1969). Some investigations of the use of successive sampling in pest and disease surveys. *Journal of the Indian Society of Agricultural* Statistics, 21, 43-57.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3e éd). New York: John Wiley.

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

KATHURIA, O.P. (1975). Some estimators in two-stage sampling on successive occasions with partial matching at both stages. *Sankhyā*, Sér. C, 37, 147-162.

KATHURIA, O.P. (1978). Double sampling on successive occasions using a two-stage design. *Journal of the Indian Society of Agricultural Statistics*, 30, 49-64.

KATHURIA, O.P. et SINGH, D. (1971). Relative efficiencies of some alternative procedures in two-stage sampling on successive occasions. *Journal of the Indian Society of Agricultural Statistics*, 23, 101-114.

SINGH, S., et SRIVASTAVA, A.K. (1973). Use of auxiliary information in two-stage successive sampling. *Journal of the Indian Society of Agricultural Statistics*, 25, 101-114.

SINGH, D. (1968). Estimates in successive sampling using a multistage design. *Journal of the American Statistical Association*, 63, 99-112.

'Résultats optimaux en situation de non-réponse' par V.P. Godambe et M.E. Thompson, Techniques d'enquête (1986), 12, 31-39.

L'équation (2.6), qui donne la définition de la fonction d'estimation optimale dans $\bar{H}''$ $(p, \mathbf{q})$, devrait être

$$h''^* = \sum_{i \in s'} (y_i - \theta x_i)\alpha_i / \pi_i q_i .$$

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; 1, l).

3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.