12-001
c.3

Statistique Canada Statistics Canada

STATISTICS CANADA STATISTIQUE CANADA
OCT 21 1988
LIBRARY
BIBLIOTHÈQUE

TECHNIQUES D'ENQUÊTE

UNE REVUE
DE
STATISTIQUE CANADA

VOLUME 14, NUMÉRO 1
JUIN 1988

Canada

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

JUIN 1988

Publication autorisée par
le ministre des Approvisionnements
et Services Canada

©Ministre des Approvisionnements
et Services Canada 1988

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.

Septembre 1988

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

Paiement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 14, n° 1

SSN 0714-0045

Ca OOS

10100290O9

Statistics Canada Library
Bibliothèque Statistique Canada

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, *U. 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*
M. Gonzalez, *U.S. Office of Management and Budget*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*
M.N. Murthy, *Applied Statistics Centre, India*
W.M. Podehl, *Statistique Canada*
J.N.K. Rao, *Carleton University*
D.B. Rubin, *Harvard 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, J. Gambino et J.-L. Tambay, *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, 4^{e} 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 14, numéro 1, juin 1988

TABLE DES MATIÈRES

Dans ce numéro 1

J. RYTEN
Les erreurs dans les statistiques du commerce extérieur 3

L. KISH
Plans de sondage à usages multiples 19

P. LAVALLÉE et M.A. HIDIROGLOU
Sur la stratification de populations asymétriques 35

D. KASPRZYK
Domaines de recherche relatifs à la SIPP (enquête sur le revenu et la participation aux programmes) 47

D. SCHNELL, W.J. KENNEDY, G. SULLIVAN, H.J. PARK, et W.A. FULLER
Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes 63

Section Spéciale - Erreur de couverture dans le recensement

G. DIFFENDAL
Test des opérations de redressement de 1986 dans le Central Los Angeles County.. 75

N. SCHENKER
Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986 93

H. HOGAN et K. WOLTER
Mesure de l'erreur dans une enquête post-censitaire 105

P.P. BIEMER
Modélisation de l'erreur d'appariement et son effet sur les estimations de l'erreur d'observation du recensement 125

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 1-2
Statistique Canada

Dans ce numéro

Quatre des neuf articles de ce numéro traitent de **l'erreur de couverture dans le recensement**. Ces articles, de même que ceux qui paraîtront dans le numéro de décembre 1988 de la revue, viennent enrichir la documentation croissante sur le sujet. L'initiative dont a fait preuve Kirk Wolter a permis de réaliser ces sections spéciales.

Nous savons que les chiffres du recensement sont inexacts à cause de l'erreur de couverture et ces problèmes ont suscité beaucoup d'intérêt ces derniers temps, aussi bien auprès des statisticiens (théoriciens et praticiens) que des décideurs. En conséquence, les méthodes permettant d'évaluer la qualité de ces chiffres de même que les limites de ces méthodes, les techniques de redressement (basées aussi bien sur le plan de sondage que sur des modèles) visant à améliorer la qualité des chiffres de population, l'incidence du sous-dénombrement sur les programmes gouvernementaux et d'autres études du même ordre ont pris une importance grandissante. Dans de nombreux pays, des études d'évaluation sont faites pour estimer l'erreur de couverture soit pendant le recensement, soit après. Ainsi, au Canada, la contre-vérification des dossiers est la plus importante étude réalisée pour mesurer le sous-dénombrement. Aux Etats-Unis, c'est l'enquête postcensitaire (EP) qui est, depuis 1950, un des principaux moyens utilisés pour mesurer le taux de couverture du recensement.

En 1986, le U.S. Bureau of the Census a mené, à Los Angeles, une étude intitulée "test des opérations de redressement" (TOR) qui mettait à l'essai un nouveau plan de sondage pour l'EP. Trois articles de la section spéciale, soit ceux de Diffendal, Schenker et Hogan et Wolter, fournissent une évaluation détaillée des méthodes et procédures utilisées dans cette nouvelle EP et analysent à fond les résultats de la recherche ainsi que les difficultés rencontrées pendant le TOR et les résultats obtenus au moyen de ce test. Diffendal donne un aperçu général de celui-ci, décrivant les méthodes utilisées et les opérations exécutées. Il présente également un bref historique des études réalisées aux Etats-Unis pour estimer le taux de couverture du recensement et décrit les événements récents qui ont amené le U.S. Bureau of the Census à effectuer des études complexes.

Schenker traite de trois méthodes utilisées pour régler le problème des données manquantes: l'imputation par hot deck, l'utilisation des modèles de régression logistique et le redressement par la pondération. Le choix d'une méthode dépend du type de données manquantes. On se sert, par exemple, de la régression logistique pour imputer les valeurs des caractéristiques binaires. Se servant des données du TOR, l'auteur compare les estimations de l'erreur de couverture obtenues au moyen de diverses méthodes d'imputation.

Hogan et Wolter examinent en détail les sources possibles d'erreur dans les nouvelles estimations produites à partir des résultats de l'EP et évaluent l'incidence de chaque type d'erreur ainsi que de l'ensemble des erreurs sur les données du TOR. D'après leurs observations, les auteurs concluent que, en pratique, "il pourrait alors y avoir des secteurs pour lesquels les estimations de l'EP seraient plus justes que les estimations du recensement et d'autres (la plupart de ceux qui restent) pour lesquels elles seraient aussi justes sinon presque aussi justes que les estimations du recensement".

Le quatrième article de la section spéciale, "Modélisation de l'erreur d'appariement et son effet sur les estimations de l'erreur d'observation du recensement", de Biemer, traite du problème d'appariement des données de l'EP et de celles du recensement. L'auteur prend trois modèles de plus en plus complexe et examine l'incidence de l'appariement sur les estimations de l'EP. Il parle ensuite des conséquences de ces résultats pour le recensement de 1990.

Les cinq autres articles dans ce numéro portent sur les erreurs dont sont entachées les statistiques sur le commerce extérieur, les problèmes que soulèvent les plans de sondage des enquêtes à usages multiples, la stratification des populations dissymétriques, l'enquête sur le revenu et la participation aux programmes menée par le U.S. Bureau of the Census et un logiciel d'ordinateur personnel pour l'estimation de la variance.

Dans son article "Les erreurs dans les statistiques du commerce extérieur", Ryten traite des sources d'erreur dans ces statistiques et des méthodes permettant de réduire ces erreurs, et il

propose que le niveau d'incertitude soit indiqué lorsque des chiffres détaillés sont communiqués. L'auteur explique les causes de divergence dans les statistiques sur le commerce correspondantes et il analyse l'importance relative de chaque cause. D'après les résultats d'une étude sur les données relatives aux importations et aux exportations tirées de la base de données sur le commerce international de Statistique Canada, l'auteur met sérieusement en doute la comparabilité des données sur le commerce correspondantes aux niveaux détaillés de la classification des marchandises. Il propose un programme qui améliorerait la qualité des statistiques sur le commerce extérieur et explique pourquoi il faudrait donner aux utilisateurs des renseignements concrets sur la qualité des données.

En pratique, il arrive souvent que les données tirées de la plupart des enquêtes servent à plusieurs usages. Cependant, les rapports de recherche et les ouvrages statistiques évitent habituellement le sujet des plans de sondage à usages multiples. Kish se penche sur cette question importante dans son article. Il présente pour commencer une hiérarchie de ces usages, puis il parle des exigences contradictoires auxquelles doivent satisfaire les plans de sondage des enquêtes à usages multiples. Il traite de dix sources d'incompatibilité, dont le choix de la taille de l'échantillon et la répartition de celui-ci en domaines et en strates, le rapport entre les biais et les erreurs d'échantillonnage, le choix des variables de stratification et la continuité des données. Pour chacun de ces problèmes, l'auteur propose des solutions et recommande fortement l'utilisation de plans de sondage qui constituent des compromis plutôt que de plans optimums pour un seul usage. Certaines propositions sont moins rigoureuses que les autres et sont avancées pour stimuler la recherche sur ce sujet.

Dans leur article "Sur la stratification de populations asymétriques", Lavallée et Hidiroglou proposent un algorithme itératif pour la stratification de populations asymétriques dans les cas où l'échantillon est proportionnel à la puissance p (c.-à-d. une répartition de l'échantillon proportionnelle au nombre total d'éléments dans chaque strate élevé à une petite puissance). Ils font une étude empirique dans laquelle la méthode de répartition suggérée est comparée à d'autres méthodes de répartition et où ils utilisent des données tirées des enquêtes annuelles sur le commerce de détail et le commerce de gros menées par Statistique Canada.

La SIPP (Survey of Income and Program Participation/enquête sur le revenu et la participation aux programmes) est une enquête permanente menée par le U.S. Bureau of the Census. Dans son article "Domaines de recherche relatifs à la SIPP", Kasprzyk passe en revue les questions d'ordre méthodologique et statistique que soulève la SIPP. Il se penche sur quatre questions en particulier qui concernent toutes les enquêtes par panel menées auprès de familles et de particuliers. Il s'agit de la conception du questionnaire, de la collecte des données, de l'erreur de réponse et enfin du plan de sondage et de l'estimation pour des variables longitudinales. L'article contient une description des problèmes importants, renvoie à des études où ces problèmes ont été examinés et résume les principaux résultats de ces études.

Dans l'article "Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes", Schnell, Kennedy, Sullivan, Park et Fuller décrivent un programme appelé PC CARP qui a été conçu pour l'analyse des données tirées d'enquêtes complexes. On a trouvé des applications à ce programme, en particulier dans plusieurs pays en voie de développement. Les caractéristiques et les fonctions du logiciel sont décrites brièvement.

Le rédacteur en chef

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 3–18
Statistique Canada

Les erreurs dans les statistiques du commerce extérieur

JACOB RYTEN[1]

RÉSUMÉ

Malgré la facilité relative de l'étude des erreurs dans les statistiques du commerce extérieur, il y a eu peu de tentatives pour quantifier leur taille, leur origine, leur distribution et leur évolution dans le temps. Les décideurs et les négociateurs commerciaux n'ont que des notions très limitées de l'imprécision de ces statistiques, en dépit de leur niveau de détail poussé. L'auteur s'est servi de la banque de données du commerce mondial mise au point à Statistique Canada pour étudier et quantifier les divergences qui existent dans les statistiques du commerce extérieur.

MOTS CLÉS: Commerce extérieur; balances du commerce bilatéral; erreurs.

1. INTRODUCTION

On examine ici quelques-unes des causes d'erreurs dans la statistique du commerce extérieur, les difficultés que pose leur détection, les façons de réduire l'imprécision des chiffres détaillés et un projet d'amélioration de la qualité des données.

Depuis l'ouvrage d'Allen et Ely (1953), publié il y a trente-cinq ans, la question des erreurs dans la statistique du commerce extérieur n'a guère soulevé l'intérêt des auteurs. On s'est intéressé quelque peu aux questions comptables, c'est-à-dire les inclusions et les exclusions, la définition des limites, l'évaluation, etc. (Nations Unies 1982), mais principalement, à la classification. En fait, l'un des plus importants changements apportés à la classification du commerce vient tout juste d'être mis en pratique (Nations Unies 1986) afin de rendre les données du commerce extérieur plus comparables entre pays. Mais, peut-être parce que ces statistiques utilisent une comptabilité complète de toutes les transactions sur les marchandises qui franchissent les frontières au cours d'une période donnée et parce que cette comptabilité est appliquée par un organisme réglementaire, à savoir l'administration douanière, on a tendance à penser en général qu'il ne reste plus guère d'erreurs mesurables. L'absence d'une analyse des erreurs de ces statistiques vient confirmer cette idée.

Régulièrement, on constate, en particulier dans le cas des bureaux de statistique des organismes internationaux, qu'il y a une erreur grave dans la comptabilisation du commerce entre deux pays. À sa dix-huitième session, la Commission statistique des Nations Unies (1974) a été officiellement notifiée du rapprochement de la statistique du commerce entre les États-Unis et le Canada. Ce rapprochement était motivé par la détection de quelques différences embarrassantes de la balance du commerce bilatéral entre les deux pays. Ensuite, et à diverses occasions, on a examiné des problèmes mettant en cause Singapour et la Malaisie, Singapour et l'Indonésie et divers pays hors CÉE et les Pays-Bas les organismes internationaux qui s'intéressent plus particulièrement aux questions du commerce. De plus, les pays qui estimaient qu'ils perdaient le contrôle de la qualité de leur statistique du commerce extérieur,essentiellement les pays du Tiers monde, ont tenté de reconstituer leurs propres chiffres en se référant à ceux de leurs principaux partenaires commerciaux. Mais rien ne prouve qu'aucune de ces manifestions d'inquiétude se soit traduite par un programme systématique de détection, de mesure et de réduction des erreurs dans les statistiques.

On ne relève que peu d'explications possibles évidentes pour ce manque d'initiative, à part celle de la croyance qu'il n'y avait aucune erreur. La statistique du commerce extérieur est l'une

[1] Jacob Ryten, Statisticien en chef adjoint, Statistique Canada, 13-B8 Immeuble Jean Talon, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

des très rares où il est possible de procéder à une comparaison de deux mesures de la même transaction calculée pratiquement au même niveau de détail avec les mêmes méthodes par deux enquêteurs indépendants. Les différences qui résultent lors de l'établissement de ces comparaisons ont été mentionnées dans la littérature consacrée à ce sujet, qui remonte au moins jusqu'à la première guerre mondiale (Coats 1926). Mais ces études n'ont pas donné lieu à des projets d'incorporation des résultats de ces comparaisons dans tout document sur la qualité des statistiques. Un des obstacles à des comparaisons systématiques de ce genre pourrait être la quantité de calculs en cause et les coûts. Un autre pourrait être le degré de connaissance nécessaire des systèmes statistiques correspondants qui, en plus d'être présentés dans certains cas dans une langue étrangère, comprennent habituellement des dispositions administratives et légales très précises qui ne sont pas comparables d'un pays à l'autre.

Les obstacles à des comparaisons systématiques ont été surmontés dans une certaine mesure à Statistique Canada, où l'on a établi une base de données du commerce mondial. Cette banque contient des statistiques détaillées du commerce des pays qui déclarent des données sous une forme lisible par une machine au Bureau de statistique des Nations Unies (BSNU). Les pays membres des Nations Unies conviennent, en vertu des conditions d'appartenance à l'Organisation, de déclarer un certain nombre de statistiques-clés au BSNU de la manière que le Secrétaire général des Nations Unies le précise. Ces statistiques comprennent celles du commerce extérieur, ventilées par pays et par produit, et dans ce dernier cas, de façon complète grâce à la Classification type pour le commerce international (CTCI) ou son équivalent, la Nomenclature du Conseil de coopération douanière (NCCD). Les rapports annuels sous une forme lisible par une machine remontent jusqu'au début des années 60.

La base de données du commerce mondial a été créée afin d'aider les négociateurs canadiens qui participent à la série actuelle des réductions multilatérales des tarifs, et aussi pour aider les exportateurs et les importateurs canadiens à mieux comprendre les marchés et les fournisseurs avec lesquels ils traitent. L'inconvénient de la base de données est qu'elle n'est pas complète. Les économies à planification centralisée ne déclarent aucune donnée ou ne fournissent que des données très agrégées. Un grand nombre des pays du Tiers monde accusent de graves retards pour le traitement de leurs dossiers douaniers, ce qui signifie que beaucoup de données manquent pour les années récentes. Tous les pays n'utilisent pas la même édition de la CTCI, et il existe encore pas mal de différences dans les concepts et les définitions utilisés par les différents pays.

Mais ces inconvénients sont neutralisés par le fait que les calculs en cause pour la comparaison de la statistique du commerce sont maintenant plus pratiques, qu'une très importante proportion du commerce mondial ne touche que les pays occidentaux et qu'elle fait l'objet d'une déclaration courante, laquelle utilise progressivement des concepts plus comparables. Si l'on prend en compte tous ces éléments, une base de données du commerce mondial pourrait servir à présenter les résultats d'une comparaison de la statistique commerciale correspondante, ce qui, à son tour, pourrait aider les organismes statistiques à s'intéresser davantage aux forces et aux faiblesses de leurs données sur les importations et les exportations de marchandises. C'est une condition nécessaire pour améliorer la fiabilité de la statistique du commerce. Compte tenu de l'attention qui est actuellement accordée à ces données, les organismes statistiques du monde entier feraient bien d'apporter les améliorations nécessaires que révèlent les comparaisons bilatérales des données correspondantes, même si cela n'est possible que de façon progressive.

Dans les sections suivantes, on examine les principales causes de divergence de la statistique correspondante et les mesures qui pourraient être prises pour estimer leur importance relative dans des cas particuliers.

2. LES DOSSIERS DES OPÉRATIONS COMMERCIALES: ERREURS ET DIFFÉRENCES DANS LES DOSSIERS CORRESPONDANTS

À la base de deux dossiers commerciaux correspondants, on retrouve dans la plupart des cas une seule transaction documentée. Un exportateur a conclu une vente et facturé l'acheteur en conséquence. Cette facture va probablement contenir des renseignements essentiels sur

la transaction, qui comprennent une description du, ou des, produits vendus, la valeur et la quantité qui y correspondent, les conditions de la vente, une identification de l'acheteur et de la résidence de celui-ci, ainsi qu'une date à laquelle la transaction a eu lieu ou aura lieu. Ce dossier produit un certain nombre de dossiers connexes, les uns obtenus par la transformation des renseignements de base sous une certaine forme prescrite, les autres, par le raccordement de ces dossiers à des dossiers connexes. On peut citer comme exemples de ce dernier cas une description de la façon dont les produits échangés passent du lieu de vente au lieu d'achat et le coût de cette opération, le coût de l'assurance de la livraison, les montants à facturer aux deux parties à la transaction à cause des droits, les taxes de vente, les frais consulaires, etc., et naturellement, la forme et la date de règlement de l'achat.

Les transformations de ces renseignements de base se rattachent aux conventions touchant la façon dont ces renseignements sont enregistrés et la documentation des différentes étapes de la transaction dans le temps. Ces transformations ne sont pas uniformes d'un pays à l'autre. Les conventions qui les régissent sont soit exposées dans la législation douanière, soit dans les règlements administratifs qui régissent la tenue des dossiers douaniers. Ils donnent lieu aux documents qui constituent la base de la statistique du commerce extérieur. Le pays de vente conserve un ensemble de documents, le pays d'achat, l'autre. Dans la pratique, ces documents vont différer en dépit du fait qu'ils se rattachent en principe et dans les faits à la même transaction commerciale.

D'abord, ces documents diffèrent dans le temps. Même dans le cas de pays limitrophes, ou lorsque le transport aérien intervient, les différences temporelles ne sont pas négligeables. Elles surgissent parce que l'ensemble des liens qui constituent la transaction est long, puisqu'il s'agit de transporter la marchandise au point de départ du transporteur international, de l'entreposer avant le départ du transporteur international, de l'amener au point de destination, de l'entreposer en attendant de remplir les formalités douanières, et pendant toutes ces diverses étapes, de remplir les documents à différentes étapes et les enregistrer en fonction des différentes conventions en vigueur.

De plus, dans un pays la date de la transaction peut être inscrite comme étant celle au moment où la facture est reçue dans le pays importateur, et dans un autre, au moment où les autorités douanières perçoivent les droits. Dans un pays, l'enregistrement de la valeur de l'achat peut comprendre tous les coûts du transport international et de l'assurance, alors que dans un autre, ces coûts peuvent être distincts. Ensuite, les unités de déclaration des quantités peuvent entraîner des incohérences. Par ailleurs, dans un pays la transaction peut être imputée non pas au pays où la facture a été émise, mais plutôt à celui où le produit a été cultivé, extrait ou fabriqué, tandis que dans un autre, ce sera le lieu de résidence du vendeur qui décidera de l'identification du pays. Les situations politiques peuvent également toucher la façon dont un pays est identifié dans les dossiers. Les règlements douaniers peuvent introduire un biais dans la façon dont les importations ou les exportations sont enregistrées. Enfin, il y a des erreurs de codage et de traitement des données. Dans les sections qui suivent, on trouvera d'autres renseignements sur ces divers facteurs.

i) *Différences entre les dossiers des exportations et des importations: la datation*

Les autorités douanières vont habituellement classer les dossiers de diverses façons: selon le pays d'origine, selon l'identification de l'entreprise importatrice ou de son agent et selon la date de réception. Mais toute opération d'importation comporte au moins quatre éléments fondamentaux, qui peuvent tous être enregistrés, mais dont un seul sera retenu comme date pour l'extraction et les statistiques. Le choix de la date ne fait pas l'objet d'une normalisation statistique, mais dépend plutôt de la façon dont les Douanes considèrent leur fonction principale et de la capacité technique de stocker d'autres données. Il est évident que si un pays considère comme date d'enregistrement des exportations le moment auquel l'organisme expéditeur prépare un document d'exportation, et si le pays d'importation choisit comme date d'importation celle à laquelle tous les droits et autres frais doivent être réglés, l'écart possible entre l'enregistrement des exportations et celui des importations correspondantes atteint un maximum.

ii) *Différences entre les dossiers des exportations et des importations: les valeurs*

Les différences de valeurs ont pendant longtemps empêché des comparaisons systématiques, et il est bon de les revoir et d'évaluer leur importance relative. L'évaluation de la transaction, c'est-à-dire le prix auquel elle est enregistrée aux fins douanières, est essentielle. Un grand nombre de pays, sinon la plupart, enregistrent la valeur d'une importation en y incluant le coût du transport international et de l'assurance rattaché à la livraison. La plupart des pays enregistrent la valeur des exportations correspondantes sans ces éléments. Il y a d'autres écarts: certains pays incluent des éléments des coûts du transport intérieur et de l'assurance, tandis que d'autres excluent les coûts portuaires du transport international. Mais ces différences ne représentent qu'une augmentation minime de la difficulté de comparer les dossiers correspondants. Les transactions faisant intervenir des partenaires commerciaux reliés, comme dans le cas des entreprises multinationales, posent un problème d'évaluation qui est résolu différemment selon les pays. Il est possible que cette cause de différence finisse par supplanter toutes les autres au cours des années à venir.

iii) *Différences entre les dossiers des exportations et des importations: le pays*

La question du pays créditeur peut introduire quelques-unes des plus bizarres différences dans tout programme systématique de comparaison. En tant qu'exportateur, un pays peut considérer comme exportation toute vente de produit qui doit franchir ses frontières douanières pour atteindre sa destination, indépendamment du fait que ce produit a été sensiblement transformé ou vendu exactement en l'état qu'il a été acheté auprès d'un autre pays. Cependant, comme importateur, un pays peut décider d'imputer un achat au pays où la dernière transformation appréciable (habituellement "appréciable" a une définition précise selon la loi) a eu lieu. Par conséquent, si l'on prend le cas de trois pays hypothétiques A, B et C, où A a exporté des biens à B et B a exporté les mêmes biens (peut-être après transformation) à C, les statistiques peuvent être enregistrées de façons très diverses avec des conséquences différentes, comme le montre le tableau ci-dessous.

Les symboles x et m dénotent respectivement la valeur des exportations vers le pays partenaire et des importations de celui-ci (deuxième lettre majuscule) telle qu'elle est enregistrée par le pays déclarant (première lettre majuscule).

Nous avons donc:

A_xB = Valeur des exportations de A à B enregistrée par A
A_mB = Valeur des importations de B à A enregistrée par A

	Enregistrées comme exportations		Enregistrées comme importations	Conséquence
i)	$A_xB + B_xC$	—	$B_mA + C_mB$	Cohérence et exhaustivité
ii)	$A_xB + B_xC$	—	$B_mA + C_mA$	Sur-imputation de A par les importateurs
iii)	$A_xC + B_xC$	—	$B_mA + C_mB$	Sur-imputation de C par les exportateurs
iv)	A_xC	—	C_mA	Cohérence mais non-exhaustivité
v)	A_xC	—	C_mB	Pas d'imputation à A par les importateurs
vi)	A_xB	—	C_mA	Pas d'imputation à C par les exportateurs.

Les différentes lettres signifient que certains pays déclarants imputent leurs exportations au premier pays de destination connue, et d'autres, au dernier pays de destination connue, que certains pays importateurs imputent les importations au pays d'origine et d'autres au pays de consignation, et que certains pays exportateurs comptent comme exportations tout ce qui sort de leur territoire national, quel que soit le degré de transformation des biens en cause. Les différences de ces approches ne sont pas négligeables à une époque où l'on parle d'accords de libre-échange, d'union douanière, de zones de libre-échange et d'autres accords pour stimuler le commerce transfrontalier. Dans chaque cas, une convention statistique distincte est nécessaire pour tenir compte de l'effet de l'accord sur la comptabilisation douanière. Créditer les pays partenaires de façon incohérente n'est qu'une cause de divergence dans les comparaisons bilatérales ou multilatérales. L'autre est due au manque d'uniformité de la classification géographique.

En fait, un grand nombre de pays incorporent leur position en matière de politique extérieure dans leurs classifications géographiques types. Par conséquent, des définitions géographiques non cohérentes des pays partenaires entraînent des différences. La plupart des pays d'Amérique latine traitent Porto Rico comme une origine ou une destination distincte des États-Unis. Pratiquement chaque pays membre de l'OCDÉ traite différemment les pays partenaires d'Afrique. Certains les regroupent selon leurs origines coloniales, tandis que d'autres le font selon le voisinage géographique. Des incohérences semblables se produisent dans le cas des îles des Caraïbes et du Pacifique sud. L'union économique d'Afrique du Sud est traitée dans les statistiques d'une façon qui traduit souvent l'opinion du pays déclarant d'un embargo sur les relations commerciales avec l'Afrique du Sud proprement dite. De plus, tous les pays n'enregistrent pas les changements de statut politique de leurs partenaires commerciaux avec le même zèle, et pour cette raison tous les pays ne suivent pas les nouveaux pays indépendants aussi rapidement que cela serait souhaitable pour effectuer des comparaisons statistiques.

iv) *Différences entre les dossiers des exportations et des importations: l'administration douanière*

Une autre différence importante surgit du fait que les administrations douanières accordent une attention moindre aux exportations que leur mandat ne l'exige pour les importations. La déclaration des livraisons d'exportations individuelles peut être consolidée pour des considérations de réduction de la paperasserie et alignée sur les bordereaux d'expédition ou les autres documents de transport des transporteurs. Dans les cas des importations, le but visé est d'obtenir une déclaration suffisamment détaillée pour permettre aux Douanes de prélever les droits et autres taxes corrects. Une conséquence est que, dans le cas des exportations, les articles de faible valeur d'une livraison complexe auront probablement plus de chance d'être classés sous la même rubrique que les principales composantes, alors que dans le cas des importations, ils risquent d'être classés de façon indépendante.

Cette différence d'intérêt que l'on peut attribuer au mandat de l'administration douanière a d'autres conséquences importantes sur la qualité des documents d'exportation et d'importation. D'une part, il est évident que l'ampleur de la sous-déclaration des exportations qui caractérise les exportations terrestres des États-Unis vers le Canada ne se limite pas à l'Amérique du Nord. Il y a presque vingt ans, le Royaume-Uni a entrepris un vaste programme qui consistait à rapprocher les bordereaux d'expédition aux documents d'exportation à cause d'un taux de sous-déclaration soupçonné d'environ un à deux pour cent du total. Par contre, on estime que la description des produits exportés est non-biaisée, à moins qu'il ne s'agisse de livraisons illégales, tandis que la description des biens importés pourrait être faussée dans le but de minimiser les taux des droits imposables.

En plus de ces raisons, qui s'expliquent par les différentes transformations juridiques et administratives que subit le dossier original, il y en a d'autres qui sont plus variables et plus sélectives quant aux dossiers visés. On peut citer comme exemples le traitement des livraisons de faible valeur, qui sont définies comme étant inférieures à différents seuils exclus, inclus ou échantillonnés à des taux différents, et le traitement des produits qui contiennent une proportion

de services élevée tels que les bandes audio et vidéo enregistrées, les plans des architectes, les logiciels enregistrés sur bandes magnétiques, les réparations et l'entretien, etc.

v) *Différences entre les dossiers des exportations et des importations: le codage et le traitement des données*

Pratiquement toutes les catégories d'information qui figurent sur les dossiers principaux tenus par les Douanes traduisent l'application d'une classification ou d'un code à une situation réelle. La cohérence du codage peut être garantie par des décisions sur les cas limites et en veillant à ce que ces décisions prises ensemble constituent en quelque sorte une jurisprudence, c'est-à-dire un ensemble de décisions à prendre, accessibles aux codeurs et qui régiraient ces derniers. Mais le seul organisme central réglementaire est le Secrétariat du Conseil de coopération douanière, qui se trouve à Bruxelles. Les pays membres ne peuvent s'y adresser sur une base quotidienne, et ses décisions ne peuvent dépasser un certain niveau de généralité. Pour cette raison, il y a des différences systématiques dans l'interprétation et l'application de codes normalisés parfois à l'intérieur d'un même pays, et à plus forte raison entre pays.

Il y a par ailleurs des incohérences imputables aux erreurs au moment du traitement des données et aux systèmes mis en place pour réduire leur impact. Ainsi, il y a des erreurs d'interprétation de la législation douanière et lors du codage des sources de renseignement qui se glissent au moment où les importateurs ou les exportateurs informent leurs autorités respectives d'une livraison imminente, des erreurs au moment de la saisie des données et des erreurs de codage au Bureau de statistique. La protection habituelle contre ces erreurs est la mise en place de systèmes de revue et de contrôle qui utilisent à des degrés différents une inspection et une revue manuelles et la détection et l'imputation informatiques. Même s'il est très probable qu'il y ait d'autres causes d'incohérences, celles mentionnées ci-dessus sont les plus fréquemment citées, depuis qu'elles ont été décrites pour la première fois (Coats 1926), et elles sont probablement les plus importantes dans le cas des différences entre les chiffres correspondants.

vi) *Différences entre les dossiers des exportations et des importations: les quantités, variable spéciale*

Contrairement aux valeurs, les quantités déclarées ne sont pas touchées par l'inclusion des coûts de transport et elles ne sont pas biaisées afin de minimiser les créances fiscales, cependant, si des valeurs sont mal codées dans les catégories à droits plus faibles, les quantités correspondantes vont suivre. Malheureusement, il y a d'autres problèmes associés à l'enregistrement et à l'utilisation des quantités qui réduisent considérablement la valeur de ces statistiques lors de la détection des erreurs. Ainsi, les quantités peuvent s'appliquer soit à une livraison complète, auquel cas elles sont habituellement exprimées comme un poids brut, soit à une marchandise en particulier, auquel cas elles sont exprimées soit en poids nets, ou en toute autre unité appropriée (longueur, surface, volume) y compris, dans le cas des marchandises complexes, le nombre d'unités.

Alors que les mesures de la quantité, en poids bruts ou nets sont comparables entre pays, leur utilisation se trouve limitée par l'hétérogénéité des livraisons en cause. Les quantités exprimées en d'autres unités se trouvent limitées par la diversité des unités utilisées, et ce qui est plus important encore, par le fait qu'elles ne peuvent être agrégées dans la classification par marchandises, et les niveaux auxquels elles s'appliquent sont beaucoup trop détaillés pour une comparaison entre pays compte tenu de nos connaissances actuelles. Ces unités trouvent également une utilisation dans la comparaison du commerce des matières brutes, en particulier si, avec les valeurs, on les utilise pour suivre les fluctuations des valeurs unitaires. En fait, on a proposé à la dix-huitième session de la Commission statistique des Nations Unies (Rapport du Secrétaire Général 1974) une étude internationale des erreurs de commerce utilisant principalement la comparaison des valeurs unitaires. Cependant, les pays membres n'ont pas estimé que les avantages éventuels justifiaient le coût prévu. À l'heure actuelle, la base de données du commerce mondial de Statistique Canada n'inclut pas des données sur les quantités, de sorte que les applications de la statistique des quantités n'ont pas encore été étudiées.

3. UN PROGRAMME DE MESURE DES ERREURS

Les causes des erreurs sont connues depuis longtemps (Coats 1926). Au début des années 70, une tentative valable de quantification a été faite lors du premier projet de rapprochement entre les États-Unis et le Canada. Mais jusqu'à présent, c'est une très importante proportion de ce que l'on connaît des erreurs dans les statistiques du commerce extérieur, et il est évident qu'elle se trouve limitée par le fait qu'il s'agit du commerce entre deux pays limitrophes, et entre ces deux pays seulement. Étant donné que les bases de données internationales comme celle du Canada vont probablement devenir plus répandues et qu'elles utiliseront divers logiciels analytiques, le moment est venu de s'interroger sur ce que l'on pourrait faire pour améliorer la statistique du commerce, ou en l'absence de toute amélioration, au moins pour informer les utilisateurs des limites des données du commerce extérieur. Il est actuellement peu probable, compte tenu des renseignements descriptifs disponibles, que les utilisateurs d'un pays vont se rendre compte du degré d'erreur des tendances à long terme de la statistique du commerce, ou de la mesure dans laquelle les mouvements mensuels de leurs balances commerciales nationales respectives se trouvent affectés, et, ce qui est plus important encore, dans quelle mesure les renseignements au niveau détaillé de marchandises sont susceptibles de contenir des erreurs.

Il est évident que le flux A_xB pourrait être le même que B_mA, tant que toutes les livraisons et leur enregistrement sont instantanés, que la base d'évaluation est la même pour les deux partenaires pour la même transaction, que les règles d'inclusion et d'exclusion sont les mêmes, qu'il n'y a pas de différences conceptuelles (géographiques, comptables ou résultant des dispositions douanières) et qu'il n'y a pas d'erreurs (de codage ou de champ). On inclut dans les «erreurs» les interprétations cohérentes des protocoles de classification d'un pays qui pourraient être contestés par d'autres pays ou par le Conseil de coopération douanière.

En principe, toutes les causes de différences autres que des erreurs devraient être résolubles, bien que la mesure de l'importance relative des différentes sources puisse être difficile dans la pratique. Une étude des différentes causes ou des facteurs est utile si l'on veut étudier la façon dont leur effet peut être pris en compte dans toute comparaison. De ces facteurs, le transport est probablement le moins difficile à traiter et presque certainement, le moins difficile à aborder. Il y a un certain nombre de pays comme les États-Unis, où les importations sont mesurées de deux façons, avec et sans le transport. En principe, les renseignements servant à estimer le coût de l'assurance et du fret existent habituellement. Les importateurs doivent, en vertu de la loi, informer leurs autorités douanières de toutes leurs dépenses au titre d'un achat à l'étranger, et les deux catégories générales de dépenses sont celles qui sont imposables (habituellement, les dépenses reliées au produit proprement dit, y compris l'emballage, ou les fils ou la monture) et toutes les autres (y compris celles reliées au transport, à l'assurance et au financement de l'importation). Par conséquent, s'il fallait exécuter une étude des coûts de transport, il y aurait des dossiers administratifs qui pourraient être reliés aux dossiers correspondants du commerce. Il y a un grand nombre de problèmes techniques reliés à la façon dont les renseignements sur le fret et l'assurance pourraient être attribués à chaque produit dans le cas des livraisons complexes, mais il y a des suggestions pour traiter ces problèmes (Ryten 1983).

On pourrait également en principe entreprendre une étude des différences de datation dans le contexte d'un flux commercial particulier entre deux pays. Dans le cas du rapprochement de la statistique du commerce entre les États-Unis et le Canada, les estimations ont utilisé des rapprochements réels de documents, ce qui a permis de comparer les dates et les délais moyens entre les exportations et les importations correspondantes. Mais il y a des méthodes moins onéreuses pour obtenir des estimations approximatives qui sont également moins limitées sur le plan de l'accès à des dossiers confidentiels et qui sont suffisamment efficaces pour permettre le calcul d'intervalles généraux de différences de datation par point de sortie et d'entrée, par mode de transport et par produit.

Ensemble, les estimations des différences de datation et la différence entre le coût de l'assurance et du fret et les évaluations franco à bord peuvent être exprimées sous la forme suivante:

$$A_mB(k) = B_xA(k) + A\,(\text{c.a.f.})\,B(k) + \theta + e$$

où $A_mB(k)$ est le flux d'importations de la marchandise k du pays B au pays A, enregistré par le pays A; $B_xA(k)$ est le flux correspondant déclaré par le pays B; A (c.a.f.) $B(k)$ est l'estimation des coûts de transport et d'assurance pour ce flux commercial, calculée à partir des dossiers du pays A; θ est un ajustement de période et e un terme d'erreur qui comprend tous les biais et erreurs aléatoires qui caractérisent la statistique des importations et celle des exportations. On suppose que toutes les autres causes de différence (géographie, inclusions et exclusions, livraisons de faible valeur, etc.) ont été réglées soit par correction, ou, de préférence, par l'exclusion de toutes les transactions qui pourraient être entachées de ces facteurs des fichiers de comparaison. Dans le temps, l'erreur moyenne devrait tendre vers zéro, et par conséquent, plus la période au cours de laquelle la comparaison est faite est longue, et plus proche est le niveau du taux moyen de variation des chiffres comparés. Si une telle comparaison devait soudainement donner des résultats pervers, on aurait alors la preuve évidente d'une détérioration de la qualité d'au moins un des deux termes de la comparaison.

3.1 Analyse utilisant la mini-base de données du commerce mondial

À des fins d'analyse, on a créé une mini-base de données à partir des données du commerce mondial afin d'entreprendre l'étude de quelques-uns de ces effets. Elle recouvre les trois principaux blocs commerciaux du monde occidental: la CÉE, définie dans ce cas par l'exclusion du Portugal et de l'Espagne, l'Amérique du Nord (Canada et États-Unis) et le Japon. En plus d'être plus simple d'utilisation en raison du nombre réduit d'enregistrements, elle évite le problème des déclarations en retard (principalement les pays du Tiers monde) et de la non-déclaration (principalement les pays à économie centralisée). La mini-base de données comprend les données sur les exportations et les importations de chaque pays, ventilées par la CTCI (jusqu'au niveau de détail à quatre chiffres) et par pays partenaire de 1978 à 1985. En plus des pays participants, elle comprend deux agrégats, la CÉE et l'Amérique du Nord. Alors que la base de données du commerce mondial utilise un certain nombre d'imputations pour simplifier l'analyse, la mini-base ne comprend que les données que les pays membres ont déclarées au BSNU, après que ce dernier eût regroupé des catégories commerciales considérées secrètes par le pays déclarant et converti les codes non normalisés déclarés par les pays en codes normalisés CTCI. Aucune de ces transformations ne devrait affecter sensiblement les conclusions résultant de la base de données.

Le groupement des pays dans la mini-base de données pose quelques problèmes statistiques. Les États-Unis déclaraient leurs importations au BSNU sur la base c.a.f., mais le Canada déclare ses importations f.o.b. Alors que les États-Unis créditent leurs pays partenaires en fonction de l'origine des biens importés, le Canada les déclare selon le pays de consignation, à l'exception des importations provenant des pays d'Amérique latine. Ceci ne serait pas un problème trop grave s'il n'y avait le fait que les États-Unis sont parfois crédités pour des exportations se dirigeant vers le Canada. Par conséquent, bien que l'addition des deux pays devrait améliorer l'appariement des flux correspondants, les différents systèmes d'enregistrement le rendent beaucoup plus difficile. On espère que cet inconvénient disparaîtra lorsque les importations f.o.b. des États-Unis seront versées dans la base et lorsque les importations canadiennes selon l'origine remplaceront les importations selon la consignation pour le plus grand nombre d'années antérieures possibles.

Dans le cas des pays de la CÉE, le rôle essentiel des Pays-Bas comme point d'entrée pour le continent européen rend les comparaisons difficiles. La zone douanière du port de Rotterdam agit non seulement comme un centre de distribution géant, mais également comme un entrepôt pour les pays desservis. L'exportateur à l'extérieur de la CÉE peut donc ne pas donc connaître à quel pays en particulier la vente est faite, et tout ce qu'il saura c'est que l'entreposage aura lieu à Rotterdam; pour cette raison, il créditera les ventes aux Pays-Bas. Mais l'importateur ultime est lié par la règle de l'origine pour attribuer l'achat au bon pays. Dans le cas des Pays-Bas, selon leurs dossiers, aucune transaction portant sur des biens avec franchissement de leurs frontières douanières n'a eu lieu. Le pays a simplement vendu des services portuaires et d'entreposage à l'une des parties.

Si les Pays-Bas ne désservaient les autres membres de la CÉE que comme un port, la création d'un total pour la CÉE suffirait à améliorer les comparaisons. Mais d'autres pays, en particulier la Suisse et l'Autriche, bénéficient également des ports et des terminaux à conteneurs hollandais. Ceci complique quelque peu le problème, puisque, dans le cas de la Suisse, l'importateur peut utiliser la règle de l'origine dans le cas des Pays-Bas lorsqu'il y a une consolidation des importations provenant d'un grand nombre d'origines, ou bien une transaction de valeur ajoutée à l'extérieur de la zône douanière de Rotterdam peut ne pas être déclarée comme du commerce extérieur hollandais de marchandises.

Un autre obstacle à l'interprétation est posé par les deux Allemagnes, puisque l'une s'abstient de déclarer ses importations au BSNU, tandis que l'autre ne considère pas comme des exportations les transactions qu'elle a avec son homologue oriental. Cela veut dire qu'il y a des exportations supplémentaires de la CÉE qui n'ont pas de dossiers d'importation, et plus particulièrement, qu'il y a des transactions commerciales non déclarées entre les deux Allemagne. L'ampleur de cette fuite non déclarée fluctue selon la prospérité relative de l'Allemagne de l'Est et ne peut être évaluée que par l'étude d'autres indicateurs. Il y a également des fuites pour le commerce avec le Japon et qui toucheront les résultats des comparaisons faisant intervenir le Japon et ses partenaires commerciaux. Elles peuvent s'expliquer par des opérations faisant intervenir des succursales de firmes étrangères situées en Asie du Sud-Est. Cependant, l'effet sur les données globales ne devrait pas être important et ne devrait pas reléguer au deuxième plan la valeur de l'analyse utilisant cette base de données.

i) *Comparaison des taux de croissance des statistiques correspondantes*

Parmi les diverses analyses effectuées à partir de la mini-base de données, une a porté sur la comparaison des taux de croissance des statistiques correspondantes pour la période 1978-1985. On a supposé qu'au cours de cette période, l'effet des erreurs et des différences de datation serait suffisamment réduit ce qui ferait ressortir les effets plus permanents. De plus, par l'étude des taux de croissance, il serait possible d'éviter dans une large mesure l'effet d'évaluations différentes. Il est peu probable que la variation du coût de l'assurance et du fret soit suffisamment différente de la variation des prix moyens des marchandises transportées pour modifier sensiblement les taux de croissance sur une période de trois ou de quatre ans. Au moins dans le cas des produits manufacturés, la proportion du fret et de l'assurance dans le coût total est bien inférieure à 10%, comme le montrent les rapports f.o.b./c.a.f des États-Unis. De plus, les coûts du transport ne se rapporteraient qu'au poids et au volume des marchandises transportées. Les coûts d'assurance, qui se rattachent à la valeur, ne représentent pas une proportion appréciable du total des coûts. La substitution du mode de transport a peu de chances de s'ajouter au coût total, sauf dans des circonstances exceptionnelles. Par conséquent, si la variation du coût correspondant était suffisante pour toucher les taux de croissance des importations par rapport aux taux correspondants des importations, les effets seraient tous à sens unique, et leur importance varierait selon la masse moyenne des marchandises transportées.

Ces considérations ne sont qu'en partie confirmées par les faits. Le tableau 1 montre les différences des taux de croissance annuels pour le total du commerce correspondant pour des groupes d'origines et de destinations obtenues à partir du commerce entre les pays de la CÉE, l'Amérique du Nord et le Japon. Même si ces différences sont relativement faibles, elles ne semblent dégager aucune tendance, malgré quelques régularités éventuelles sous-jacentes échappant à une inspection superficielle.

Le tableau 2 présente les taux de croissance de certaines sections de la CTCI entre le Japon et ses deux partenaires commerciaux. Pour simplifier le tableau, on a posé comme principe d'ignorer les flux inférieurs à un milliard de dollars en 1982, puisque de tels flux ne semblent pas être suffisamment stables pour justifier une interprétation.

Les participants aux discussions portant sur les classifications de marchandises comparables sur le plan international ont immanquablement exigé davantage de détails, plutôt que moins. La collecte de statistiques aux fins d'une comparaison internationale a poussé les pays

Tableau 1

Différences des taux de croissance annuels du commerce total correspondant pour le Japon, l'Amérique du Nord et la CÉE, 1978-1985

Pays A - Pays B	Différences des taux de croissance pour la période[1]			Différence exprimée en valeur des exportations en 1982/ arrondie au cinq millions de dollars le plus proche
	1982/78	1985/82	1985/78	
A.N. - CÉE	.6	−.5	−.5	265
A.N. - Japon	−.4	.5	-	-
CÉE - A.N.	−.8	−.7	−.8	365
CÉE - Japon	1.1	1.9	1.5	90
Japon - A.N.	−.7	−.2	−.5	200
Japon - CÉE	−1.2	−.6	−.9	155
Moyenne	.8	.7	.7	

[1] Différence définie comme (taux de croissance en % de A_xB) - (taux de croissance en % de B_mA.
[2] Différence entre A_xB et B_mA arrondie au cinq millions de dollars le plus proche.
Le symbole (-) indique une valeur insignifiante.

Tableau 2

Différences des taux de croissance annuels du commerce total correspondant par section CTCI Japon, en 1978-82 et 1982-85

Section de la CTCI	Japon - Amérique du Nord			Japon - CÉE		
	Taux de croissance annuel en % pour la période[1]		Différence de la valeur des exportations en 1982[2]	Différence de croissance annuel en % pour la période		Différence de la valeur des exportations en 1982
	1982/78	1985/82		1982/78	1985/82	
5. Produits chimiques	.7	−1.6	15	−1.5	.5	5
6. Semi-produits	−2.5	.9	60	1.9	−.5	5
7. Matériel de transport	−1.0	−1.0	275	−2.0	−.7	85
8. Articles manufacturés	1.4	−.9	35	−.8	.8	20

[1] Différence définie comme (taux de croissance en % A_xB) - (taux de croissance en % de B_mA) (A est le Japon).
[2] Différence entre A_xB et B_mA arrondie au cinq millions de dollars le plus proche.

à publier des données bien au-delà du niveau à trois chiffres de la CTCI ou son équivalent. Un certain nombre de pays du Tiers monde publient des données ventilées selon dix chiffres (classification internationale adaptée au pays) et selon le pays. L'étude permet cependant de constater que les flux codés à un chiffre, où il y a rarement eu des controverses, se caractérisent par des différences très considérables lorsqu'on les compare à leurs homologues dès que leur valeur absolue tombe, par exemple, en-dessous de 50 millions de dollars. Au-delà du premier chiffre de la classification, les différences s'accroissent très rapidement.

Le cas des exportations japonaises vers l'Amérique du Nord et les importations correspondantes figurant aux tableaux 1 et 2 mérite qu'on l'examine de plus près. Au point central (1982), ce commerce était évalué à environ quarante milliards de dollars (É.-U.). Le total des importations a augmenté en moyenne de 0.5% par an de plus que les exportations. C'est un montant d'environ 200 millions de dollars par an, pour l'année centrale. Une étude détaillée permet de conclure qu'une partie appréciable de l'explication est attribuable à la section 7 de la CTCI,

Tableau 3

Variation des rapports X/M entre 1978 et 1985 et comparaisons avec les rapports normalisés X/M, les parts des sections CTCI étant supposées constantes[1]

	Amérique du Nord			CÉE			Japon		
	Rapport simple		Rapport normalisé	Rapport simple		Rapport normalisé	Rapport simple		Rapport normalisé
	1978	1985	1985	1978	1985	1985	1978	1985	1985
Amérique du Nord				.90	.91	.92	.85	.86	.69
CÉE	.96	.92	.69				.78	.86	.89
Japon	.95	.98	.91	1.00	.94	.86			

[1] Le rapport simple est $X/M = (A_x B/B_m A)$. Le rapport normalisé utilisant des parts constantes est

$$\text{Rapport simple} = \frac{1}{M_{78}} \sum_{i=0}^{n} \frac{x_{it}}{m_{it}} \cdot m_{i78},$$

où m_{it} = importations courante de la section i de la CTCI ($i = 0, 1, \ldots, n$),
m_{i78} = importations en 1978 de la section i de la CTCI,
x_{it} = exportations courantes de la section i de la CTCI, et
M_{78} = total des importations en 1978.

qui regroupe, entre autres, tous les types de matériel de transport. La différence des taux de croissance dans ce cas-là est de un pour cent par an en moyenne. Il serait intéressant de pousser davantage cette étude afin de déterminer si la divergence se répartit de façon uniforme, ou si elle caractérise une marchandise en particulier.

Quelles qu'en soient les causes, ces comparaisons semblent indiquer que sur un nombre suffisamment grand d'années et pour des parties relativement importantes du total des flux commerciaux, les différences des taux de croissance ne sont pas élevées en termes absolus. Néanmoins, même de petites différences peuvent bouleverser les variations d'une période à l'autre de la balance commerciale globale, en particulier lorsqu'elle est proche de zéro. De plus, lorsque l'on traite avec un partenaire commercial comme le Japon, dont les exportations sont concentrées principalement dans une ou deux catégories à un chiffre de la classification des marchandises, les possibilités de corriger une mauvaise classification systématique sont comparativement peu nombreuses. Il est donc d'autant plus important de comprendre pourquoi le commerce bilatéral, mesuré par les deux dossiers homologues, n'a pas suivi le mouvement.

ii) *Comparaison des rapports des exportations aux importations annuelles*

Un autre type d'analyse est également très révélateur. Tout flux d'importations devrait être égal au flux d'exportations correspondant, plus le coût du fret et de l'assurance, plus un terme qui traduise l'ensemble des différences de concepts, de datation et des erreurs. Alors que la datation et les erreurs se font surtout sentir à court terme, les différences de concepts devraient se révéler être la cause principale à plus long terme. Pour cette raison, si le rapport des exportations annuelles aux importations annuelles varie dans le temps, cela peut s'expliquer par une combinaison des facteurs suivants: une variation des parts des composantes avec un c.a.f. relativement élevé par rapport à celles avec un c.a.f. bas, un changement dans la composition des marchandises avec de petites différences quant aux marchandises et d'importantes différences de datation, un changement dans la proportion du c.a.f. à la valeur totale et d'autres facteurs.

Le tableau 3 présente quelques résultats agrégés de cette analyse. En regard de chacun des flux touchant le Japon, la CÉE et l'Amérique du Nord, il y a trois chiffres: le rapport simple (à pondération de l'année courante) de l'agrégat des exportations à celui des importations en 1978, le rapport correspondant en 1985 et le rapport normalisé pondéré selon l'année de base,

Tableau 4

Variations des rapports simples *x/m* entre 1978 et 1985 comparées aux rapports normalisés *x/m* avec des parts de division CTCI constantes

Exportations de ... à ...	A.N. CÉE	CÉE A.N.	CÉE Japon	Japon CÉE	A.N. Japon	Japon A.N.
Sections CTCI						
0 Aliments	107	102	100	98	98	90
1 Boissons et tabacs	99	100	99	100	99	100
2 Matières brutes	100	100	96	93	100	102
3 Combustibles minéraux	102	117	304	93	103	109
4 Huiles animales et végétales	99	98	107	100	101	92
5 Produits chimiques	102	101	101	97	100	96
6 Produits manufacturés	99	101	99	96	98	100
7 Machines et transport	96	91	95	97	92	100
8 Articles manufacturés divers	100	100	100	98	97	99
9 Transactions diverses	150	176	163	157	86	92

en supposant que les proportions des importations par section au total des importations pour chaque flux de marchandises demeurent constantes depuis 1978. Ces rapports normalisés sont une approximation d'une estimation qui élimine l'impact des variations dans l'ensemble du c.a.f. de la variation du rapport dans le temps. Toute différence entre les deux rapports de 1985 doit donc être attribuée à d'autres facteurs.

L'évolution des rapports dans le temps en raison de l'augmentation de la part des produits hautement manufacturés dans certaines exportations de flux est assez prévisible. Ainsi, les exportations de la CÉE vers l'Amérique du Nord et le Japon, les exportations du Japon vers la CÉE et l'Amérique du Nord devraient inclure proportionnellement davantage d'articles manufacturés. Le rapport qui traduit les variations de composition est donc supérieur au rapport normalisé, car l'importance relative du c.a.f. diminue à mesure que la valeur d'une unité de poids ou de volume augmente.

Mais il y a des exceptions *a priori* à cette certitude dans le tableau. Ainsi, les exportations de l'Amérique du Nord vers le Japon se caractérisent par un très grand écart entre le rapport simple et le rapport normalisé, même si la part des articles manufacturés a augmenté relativement moins.

Le tableau 4 contient une ventilation par section de la CTCI pour les rapports correspondant aux flux commerciaux entre chacun des six couples de blocs commerciaux retracés dans la mini-base de données. Les chiffres indiqués sont des rapports de l'indice simple au niveau de la section à l'indice calculé à partir de la part des importations au niveau de la division. Ils indiquent l'apport de la variation des rapports représenté par les variations de la composition de marchandises. Ce ne sont rien d'autre que des indicateurs, en partie parce qu'ils diminuent d'un niveau dans la classification des marchandises. On ne peut facilement déceler de tendances, il y a en gros autant de cas de dépassement que de rabage.

Pour les flux plus importants, tels que ceux de l'Amérique du Nord vers la CÉE ou de la CÉE vers le Japon, la composition par marchandise est relativement stable, et il n'y a guère de différence entre les rapports à pondération de base et à pondération courante. De plus, ces rapports ne fluctuent guère au cours de la période étudiée. D'autres flux sont plus sensibles à la composition par marchandises, ce qui semble indiquer qu'à des niveaux plus bas de la classification les différences f.o.b./c.a.f. expliquent une petite partie de la variation des rapports *x/m* dans le temps.

On obtient les chiffres du tableau en prenant l'indice qui mesure la variation de chaque section (1 chiffre) du rapport simple X/M de 1978 à 1985, c'est-à-dire (x/m) 1985 divisé par (x/m) 1978, et en le divisant par un indice correspondant dans lequel le rapport normalisé (x/m) pour 1985 a été utilisé et où les rapports de divisions (2 chiffres) ont été agrégés à partir de leurs parts de 1978 dans leur division correspondante. Algèbriquement, le rapport obtenu R_i est:

$$R_i = 100 \cdot M_{i78} \cdot \frac{X_{i85}}{M_{i85}} \div \sum_{j=0}^{n_i} \frac{x_{ij85}}{m_{ij85}} \cdot m_{ij78}.$$

Un chiffre de 104, par exemple, signifie qu'une augmentation de quatre pour cent de la valeur courante des exportations par rapport aux importations correspondantes s'explique par des raisons autres que l'effet des changements de la composition des marchandises sur la composante c.a.f.

3.2 Analyses utilisant la base de données du commerce mondial complète

Possibilités de mauvaise classification par pays et marchandise:

Les tableaux 5 et 6 tirés de la base de données du commerce mondial complète présentent le nombre de cas de mauvaise classification possibles par pays et marchandise. Le tableau 5 contient le nombre de cas pour 1983 pour lesquels il y a des échanges bilatéraux d'une marchandise selon l'un des pays déclarants d'un couple partenaire commercial, mais non selon l'autre. Ce nombre de cas est exprimé pour chaque niveau de détail de la CTCI comme une proportion de tous les cas. Le tableau 5A présente l'impact sur la valeur, là encore pour chaque niveau de la CTCI. En plus de fournir une mesure sommaire de la taille des erreurs, les tableaux donnent aussi une idée de la rapidité avec laquelle le nombre de situations anormales augmente en fonction du niveau de détail de la classification.

Tableau 5

Comparaison des statistiques du commerce extérieur correspondantes en 1983 – Nombre de dossiers[1]

Niveau de détail de la CTCI	Pourcentage ne rapportant pas d'exportations	Percentage ne rapportant pas d'importations	Pourcentage du total
0 (combiné)	11	4	15
1 chiffre	14	7	21
2 chiffres	16	10	26
3 chiffres	19	13	32

[1] Pourcentage du nombre de dossiers de couples partenaires commerciaux où un membre déclare 0 exportations/importations et les autres membres déclarent un autre chiffre

Tableau 5A

Comparaison des statistiques du commerce extérieur correspondantes en 1983 – Valeurs des dossiers[1]

Niveau de détail de la CTCI	Pourcentage ne rapportant pas d'exportations	Percentage ne rapportant pas d'importations	Pourcentage du total
0 (combiné)	.1	-	.1
1 chiffre	.3	.1	.4
2 chiffres	.6	.4	1.0
3 chiffres	1.1	.9	2.0

[1] Pourcentage de la valeur des dossiers des couples partenaires commerciaux, dont un membre déclare 0 exportations/importations et l'autre un commerce non nul

Le symbole (-) indique une valeur insignifiante.

Tableau 6

Comparaison des statistiques du commerce extérieur correspondantes pour deux années choisies

	1979	1983
Nombre de dossiers avec $x > m$ en pourcentage de tous les dossiers	35	32
Valeur des exportations avec $x > m$ en pourcentage du total des exportations	41	42
Rapport x/m pour $x > m$	1.18	1.15
Rapport x/m pour $x < m$	.87	.85

Tableau 7

Rapports X/M en 1985
de trois pays déclarants sélectionnés à neuf partenaires commerciaux

À	De Canada	États-Unis	Japon
C.É.E.	.84	.92	.94
Pays-Bas	1.93	1.34	1.33
Belgique - Luxembourg	1.47	1.51*	1.26
Danemark	1.20*	.74	1.05*
France	.70	.74*	.69*
République fédérale d'Allemagne	.69	.81	.98
Irelande	.55	.78	.72*
Italie	.75	.84	.75*
Royaume-Uni	.74	.86	.89
Grèce	1.00	1.23*	.89*

Le tableau 6 présente les variations entre deux années choisies d'un certain nombre d'indicateurs reliés à des cas où les exportations ont dépassé les importations correspondantes. Même si sur une période de quatre ans il y a eu certains changements dans le nombre de dossiers pour lesquels les exportations ont dépassé les importations ainsi que la valeur unitaire, les changements en question sont mineurs. Assez curieusement, les cas de x/m représentent plus de 40% de la valeur totale du commerce, et comme ce chiffre a augmenté légèrement, le nombre de dossiers qui l'explique a baissé de 10%.

Au tableau 7, on soumet à l'épreuve des faits un certain nombre de prédictions *a priori*. On étudie trois exportateurs déclarants, le Canada, les États-Unis et le Japon, et neuf partenaires commerciaux déclarants, c'est-à-dire les pays membres de la CÉE autres que l'Espagne et le Portugal. Les tableaux donnent la liste des rapports simples x/m de 1985 pour le commerce pays à pays. Toutes choses étant égales par ailleurs, les prédictions suivantes semblent plausibles:

- plus la part de la fabrication d'un flux commercial est élevée, plus le rapport x/m est élevé, ce qui revient à dire que plus le rapport c.a.f./valeur totale est plus petit, et plus la valeur ajoutée incorporée dans une marchandise est élevée. Pour cette raison, le classement par ordre *croissant* des rapports devrait être Canada, États-Unis, Japon;
- dans le cas du commerce avec les pays d'entreposage, soit les Pays-Bas, et dans une moindre mesure, l'Union belgo-luxembourgeoise, le mauvais codage du pays par l'exportateur ne devrait toucher principalement que les livraisons en vrac. Pour cette raison, le rapport x/m par ordre décroissant devrait être Canada, États-Unis, Japon; et
- des rapports x/m supérieurs à un ne devraient s'observer que pour les pays d'entreposage.

Pour trente rapports x/m (en comptant les trois rapports pour la CÉE dans son ensemble), on relève neuf cas (chiffres suivis d'un astérisque dans le tableau) pour lesquels les prédictions ne se vérifient pas. En supprimant les deux chiffres pour la Grèce, parce que les flux commerciaux correspondants sont beaucoup trop faibles, sept rapports ne se comportent pas selon les prévisions, ce qui représente quand même plus de vingt pour cent de tous les cas.

La conclusion essentielle de ces analyses est que toute augmentation du niveau de détail de la hiérarchie de la classification au-delà du niveau combiné rend les comparaisons avec le commerce correspondant très difficiles. Ceci n'est pas incompatible avec les tentatives progressives, faites à la fois au pays et à l'étranger, pour accroître le niveau de détail de la classification par marchandise et augmenter le nombre de ventilations grâce à des variables de classification supplémentaires. Même lorsqu'on les regroupe dans le temps, les transactions de ces cellules détaillées ne correspondent guère avec leurs homologues. Comme on ne peut affirmer que les deux rapports d'une comparaison bilatérale sont simultanément corrects, il est possible que les deux contiennent une erreur appréciable.

4. SIGNALER LES ERREURS AUX UTILISATEURS

Il y a deux questions distinctes. L'une est d'informer les utilisateurs que, contrairement à une opinion répandue, les chiffres du commerce extérieur, et en particulier les chiffres détaillés, peuvent être biaisés. L'autre est de mettre sur pied un programme d'amélioration de la qualité des données du commerce extérieur qui bénéficierait du fait que des mesures correspondantes de la même transaction existent. Voici quelques propositions pour lancer un tel programme.

L'analyse contenue dans cet article prévoit qu'au-delà du niveau à deux chiffres de la classification par marchandise par pays, même annuellement, il n'est pas possible d'accorder une confiance complète aux niveaux ni aux variations d'une année à l'autre. Les utilisateurs auront probablement une réaction négative à cette conclusion, puisqu'ils ont déjà des raisons de mettre en doute le champ des agrégats dans le cas des exportations. Les résultats du programme de rapprochement de la statistique des États-Unis et du Canada ne doivent pas être considérés comme étant limités à ces deux pays. D'autres pays ont à faire face à la même catégorie de problèmes à des degrés divers. La révélation que, en plus de ces lacunes, les données par marchandise au-delà d'un certain niveau ne peuvent être utilisées qu'avec beaucoup de prudence, pourrait se traduire par un changement fondamental de la perception des utilisateurs de la statistique du commerce extérieur.

Mais, si cette mesure n'est pas adoptée, quel qu'en soit le degré d'impopularité, on maintient une croyance qui n'est pas entièrement justifiée par les faits. Les chiffres détaillés du commerce de marchandises servent à divers usages, dont le principal est celui des tarifs. Les discussions sur ces questions utilisent beaucoup les chiffres détaillés, rarement les différences entre les données nationales et correspondantes et aussi rarement la statistique de la consommation intérieure comme mesure de vérification de l'ordre de grandeur que semblent indiquer les données douanières. De plus, dans une autre utilisation de données détaillées des marchandises, on tire des opinions sur des politiques industrielles et régionales, et des mesures peuvent être prises à partir d'éléments de preuve qui, selon cette analyse, ne seraient pas solides. Il est évident qu'il appartient aux organismes statistiques de communiquer aux utilisateurs les lacunes relevées des données afin d'empêcher la généralisation d'une mauvaise utilisation.

5. UN PROGRAMME D'AMÉLIORATION DE LA STATISTIQUE DU COMMERCE EXTÉRIEUR

En plus de fournir aux utilisateurs plus de renseignements concrets sur les erreurs dans la statistique du commerce extérieur, il faudrait mettre sur pied un ou plusieurs programmes destinés à améliorer la qualité de ces statistiques dans le temps. Voici des mesures qui auraient dû probablement être prises il y a quelque temps:

i) la composante c.a.f. des importations doit faire l'objet d'une mesure systématique. Sinon, il ne serait pas possible de comparer les exportations et les importations en général. Les renseignements deviennent disponibles au moment où l'importation est déclarée aux Douanes. Les questions telles que la fréquence et le niveau de détail dépendront des ressources et de l'urgence à améliorer les connaissances des utilisateurs;

ii) il faudrait entreprendre une étude des retards entre les exportations et les importations par catégorie de marchandises et par pays d'origine. Pour que cette étude soit rentable, il faudra probablement faire appel à la collaboration des pays partenaires, mais si cette étude n'est pas prévue dans un avenir rapproché, on pourrait utiliser comme remplacement acceptable les factures commerciales;

iii) à partir de ces deux éléments, il faudrait utiliser une méthode officielle d'estimation des importations correspondantes sur la base des exportations, et totaliser l'erreur d'estimation pour étude ultérieure. Si l'erreur d'estimation n'a pas d'autocorrélation appréciable, les erreurs de codage et connexes pourraient expliquer la différence entre l'importation enregistrée et son estimation statistique. Si, par contre, le terme d'erreur ne répond pas à ces critères, il faudrait en faire l'objet d'une étude ultérieure en collaboration avec le pays partenaire;

iv) il faudrait vérifier les excédents ou les déficits évidents par rapport aux pays qui ont probablement un rôle d'intermédiaire commercial ou d'entreposage. Ainsi, un excédent d'exportation avec les Pays-Bas pour les États-Unis pourrait faire l'objet d'une vérification par rapport aux déficits correspondants avec des pays tels que la République fédérale d'Allemagne ou la France. Les méthodes économétriques peuvent servir à séparer l'effet global des services d'entreposage, même si ces derniers vont servir plus probablement aux marchandises volumineuses et entreposables, des effets de courte durée tels qu'une erreur de codage.

v) pour les marchandises qui sont des valeurs systématiquement des valeurs aberrantes, une fois toutes les corrections faites, parce qu'elles persistent dans le temps ou parce qu'elles touchent tous les pays, il faudrait se servir du Système harmonisé et obtenir l'aide du Conseil de coopération douanière pour l'interprétation de ses notes explicatives.

Il est évident que le lancement d'un tel programme nécessite des préparatifs, les approbations nécessaires et des ressources. Il ne peut être exécuté immédiatement, et la plupart des pays ne vont pas le parrainer tout de suite. Mais les propositions ne devraient pas être laissées de côté, comme cela a été le cas il y a quelque treize ou quatorze ans pour des propositions semblables. On accorde beaucoup trop d'attention aux statistiques du commerce pour prendre le risque de retarder leur amélioration. La comparaison avec les données correspondantes révèle qu'on ne pourra y accorder plus d'attention que si elles sont sensiblement améliorées, ou si leurs analystes deviennent plus conscients des limites des données qui servent à tester leurs hypothèses.

BIBLIOGRAPHIE

ALLEN, R.G.D., et ELY, J. EDWARD (1953). *International Trade Statistics*, New York: John Wiley & Sons.

COATS, R.H. (1926). Canadian Trade Statistics. *Journal of the Canadian Bankers' Association*.

RYTEN, J. (1983, 1984, 1986). Reports on the Reorganization of Bolivian Foreign Trade Statistics (En espagnol). United Nations Development Programme. New York.

UNITED NATIONS (1982). International Trade Statistics, Concepts and Definitions. *Statistical Papers Series* M, 52. New York.

UNITED NATIONS (1986). Standard International Trade Classification. *Statistical Papers Series* M, 34. New York.

UNITED NATIONS STATISTICAL COMMISSION (1974). International Trade Reconciliation Study. Rapport du Secrétaire Général. Dix-huitième session, Genève.

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 19–33
Statistique Canada

Plans de sondage à usages multiples[1]

LESLIE KISH[2]

RÉSUMÉ

La plupart des enquêtes ont de nombreux usages et nous proposons dans cet article une hiérarchie de ces usages en six niveaux. Toutefois, la plupart des théories et des ouvrages statistiques mettent l'accent sur les enquêtes à usage unique pour éviter la complexité des plans de sondage à usages multiples et les problèmes que ces plans soulèvent. Nous exposons tout d'abord dix sources d'incompatibilité entre les usages de ces plans, puis nous posons des problèmes et donnons des solutions pour chaque cas. Heureusement, des compromis et des solutions communes sont possibles puisque la plupart des optimums sont très peu prononcés et que la plupart des "exigences" relatives à la précision sont en réalité très souples. Il est préférable de parler d'usages multiples et de composer avec eux que de se limiter à quelques usages choisis arbitrairement; en outre, la venue de l'informatique a facilité l'application des plans de sondage à usages multiples.

MOTS CLÉS: Répartition entre les domaines; erreurs quadratiques moyennes; répartition à usages multiples; plan à usages multiples; répartition optimale; échantillons périodiques; taille de l'échantillon.

1. INTRODUCTION

Dans la plupart des cas, la planification des enquêtes révèle plusieurs usages, puis beaucoup d'autres usages apparaissent au moment de l'analyse des données, et d'autres encore au moment de leur interprétation et de leur utilisation. Toutefois, la nature polyvalente de la plupart des enquêtes tend à être dissimulée par des analyses de plans d'enquête trop simplifiées, qui ne tiennent compte que d'une variable. Cette lacune ressort particulièrement dans le cas des enquêtes par sondage, auxquelles nous nous intéresserons dans le présent article, mais elle est observable aussi dans le cas d'autres plans statistiques comme les études expérimentales et les études d'évaluation.

En pratique, les enquêtes visent habituellement plusieurs objectifs. Pourquoi alors la théorie de l'échantillonnage ne traite-t-elle pas les plans de sondage à usages multiples? Parce qu'une théorie des plans à usages multiples serait trop complexe et que la théorie de l'échantillonnage est déjà assez complexe comme cela; nous relèverons plus loin des exceptions particulières. Par surcroît, les descriptions de plans de sondage que nous lisons couramment tendent à refléter le caractère "prestigieux" de la théorie de l'échantillonnage à usage et à variable unique au lieu de représenter fidèlement les nombreux compromis auxquels il faut arriver dans la réalité. De nombreux plans de sondage courants (notamment, la méthode d'échantillonnage avec probabilités égales) servent probablement à de nombreux usages de façon efficace; selon nous, la conception *explicite* de plans de sondage à usages multiples semble être une chose rare mais combien nécessaire.

La *polyvalence* des échantillons d'enquête revêt plusieurs aspects qui sont hiérarchisés en *six niveaux* dans la section suivante. Nous définissons ensuite *dix sources d'incompatibilité* entre des objectifs. Ces sources d'incompatibilité sont traitées séparément dans les sections 4 à 9, où nous présentons des façons de résoudre l'incompatibilité. Certaines des solutions proposées puisent dans des éléments bien connus de la théorie des sondages tandis que d'autres sont plus inédites, donc moins bien élaborées et documentées.

Par cet article, nous visons tout particulièrement à fournir aux utilisateurs de la documentation utile sur les méthodes de conception et d'application des plans de sondage à usages multiples

---

[1] Communication-thème présentée au Symposium international sur la statistique de Taipei, Taïwan, août 1986, et à un colloque de Statistique Canada, 7 octobre 1987.

[2] Leslie Kish, Institute for Social Research, University of Michigan, Ann Arbor, MI 48104 É.-U.

et à les sensibiliser à l'importance et à la réalité de ces plans. En second lieu, nous voulons poser les jalons d'une recherche intégrée sur les nombreux problèmes que soulèvent les plans de sondage à usages multiples. Souhaitons que les imperfections de nos méthodes incitent d'autres personnes à poursuivre la recherche et à élaborer de nouvelles méthodes.

2. UNE HIÉRARCHIE D'USAGES

Au départ, il nous faut clarifier le sens de l'expression "à usages multiples" car celle-ci renvoie à un trop grand nombre de concepts dans la documentation statistique. À cette fin, nous avons réparti une vingtaine d'usages en six niveaux hiérarchiques (tableau 1). Les niveaux 3 (variables multiples) et 4 (enquêtes portant sur plusieurs sujets à la fois) sont ceux qui font le plus souvent l'objet d'analyses et parfois, ils sont considérés séparément pour les mêmes variables ou des variables étroitement liées (Murthy 1967). Chacun des six niveaux est subdivisé selon les formes d'usage, celles-ci étant analysées plus en détail dans d'autres ouvrages (Nations-Unies 1980; Lahiri 1963); les listes du tableau 1 ne sont pas exhaustives.

Les opérations d'enquête intégrées (niveau 5) ont un lien avec les enquêtes portant sur plusieurs sujets mais doivent être distinguées de celles-ci parce qu'elles se rattachent à des organismes et à des établissements qui réalisent de nombreuses enquêtes dans divers domaines sur de plus longues périodes (Nations-Unies 1980; Foreman 1983). Le cinquième niveau était désigné comme les "opérations d'enquêtes permanentes" jusqu'à ce qu'on reconnaisse que la plupart des grandes enquêtes par sondage étaient réalisées par des organismes d'enquête permanents comme le U.S. Census Bureau, Statistique Canada ou le Survey Research Center. Cette permanence présente des avantages notables au point de vue des coûts et de la qualité, sans compter les effets restrictifs sur les plans de sondage (Kish 1965).

Les bases de sondage principales ou les échantillons principaux (niveau 6) représentent une forme plus poussée et plus spécialisée de plans de sondage à usages multiples. Cela peut tout simplement signifier que l'on utilise les mêmes cartes ou listes d'îlots ou les mêmes segments de secteur pour des enquêtes différentes ou que l'on recourt à l'"échantillon principal de l'agriculture" (King et Jessen 1945), méthode selon laquelle les régions rurales de tous les comtés des É.-U. sont divisées en segments comptant environ quatre fermes chacun; il peut aussi s'agir de l'entreprise qui vend la liste courante des logements pour la formation de la plupart des échantillons utilisés en Allemagne de l'Ouest. Ces exemples très variés ont toutefois quelque chose en commun: dans les trois cas, le partage des "frais de mise en marche" (conception du plan, stratification, listage, etc.) pour la construction de bases de sondage permet de réaliser des économies.

Les ouvrages portant sur l'échantillonnage à usages multiples ont toujours passé sous silence les paramètres statistiques fondés sur une seule variable et des domaines variés (niveaux 1 et 2) même si ces catégories sont les plus courantes; toutefois, comme nous le verrons plus loin, ces catégories peuvent avoir les conséquences les plus notables et soulever les problèmes les plus sérieux. L'effet du plan sur des paramètres statistiques comme les médianes, les quantiles et les coefficients de régression peut être très différent de l'effet du même plan sur les moyennes et les agrégats (Kish 1961; Kish 1965; Kish et Frankel 1974). De plus, la formation d'échantillons périodiques soulève de nouvelles questions (Section 8). Toutefois, l'effet du plan se fait sentir surtout sur les moyennes des petites "sous-classes" (c'est-à-dire, aussi petites que 0.10 ou 0.01) de l'échantillon intégral, ces sous-classes correspondant à des "domaines" similaires dans la population (Section 5).

Chacun des niveaux du tableau 1 présente des aspects différents des plans de sondage et peut être approfondi avec profit afin d'en préciser la signification et d'en tirer des exemples plus précis (dont certains figurent déjà dans le tableau 1).

Les lacunes qui ont valu aux plans de sondage à usages multiples d'être omis sont nombreuses. Premièrement, il est nécessaire de formuler les différents usages *en termes statistiques explicites* de sorte qu'ils puissent être utilisés dans les formules qui permettent de les comparer et celles qui servent à établir des compromis mais voilà, la principale difficulté est peut-être d'obtenir une

Tableau 1
Hiérarchie d'usages

1. Statistiques diverses établies à partir des mêmes variables
   - Totaux, moyennes, médianes et quantiles; distributions
   - Statistique analytique: régressions, analyses de catégories
   - Aspects chronologiques: statique, macro-variation, micro-variation, cumulatif
2. Populations et domaines divers (sous-classes)
   - Classes propres et classes de recoupement
   - Comparaison de sous-classes
3. Variables multiples sur le même sujet
   - Différentes mesures d'une même variable (par ex.: revenu, chômage)
   - Périodes diverses — par jour, par semaine, par mois, par année
   - Plusieurs aspects d'un même sujet: revenu, épargne, patrimoine
4. Enquêtes portant sur plusieurs sujets
   - Plusieurs sujets traités sur le même questionnaire, durant la même interview ou au cours de la même enquête
   - Enquêtes-santé portant sur de nombreuses maladies
   - Études de marché pour de nombreux clients et portant sur de nombreux produits
   - Enquêtes agricoles portant sur de nombreuses cultures
   - Enquêtes sociales à usages multiples
5. Opérations d'enquête intégrées (enquêtes permanentes)
   - La NSS en Inde, la CPS aux É.-U., le NHSCP des N.-U.
   - Enquêtes distinctes confiées à des divisions et à des bureaux régionaux différents
   - Source commune des enquêtes
   - Diversité des méthodes, des coûts, des opérations, des répartitions, des répondants
6. Bases de sondage principales
   - Plusieurs échantillons tirés d'une base de sondage ou d'une série de listages
   - Établissements, organismes distincts
   - Effectifs régionaux distincts? Mêmes UPÉ?

liste complète de ces termes statistiques. Deuxièmement, il faut connaître les estimations des *facteurs de variance* et *des facteurs de coût* pour chaque usage. Troisièmement, certaines méthodes exigent que l'on établisse des valeurs *au degré de précision "requis"* pour tous les usages (Section 5). Quatrièmement, les valeurs et les estimations précitées doivent être incluses dans une formule mathématique qui aura pour solution *un plan "optimal" unique* qui sera prêt à être utilisé. Les ordinateurs ont simplifié les calculs nécessaires à une telle solution mais l'analyse conceptuelle et théorique est toujours aussi exigeante (Section 5).

Les exigences qui se rattachent à l'étude des plans à usages multiples nous aident à comprendre pourquoi les manuels font peu souvent mention de ce genre de plans. Il convient toutefois de signaler les ouvrages qui figurent dans la bibliographie ci-jointe de même que les notes et la bibliographie qui paraissent dans Rodriquez-Vera (1982), Cochran (1977) et Chatterjee (1967). Par ailleurs, les descriptions d'enquêtes présentent souvent un paramètre statistique (par ex., la moyenne) d'une variable principale comme *le* seul objet (principal) de l'enquête. Dans le contexte des plans à usages multiples, cela revient à enlever toute importance aux autres buts de l'enquête. On peut atténuer l'invraisemblance de cette affirmation en supposant que les autres objets principaux donneraient lieu à des répartitions semblables; il n'en reste pas moins que cette affirmation devrait être étayée à l'aide des calculs prévus dans les quatre étapes exposées précédemment.

3. APERÇU DE DIX SOURCES D'INCOMPATIBILITÉ

Il serait utile d'examiner brièvement les dix sources d'incompatibilité énumérées dans le tableau 2 avant d'aborder des problèmes spécifiques et de proposer des solutions. La liste du tableau 2

n'est probablement pas exhaustive et le lecteur pourrait bien lui ajouter des éléments. Mieux encore, il pourrait définir à l'intérieur de chacun des points du tableau des problèmes et des solutions qui ne sont pas traités dans le présent document. Il serait souhaitable de mettre les dix sources d'incompatibilité en relation logique avec les vingt usages répartis en six niveaux; nous pourrions alors ramener cet exposé à vingt noeuds usage/source d'incompatibilité, par exemple, ou à dix noeuds niveau/source d'incompatibilité. Malheureusement, les sources d'incompatibilité dénotent une dimension perpendiculaire par rapport à l'usage et les soixante (10 × 6) cellules (ou la plupart d'entre elles) ont un contenu significatif.

Heureusement, les dix sources d'incompatibilité ne sont pas toutes présentes dans chaque plan de sondage. Nous estimons qu'il faudrait toujours considérer, du moins de façon informelle, la possibilité d'un problème au point de vue des tailles d'échantillon m_g et du rapport entre les biais et les erreurs d'échantillonnage parce qu'il s'agit là des problèmes les plus fréquents. Il faudrait aussi se pencher, au moins brièvement, sur la répartition de la taille de l'échantillon entre les domaines et la répartition entre les strates; de fait, il faudrait s'arrêter encore plus souvent sur ces points. Le calcul des erreurs d'échantillonnage (j) devrait aussi être examiné dans la plupart des enquêtes. En revanche, pour ce qui a trait aux enquêtes uniques, on ne s'intéressera pas aux problèmes qui touchent la permanence du plan de sondage. Par contre, dans le cas d'une enquête permanente fondée sur une base de sondage permanente, les décisions concernant e), f), g) et h) (stratification, taille des grappes et mesures) peuvent avoir été prises il y a longtemps pour un plan de sondage fixe. Malgré cela, les tailles de grappe (f) utilisées dans les étapes intermédiaires (îlots et segments) peuvent faire l'objet de modifications opérationnelles souples.

Il est de plus rassurant de savoir que des compromis fondés sur des méthodes statistiques peuvent produire des résultats tout à fait acceptables et ce, pour plusieurs raisons (Section 5 à 8). La première est qu'un faible écart entre la répartition observée et la répartition optimale n'entraîne qu'un très faible accroissement de la variance. Les courbes d'efficacité tendent à être horizontales sur de larges intervalles centrés sur les points optimaux; ainsi, une grande précision pour des plans de sondage distincts, par ailleurs impossible à obtenir, n'est pas nécessaire. La seconde raison est qu'un écart appréciable entre la répartition observée et la répartition optimale peut entraîner un accroissement notable de la variance. Ainsi, le fait d'ignorer les usages importants peut entraîner des pertes d'efficacité appréciables dans leur cas, ce qui nous oblige à élaborer des plans de compromis pour ces usages. La troisième raison est que les plans de compromis, en conformité avec les méthodes statistiques, peuvent réduire sensiblement les pertes élevées que pourrait engendrer une optimalisation des répartitions pour d'autres usages, et produire des variances à peine plus élevées que celles découlant des plans optimaux pour chaque usage (Section 5).

4. TAILLE DES ÉCHANTILLONS ET RAPPORTS DE BIAIS (B/σ)

Ces deux sources d'incompatibilité (a et b dans le tableau 2) passent probablement pour les plus importantes de toutes à cause des conséquences qu'elles peuvent avoir. Si nous les considérons ensemble ici, c'est uniquement parce qu'elles peuvent être liées étroitement l'une à l'autre par les effets des sous-classes. Commençons par le cas le plus connu (échantillonnage aléatoire simple avec remise), où $m = S^2/V^2$ représente la taille d'échantillon nécessaire pour obtenir un degré de précision "requis" V^2 pour une moyenne d'échantillon y avec variance S^2 des unités élémentaires. Or, S_g^2 dépend largement des variables et des domaines, qui varient en fonction de g pour la moyenne $\bar{y}_g$, et la valeur "requise" V_g^2 peut varier davantage. Nous considérons aussi les effets de plan D_g^2, qui varient également, et par voie de conséquence, $m_g = S_g^2 D_g^2 / V_g^2$ représente la taille d'échantillon nécessaire pour la moyenne de la variable g. En ce qui concerne la moyenne $\bar{y}_g$ d'un domaine g qui ne comprend qu'une proportion P_g de la population, la taille d'échantillon *globale* requise pour le domaine devient $n_g = m_g/P_g$ et il est plus pratique d'exprimer la fraction de sondage requise sous la forme $f_g = n_g/N = S_g^2 D_g^2 / V_g^2 P_g N$. On peut omettre le facteur (1-f) ou l'inclure dans D_g^2. P_g devient faible et déterminante s'il est nécessaire d'avoir un degré de précision élevé pour les petites sous-classes.

Tableau 2

Dix sources d'incompatibilité (a-j)

a.(4)[1] Des tailles m_g ou des taux f_g sont nécessaires pour les usages g

$V_g^2 = S_g^2 D_g^2 / m_g$ et $m_g = S_g^2 D_g^2 / V_g^2$ ou $f_g = S_g^2 D_g^2 / V_g^2 P_g N$

où m_g désigne la taille des sous-classes et $f_g = n_g/N = m_g/P_g N$ désigne le taux de sondage

b.(4) Rapport entre les biais et les erreurs d'échantillonnage dans EQM $= \sqrt{(\sigma^2 + B^2)}$
- Le rapport de biais B/σ diminue lorsque σ s'accroît pour les sous-classes
- À des fins de comparaison, B/σ tend à diminuer lorsque B diminue ou σ augmente

c.(5) Répartition de m_g entre les domaines

$m_t = \Sigma_g m_g$

d.(6) Répartition de m_{gh} entre les strates h

$m_g = \Sigma_h m_{gh}$

e.(6) Choix de variables pour la stratification
Stratification à plusieurs variables

f.(7) Tailles de grappe optimales

$D_g^2 = [1 + \rho_g(\bar{b}_g - 1)]\bar{b}_g = P_g n_t / a$ pour les classes de recoupement

g.(7) Mesures pour les tailles de grappe

h.(7) Conservation des unités d'échantillonnage (UPÉ) pour de nouveaux sujets, de nouvelles mesures et de nouvelles strates et pour des sujets variés.

i.(8) Évolution du plan de sondage
Quel est le degré de chevauchement? Doit-il y avoir des panels? Variation vs agrégation.

j.(9) Calcul et présentation des erreurs d'échantillonnage.

[1] Les chiffres (4) à (9) renvoient aux sections où sont traités ces problèmes.

Pour la comparaison de sous-classes, on observe une augmentation encore plus forte des variances: $m_g = (m_a^{-1} + m_b^{-1})^{-1} = n(P_a^{-1} + P_b^{-1})^{-1}$, où P_a et P_b désignent des proportions de l'échantillon n (en supposant que $S_a^2 = S_b^2$). Si nous comparons par exemple deux moyennes de sous-classes ($0.01n$ et $0.10n$), nous avons la "taille effective" $m_g = n(0.01^{-1} + 0.10^{-1})^{-1} = n/110$. Pour ce qui a trait à d'autres paramètres statistiques comme les médianes et les coefficients de régression, il serait trop complexe de définir des tailles d'échantillon "requises". Cela déborde le cadre de notre étude mais il n'est pas exclu que nous présentions certains chiffres.

L'analyse des paramètres statistiques des sous-classes prend une toute autre dimension si nous incluons le biais B^2 dans l'erreur quadratique moyenne (EQM = RMSE $= \sqrt{(\sigma^2 + B^2)}$). La figure 1 sert à illustrer une tendance courante dans l'exactitude des données d'enquête bien qu'il puisse y avoir des écarts appréciables entre les rapports du biais à l'erreur d'échantillonnage; nous recommandons fortement de lire la légende qui accompagne la figure. Il arrive souvent que le biais potentiel B^1 soit supérieur à l'erreur d'échantillonnage mesurable σ_1 pour tout l'échantillon. Cependant, si nous prenons une sous-classe qui représente environ le dixième de l'échantillon,

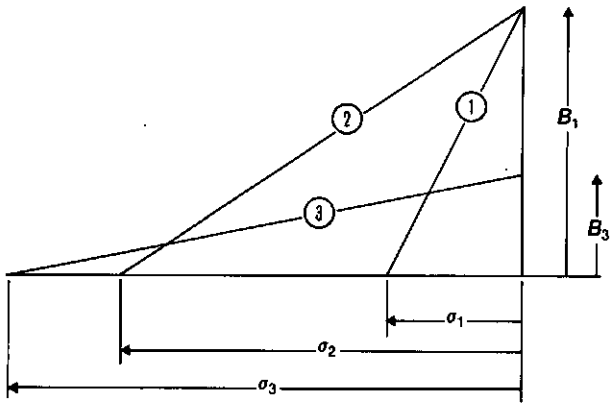

Figure 1. Erreurs variables (σ) et biais (B) dans l'erreur quadratique moyenne (EQM)

Les bases des triangles représentent les erreurs d'échantillonnage et d'autres erreurs variables (σ). Par exemple, σ_1 peut désigner l'erreur type de la moyenne $\bar{y}$ de l'échantillon intégral, $ste(\bar{y}_t)$ peut désigner l'erreur type (plus élevée) d'une moyenne de sous-classe ($ste(\bar{y}_c)$) et σ_3 peut désigner l'erreur type de la différence entre deux moyennes de sous-classes $ste(\bar{y}_c \text{ ⅛ } \bar{y}_b)$).

La hauteur des triangles représente les biais (B) et l'hypoténuse représente l'EQM = $\sqrt{(\sigma^2 + B^2)}$. (1) Pour l'échantillon intégral, le biais B_1 peut être élevé par rapport à l'erreur variable σ_1 ; par conséquent, le fait d'utiliser de plus grands échantillons ne réduirait pas l'EQM de beaucoup. (2) Par ailleurs, avec le même biais B_1 mais un échantillon moindre, tiré d'une sous-classe, le rapport change et l'erreur variable (σ_2) constitue désormais la majeure partie de l'erreur quadratique moyenne (EQM_2); du reste, EQM_2 n'est pas beaucoup supérieure à EQM_1 même s'il s'agit d'un échantillon beaucoup plus petit. (3) En ce qui concerne la différence de moyennes, le biais net B_3 peut être beaucoup moindre de sorte que même si est plus élevée que et σ_3, EQM_3 n'est que légèrement supérieure à EQM_2. Ce changement notable du rapport de biais B/σ tend à se produire non seulement dans le cas des différences de moyennes de sous-classes à l'intérieur d'un même échantillon, mais aussi dans le cas de différences de valeurs entre des enquêtes successives.

nous constatons que l'erreur type est environ trois fois plus élevée que dans le premier cas. Si nous comparons les moyennes de deux sous-classes de ce genre, l'erreur type (σ_3) est environ 1.4 fois plus élevée que dans le second cas.

En revanche, nous pouvons constater que les hypoténuses, qui représentent l'EQM, augmentent beaucoup moins fortement. Pour ce qui est de l'EQM_1, le biais B_1 est clairement supérieur à l'erreur type et cela peut être observé pour certaines variables dans de grands échantillons. Par contre, le fait de conserver le même biais ($B_2 = B_1$) pour la sous-classe a fait s'accroître l'EQM modérément et nous voyons que l'erreur type (σ_2) est maintenant supérieure au biais. La supériorité de l'erreur type est encore plus forte dans le cas de EQM_3, où σ_3 est supérieur à σ_2 mais où le biais (qui est censé être affecté du même signe qu'auparavant parce qu'il s'agit d'une tendance courante) est tombé à B_3 (comparaison des moyennes).

Des exemples de ces phénomènes se retrouvent partout et figurent à toutes fins dans le tableau 1. Nous choisissons l'exemple le plus courant — statistiques du chômage, où les biais observés peuvent éclipser entièrement les plus faibles valeurs des variations mesurables (par ex. 0.1 %). Toutefois, en ce qui a trait aux petites sous-classes (par ex.: adolescents de race noire), les erreurs d'échantillonnage pour les petites bases de sondage excèdent les biais. Dans le cas des comparaisons périodiques, les variations d'échantillonnage sont encore plus prononcées.

Les rapports que nous venons de décrire et qui mettent en relation les biais et les erreurs variables ne sont pas imposés par des considérations logiques; ils relèvent plutôt de considérations empiriques et courantes. Le fait de négliger ces rapports élémentaires créé beaucoup de confusion quant à la nécessité d'avoir des enquêtes par sondage d'une grande précision (c.-à-d., erreurs d'échantillonnage faibles(σ)). Par la figure 1, nous voulons donner des réponses pratiques à des questions courantes comme celles qui suivent: pourquoi dépenser de l'argent pour de grands échantillons et des méthodes d'échantillonnage rigoureuses devant des biais de mesure élevés? Pourquoi se donner la peine de calculer les erreurs d'échantillonnage si les biais de réponse constituent la majeure partie de l'erreur totale? Les réponses à ces questions se trouvent dans la supériorité

des erreurs d'échantillonnage pour les sous-classes et pour la comparaison des moyennes des sous-classes. Faisons en sorte que ces réponses soient plus explicites dans les plans de sondage à venir.

5. RÉPARTITION ENTRE LES DOMAINES

Cette source d'incompatibilité des plus importantes et des plus fréquentes revêt plusieurs aspects. Considérons tout d'abord la répartition de la taille globale d'un échantillon (ou la répartition des efforts ou des coûts) entre des domaines qui correspondent chacun à une partie de la population. Un exemple type est la répartition de la taille d'un échantillon entre les nombreuses (5, 10, 20 ou 50) provinces ou régions ou les nombreux états d'un pays; ces domaines renferment habituellement des populations N_d très différentes (dont le rapport peut varier de 1 à 100) même si les territoires eux peuvent être de même superficie ou presque. La question que l'on se pose souvent dans les circonstances est la suivante: les tailles d'échantillons (n_d) devraient-elles être à peu près les mêmes? Ou devraient-elles être proportionnelles à N_d avec des taux de sondage constants $f_d = f$? Des tailles identiques tendent à produire des erreurs de même valeur pour les moyennes (e.t.($\bar{y}_d$)). En revanche, des taux de sondage constants tendent à produire l'erreur type la plus faible pour la moyenne globale $\bar{y}_w = \Sigma W_d \bar{y}_d$ parce qu'ils produisent des erreurs plus faibles pour les grands domaines. L'erreur type obtenue peut être moins élevée que la "normale", surtout si l'on tient compte des biais potentiels (figure 1), et peut ne pas justifier la taille élevée des échantillons et les coûts afférents. C'est ce que prétendent ceux qui préconisent la même taille d'échantillon (n_d) pour toutes les provinces. Cependant, on relève des erreurs d'échantillonnage plus élevées pour $\bar{y}_w$ dans la plupart des sous-classes, notamment celles formées par le recoupement de variables telles l'âge, le sexe, la classe socio-économique, etc., et dont l'effectif tend à être proportionnel à l'effectif total. Cela est le genre d'inconvénients qui découlent de la forte disparité des taux de sondage $f_d = n_d/N_d$ applicables aux provinces à cause de l'uniformité des tailles d'échantillons n_d.

Dans la Current Population Survey des États-Unis par exemple, on attribue un taux de sondage plus élevé aux États moins populeux. La pondération qui s'ensuit a pour effet d'accroître la variance (pour un coût total fixe) des moyennes globales et des "classes de recoupement" comme celles formées des jeunes hommes et des jeunes femmes, et plus particulièrement des adolescents et adolescentes de race noire (touchés par des taux de chômage extrêment élevés). Ce genre d'incompatibilité entre les besoins nationaux et régionaux (ou provinciaux) se retrouve dans tous les pays à cause des fortes variations de population d'une région à l'autre. Ainsi, le fait de vouloir accroître la qualité des données provinciales, pour un coût total fixe, est en contradiction avec le fait de vouloir accroître le degré d'exactitude des données nationales et des données recoupées.

Pour réduire les risques de confusion, nous définissons les "domaines" comme des sous-ensembles de la population et les "sous-classes" comme des sous-ensembles de l'échantillon. Nous faisons ensuite la distinction entre les "domaines du plan" (et les sous-classes du plan), qui sont des sous-ensembles (par ex.: provinces et régions) contenus dans des strates définies par le plan de sondage, et les "classes de recoupement" (par ex.: âge, sexe, profession, revenu, etc.), qui s'étendent souvent presque aléatoirement sur plusieurs strates à la fois. Les effets du plan ne sont pas les mêmes pour ces deux genres de sous-classes (Kish 1961, 1980, 1987).

Par ailleurs, des difficultés peuvent surgir au point de vue de la répartition des variables et de la précision des variances $D_d^2 S_d^2$ pour des raisons autres que les différences de taille entre les *domaines*; cependant, il n'est pas approprié ici d'approfondir cet aspect de la question. Nous voulons plutôt attirer l'attention sur deux méthodes techniques qui permettent de résoudre simultanément tous les problèmes de répartition (la quatrième étape décrite à la fin de la Section 2). L'une de ces méthodes consiste à utiliser la programmation non linéaire itérative pour satisfaire simultanément à toutes les exigences de fiabilité relatives *aux usages énoncés* et ce, à un coût minimum. Cette solution élégante, qui convient à divers problèmes, fait appel à l'ordinateur et a fait l'objet de nombreux articles depuis 1963 (voir les comptes rendus et les bibliographies dans Bean et

Burmeister 1978; Rodriquez-Vera 1982 et Cochran 1977). Le "coût minimum requis" est souvent beaucoup trop élevé parce que les exigences de fiabilité sont irréalistes. Il faut alors ramener la solution à un niveau beaucoup plus modeste. Or, un tel geste met au jour les prétentions (fausses à notre avis) de cette solution élégante qui repose sur des exigences de fiabilité irréalistes. Nous nous interrogeons principalement sur le bien-fondé de l'utilisation de "fonctions en escalier" pour illustrer le degré de fiabilité "requis"; suivant ce genre de fonctions, toute variance inférieure à V^2 (valeur requise) prend une valeur constante et toute variance supérieure à V^2 prend une valeur nulle.

Une méthode très différente de celle décrite ci-dessus consiste à établir un genre de *moyenne* pour toutes les répartitions "optimales" (c.-à-d. privilégiées) se rapportant à divers usages, en *minimisant la variance combinée (pondérée)* soit pour le coût fixe ou la taille d'échantillon fixe. Certes, si la variance ainsi obtenue est trop élevée (ou trop faible), on peut modifier les solutions en conséquence (c.-à-d. les rehausser ou les rabaisser). Nous préférons cette dernière méthode, qui est un compromis entre diverses répartitions qui existent chacune sous une forme optimale pour un seul usage (Yates 1981; Dalenius 1957). Cette méthode prévoit l'attribution de poids relatifs I_g à tous les paramètres statistiques énumérés, exercice qui peut paraître difficile à première vue (quoiqu'un décideur "inexpérimenté" pourra attribuer la même valeur à tous les paramètres). Toutefois, les deux autres solutions qui s'offrent à nous sont plus rigoureuses et devraient s'avérer encore plus difficiles à appliquer: la première consiste à définir le degré de fiabilité "requis" de tous les paramètres statistiques pour la première méthode, puis à attribuer arbitrairement des poids de même valeur à tous ces paramètres; la seconde méthode consiste à attribuer le poids unitaire à un paramètre statistique en particulier et des poids nuls à tous les autres paramètres.

En outre, il est facile de montrer que des compromis relatifs à la moyenne sont ordinairement réalisables et utiles parce que les répartitions ne sont pas influencées par les variations modérées de poids (comme c'est souvent le cas en statistique). Après tout, il est plus raisonnable de modifier l'importance relative d'une variable dans un rapport de 2 ou 5 par exemple que d'attribuer un poids de 1 à une variable et un poids nul à toutes les autres variables, cette dernière solution produisant des rapports de poids de valeur infinie.

Premièrement, définissons $\Sigma_i V_{gi}^2/n_i$ comme la variance possible d'un paramètre statistique g avec les résultats n_i de la répartition de la taille de l'échantillon pour la ième composante de la variation. Ensuite, posons $1 + L_g(n) = (\Sigma_i V_{gi}^2/n_i)/V_g^2(\min) = \Sigma_i C_{gi}^2/n_i$ comme le taux d'accroissement (avec le résultat n_i de la répartition) de la variance du paramètre statistique g par rapport à la valeur minimum de cette variance; dans les deux cas, nous considérons la même valeur fixe Σn_i. Par conséquent, $L_g(n)$ représente la perte par rapport à la valeur minimum 1; à ce stade-ci, nous devons décider si les fonctions à minimiser seront les variances relatives C_{gi}^2/n_i; il s'agit là d'une décision fondamentale. À notre avis, il n'y a pas de fonction plus appropriée que celles-là pour les besoins de l'équation (5.1) ci-dessous. Par exemple, nous préférons ces fonctions à V_{gi}^2, qui repose sur des unités de mesure arbitraires, lesquelles sont effacées par $V_{gi}^2(\min)$. Il arrivera toutefois à de rares occasions que $V_g^2(\min)$ soit nulle ou très faible, auquel cas C_{gi}^2 pourrait être très élevée et instable; si cette situation devait se présenter, on attribuera des valeurs arbitraires à C_{gi}^2 ou à I_g dans l'équation ci-dessous. Ces paramètres et ceux que nous verrons plus loin, y compris ceux du tableau 3, sont définis et analysés par Kish (1976).

Alors, étant donné les poids I_g servant à mesurer l'importance relative du paramètre statistique g pour n'importe quelle série de résultats n_i de la répartition des tailles d'échantillons, nous avons:

$$1 + L(n) = \Sigma_g I_g(1 + L_g(n)) = \Sigma_g I_g \Sigma_i C_{gi}^2/n_i$$

$$= \Sigma_i \Sigma_g I_g C_{gi}^2/n_i = \Sigma_i Z_i^2/n_i. \qquad (1)$$

Après avoir modifié l'ordre des sommations, nous avons créé la nouvelle variable $Z^2{}_i = \Sigma_g I_g C_{gi}^2$. Nous pouvons minimiser cette fonction afin d'obtenir des solutions de compromis pour

Tableau 3

Fonctions de perte (1 + L) pour deux populations (Kish 1976)

Résultats de répartition m_i	(A) (1 + L) pour $W_2/W_2 = 4$			(B) (1 + L) pour 133 pays: 0.2 à 100 millions			
						combinée avec poids	
	$\Sigma W_i \bar{y}_i$	$\Sigma \bar{y}_i/2$	combinée	$\Sigma W_i \bar{y}_i$	$\Sigma \bar{y}_i/133$	1:1	I_c/I_d:1
MW_i	1	1.56	1.28	1	6.86	3.93	
M/H	1.36	1	1.18	3.34	1	2.17	
$\propto\sqrt{(W_i)}$	1.08	1.125	1.102	1.35	1.54	1.44	
$\propto\sqrt{(W_i^2 + H^{-2})}$	1.116	1.080	1.098	1.31	1.28	1.295	
$\propto\sqrt{(0.5W_i^2 + H^{-2}}$				1.47	1.17	(1.32)	1.27
$\propto\sqrt{(2W_i^2 + H^{-2})}$				1.20	1.44	(1.32)	1.28
$\propto\sqrt{(4W_i^2 + H^{-2})}$				1.12	1.66	(1.39)	1.23

Dans le cas (A), il y a deux strates et deux domaines ($W_1 = 0.8$ et $W_2 = 0.2$); il convient de souligner que $m_i = \sqrt{W_i}$ produit un résultat presque aussi bon que la répartition optimale dans le cas de la fonction de perte combinée.

Dans le cas (B), nous considérons les populations de 133 pays, dont la taille varie de 200,000 à plus de 100,000,000, soit un rapport de 1 à 500. Pour des raisons pratiques, nous avons exclu de notre analyse (par rapport au traitement qui a été fait pour l'enquête mondiale sur la fécondité) les quatre plus importants pays et les quelques pays qui comptent moins de 200,000 habitants. Leur inclusion aurait fait passer la variance des rapports de tailles, W_i, de 2.5 à 12 et aurait produit des résultats spectaculaires. Il convient de souligner que la répartition $\sqrt{W_i}$ réduit sensiblement les pertes. Un compromis est préférable à rien du tout. Néanmoins, le répartition optimale, $\sqrt{(W_i^2 + H^{-2})}$, est bien supérieure. Diverses valeurs de I_c/I_d (= 1/2, 2/1 et 4/1) augmentent légèrement la variance de la fonction de perte combinée avec des poids (1:1) mais produisent des résultats stables pour les fonctions de perte combinées avec poids (I_c/I_d:1).

Deux exemples dans le tableau 3 illustrent les compromis étonnamment satisfaisants qui ont pu être établis entre des résultats de répartition contradictoires grâce à la méthode de la moyenne pondérée: les données pertinentes, qui se trouvent à la quatrième ligne du tableau 3, se comparent très avantageusement aux autres données du tableau. Ces résultats s'expliquent par l'existence d'une très longue courbe horizontale pour la répartition optimale, comme nous l'avons souligné dans la Section 2 et dans d'autres ouvrages (Kish 1976; Kish 1987). Au Canada, par exemple, le rapport de tailles de population entre les provinces peut atteindre 1/70, ce qui nous amène à dire que le cas des dix provinces canadiennes peut être apparenté au cas B du tableau 3; par ailleurs, le graphique de la figure 2 a été construit à l'aide des chiffres de population des dix provinces. (Vois aussi Fellegi et Sunter 1974.)

le coût total fixe $\Sigma c_i n_i$. En ce qui a trait à la controverse entourant la définition de n_d ($n_d = n/H$ — échantillons de même taille pour les domaines *ou* $n_d = nW_d$ — taille d'échantillon proportionnelle à l'effectif W_d du domaine), nous pouvons affirmer que les résultats de la répartition optimale (de compromis) sont proportionnels à $\sqrt{(W_d^2 + H^{-2})}$, avec des poids identiques I_g.

Les enquêtes mondiales sur la fécondité illustrent bien le cas qui nous occupe; dans ces enquêtes, on utilise des échantillons de taille comparable pour les petits et les grands pays. Les tailles d'échantillon se situaient uniquement entre 3 et 10 000, sans corrélation apparente avec la taille de la population. Les variances des moyennes continentales des enquêtes nationales ont par conséquent doublé ou triplé. On décrit dans ces termes l'apport des enquêtes sur la fécondité:

> "Le principal apport de ces enquêtes a été jusqu'à maintenant de confirmer la baisse soutenue de la fécondité, qui a caractérisé une bonne partie de l'Asie et de l'Amérique latine dans les années 1970, et de faire ressortir le contraste entre ces continents et l'Afrique, où le taux de fécondité demeure élevé et où la volonté d'avoir de nombreux enfants est encore très réelle" (Macura et Cleland 1985). (TRADUCTION).

6. RÉPARTITION ENTRE LES STRATES ET CHOIX DES CRITÈRES DE STRATIFICATION

On confond souvent les domaines et les strates dans les discussions mais il importe de bien les distinguer dans les études pratiques sur les plans de sondage. Les domaines sont des sous-

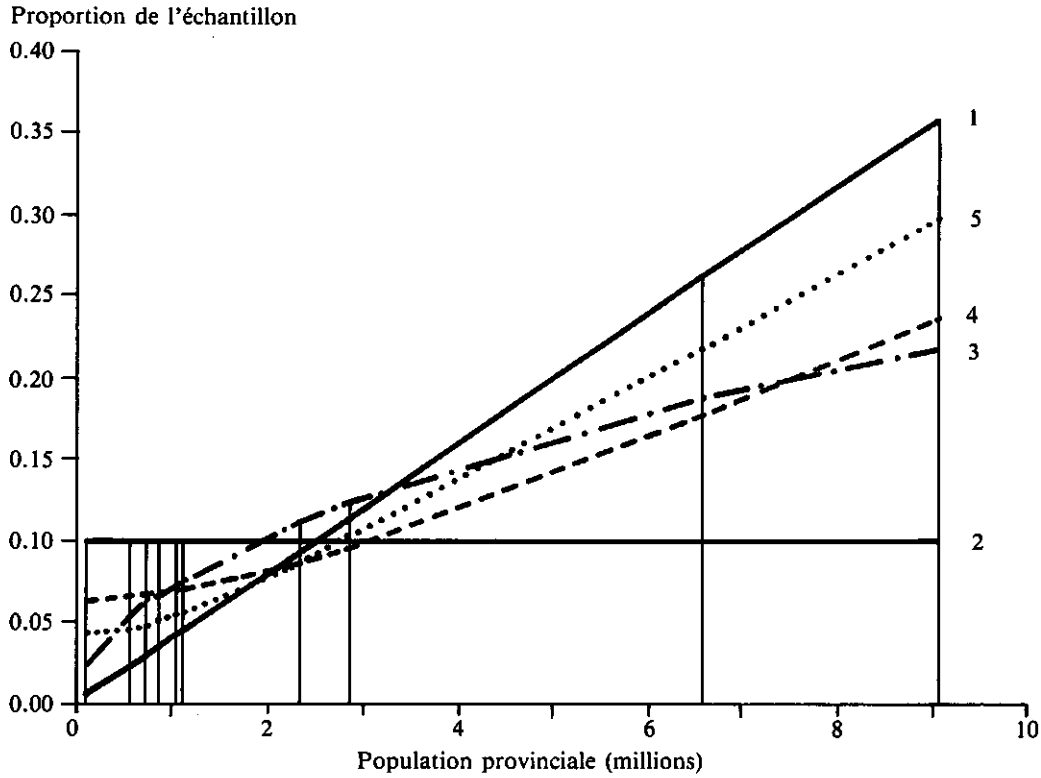

Figure 2. Cinq modes de répartition de la taille n_h de l'échantillon (pour un total fixe Σn_h)

Les dix provinces du Canada permettent d'illustrer dans le graphique ci-dessus les problèmes qui découlent habituellement de l'existence de grands domaines de tailles inégales et les compromis avantageux qu'il est possible d'établir.

1 La répartition proportionnelle à la taille des domaines ($n_h \propto W_h$) est représentée par une diagonale.
2 La répartition uniforme ($n_h \propto 1/H$) est représentée par une droit horizontale.
L'écart entre ce mode de répartition et le précédent est appréciable à l'extrémité des courbes.
3 La répartition par la racine carrée ($n_h \propto \sqrt{W_h}$) produit des compromis aux deux extrémités.
4 La répartition optimale ($n_h \propto \sqrt{(W^2 + 1/H^2)}$) améliore les deux extrémités mais on constate un "aplatissement" intéressant à l'extrémité gauche de la courbe.
5 La répartition optimale "pondérée" ($n_h \propto \sqrt{(.8W^2 + .2/H^2)}$) améliore considérablement la partie supérieure de la courbe.

populations pour lesquelles on calcule des estimations propres tandis que les strates sont habituellement des divisions de ces sous-populations, qui sont créées pour réduire les variances. Si l'on considère par exemple les provinces comme des domaines, on y créera plus de strates dans le but de réduire les variances à l'échelle de la province; en revanche, les domaines de recoupement comme l'âge, le sexe et la situation économique tendent à chevaucher les strates. Bien qu'elle ne soit pas aussi capitale que la répartition entre les domaines, la répartition de la taille de l'échantillon entre les strates peut être essentielle dans le cas des répartitions optimales efficientes disproportionnées. Les deux méthodes que nous avons décrites dans la Section 5 pour la répartition de la taille de l'échantillon entre les domaines peuvent aussi servir à la répartition entre les strates mais dans des buts différents. Quelques-unes des sources sur la programmation non linéaire concernent les domaines et d'autres les strates tandis que d'autres encore confondent les deux.

La présence de plusieurs variables d'enquête et de plusieurs paramètres statistiques parmi les usages justifie clairement l'utilisation d'un plus grand nombre de variables de stratification. Des

variables d'enquête différentes tendront à avoir des relations optimales variées avec les critères de stratification; il est donc fortement recommandé d'utiliser un grand nombre de critères même si chacun d'eux n'est utilisé qu'avec quelques divisions de strates (catégories). Le plan de sondage à usages multiples est ce qui se prête le mieux à la stratification à plusieurs variables (Kish et Anderson 1978). C'est peut-être aussi ce qui justifie le mieux le besoin de méthodes "d'échantillonnage contrôlé". Le choix des limites de strates, appelé "stratification optimale", est un sujet connexe mais secondaire dans le cadre du présent article.

7. TAILLES DE GRAPPES; MESURES DE LA TAILLE; CONSERVATION D'UNITÉS

Nous constatons parfois dans la description des plans de sondage que l'effet du plan a été estimé au moyen de la formule $D_g^2 = [1 + \rho_g \bar{b}_t - 1)]$, où ρ désigne une corrélation intraclasse synthétique entre la variable "la plus importante" g et $\bar{b}_t = n/a$, la taille de grappe moyenne. Il est alors possible de calculer la variance effective des unités élémentaires $S_g^2 D_g^2$ et la variance $S_g^2 D_g^2/n$ pour la moyenne de la variable g. Il faut toutefois s'interroger sur la composition de n et de $\bar{b}_t$. Si la population à l'étude est constituée de femmes mariées en âge de procréer, il se peut que cette catégorie de personnes ne représentent que 10% de la population totale et ne se retrouvent que dans 30% des logements et même beaucoup moins en ce qui a trait à quelques rares populations. Cette question a été traitée dans l'échantillonnage pour les caractéristiques peu courantes (Kish 1965). Nous évitons ordinairement les grandes grappes à cause des effets négatifs qu'elles ont sur la variance. Cependant, même de grandes grappes de la population ne produiront que de petites grappes d'individus ayant une caractéristique peu courante, si cette catégorie de personnes sont disséminées un peu partout. Par exemple, on peut échantillonner des îlots complets pour les personnes âgées de plus de 65 ans; on peut aussi sonder des villages complets pour trouver des personnes atteintes d'une maladie bien définie. Par contre, si la caractéristique se retrouve surtout dans de petites régions, il est facile d'identifier ces régions et de les stratifier en conséquence.

Dans les plans à usages multiples, les classes de recoupement de l'échantillon ont des tailles variées qui représentent chacune une partie de la taille n_t de l'échantillon et $\bar{M}_g$ représente la proportion de l'effectif de ces classes par rapport à la population. Ainsi, par le plan de sondage nous voulons estimer $[1 + \rho_g(\bar{b}_t - 1)]$ non seulement pour diverses variables g pour l'échantillon intégral n_t mais aussi pour de nombreuses classes de recoupement. Comme dans la section précédente, l'indice g sert ici à désigner les variables aussi bien que les sous-classes de manière à simplifier la notation. Nous posons ensuite quelques hypothèses qui se sont avérées de bonnes approximations dans des milliers de calcul empiriques pour un très grand nombre d'échantillons:

$$[1 + \rho_g(\bar{b}_g - 1)] \simeq [1 + \rho_g(\bar{M}_g \bar{b}_t - 1)] \simeq [1 + \rho_t(\bar{M}_g \bar{b}_t - 1)] \qquad (2)$$

En d'autres termes, nous utilisons $\bar{b}_g = \bar{M}_g \bar{b}_t$ et $\rho_g \; \rho_t$ comme approximations grossières. Il est vrai que cela contribue à sous-estimer quelque peu les valeurs moyennes de D_g^2 pour les classes de recoupement à cause de la variabilité des tailles de grappes. Cependant, cette sous-estimation a des effets négligeables comparativement à ceux qui découlent des fortes variations de ρ_g d'une variable à l'autre (Kish 1987; Verma et coll. 1980; Kish et coll. 1976), et influe peu sur l'efficacité des plans de sondage. Il est important de considérer l'efficacité des estimations pour les sous-classes comme pour l'échantillon au complet; cet examen révèle une efficacité beaucoup plus forte pour les grandes grappes que celle que nous aurions observée pour $\bar{b}_t$ et n_t si nous nous en étions tenus à l'échantillon proprement dit.

Les mesures de la taille ont un rapport avec la taille des grappes mais varient à cause des erreurs dont sont entachées les mesures existantes, lesquelles erreurs sont surtout attribuables à une variation de la composition de la population et à l'obsolescence des données. Signalons aussi les problèmes relatifs aux mesures de la taille pour les enquêtes portant sur plusieurs sujets à la fois et les opérations d'enquête intégrées pour des populations différentes; ces problèmes peuvent exiger plus

particulièrement des compromis rigoureux. Il est nécessaire de distinguer ces deux niveaux d'usages (tableau 1) parce que les enquêtes portant sur plusieurs sujets utilisent chacune un seul échantillon dans une opération tandis que les opérations d'enquête intégrées peuvent utiliser des unités de sondage de tailles différentes pour des enquêtes différentes (Nations-Unies 1980). Considérons par exemple des plans intégrés pour les populations totales et l'agriculture, peut-être aussi pour les sous-populations ethniques et les activités industrielles ou commerciales: les mesures de la taille pour chacun de ces cas peuvent différer largement l'une de l'autre. Il est néanmoins possible de trouver un solution de compromis de manière à obtenir un niveau d'efficacité raisonnable dans chaque cas.

Les mesures de la taille ont également un rapport étroit avec les problèmes qui se rattachent à la conservation des unités par suite d'une modification des strates et des probabilités (Kish et Scott 1971). Ces méthodes ont été conçues pour tenir compte de l'évolution des unités de sondage tant au point de vue des mesures de la taille qu'au point de vue des variables de stratification, mais elles peuvent aussi tenir compte des différences entre les variables d'enquête:

"On attribue souvent des probabilités d'échantillonnage inégales aux unités de sondage. Bien qu'elles aient une application plus vaste, nos méthodes servent spécialement à choisir des unités primaires d'échantillonnage pour les enquêtes. Le plus souvent, ces unités sont prélevées séparément dans de nombreuses strates à raison d'une unité par strate."

"Après l'échantillonnage initial, les unités peuvent servir à de nombreuses enquêtes pendant plusieurs années mais progressivement, les besoins des nouvelles enquêtes seront probablement mieux servis par des strates et des probabilités d'échantillonnage fondées sur de nouvelles données que par les strates et les probabilités qui ont servi à l'échantillonnage initial. La différence entre les données initiales et les nouvelles données peut être imputable à des changements particuliers dans les unités de sondage, rapportés par le dernier recensement ou à une modification des objectifs et des populations de l'enquête. Par exemple, un échantillon conçu à l'origine pour les ménages et les personnes peut devoir servir ultérieurement à une enquête sur les agriculteurs ou les étudiants de niveau collégial. *Évidemment, nos méthodes peuvent aussi servir à constituer un groupe d'échantillons connexes qui n'ont pas les mêmes objectifs.*"

La méthode proposée permet d'appliquer les meilleures mesures (pour la taille et les strates) pour chaque usage de l'échantillon tout en conservant le plus grand nombre possible d'unités de sondage (surtout des UPÉ) communes à plusieurs échantillons pour des usages différents. Il serait toutefois possible d'imaginer un compromis par lequel on ferait la moyenne des mesures de manière à obtenir un chevauchement complet des unités même si cela devait impliquer une perte d'efficacité pour chacun des usages. Une combinaison des deux méthodes peut s'avérer encore plus efficace, soit accroître le degré de chevauchement avec une faible perte d'efficacité dans chaque cas en considérant uniquement les différences de mesures qui excèdent une valeur minimum arbitraire (Kish et Scott 1971).

8. USAGES ET PLANS DE SONDAGE DES ENQUÊTES PÉRIODIQUES

Les enquêtes périodiques soulèvent des difficultés de plus en plus grandes à mesure que leur nombre et leur importance s'accroissent. Il est faux de supposer que ces enquêtes coûteuses et déterminantes servent à un seul des cinq usages énumérés dans le tableau 4 puisqu'elles sont destinées habituellement à tous ces usages ou à plusieurs d'entre eux si le plan de sondage le permet.

Le tableau 4 contient cinq usages et six plans. Les quatre premiers éléments de chaque groupe sont désignés par des lettres identiques d'une ligne à l'autre. Cette correspondance signale les plans de sondage qui conviennent le mieux (par des variances réduites) à chacun des quatre usages. Comme la plupart des enquêtes périodiques poursuivent plusieurs objectifs, il est normal que nous nous attaquions au délicat problème des plans de sondage à usages multiples. Dans la réalité, n'importe quel des six plans énumérés dans le tableau 4 peut servir à calculer les niveaux courants (A) et les variations nettes (C) mais cela se fait difficilement sans une augmentation de la variance

Tableau 4

Usages et plans de sondage des enquêtes périodiques

Usages	Plans	Mode de renouvellement
A. Niveaux courants	A. Chevauchement partiel $0 < P < 1$	abc-cde-efg
B. Valeurs cumulatives	B. Aucun chevauchement P = 0	aaa-bbb-ccc
C. Variations nettes (moyennes)	C. Chevauchement complet P = 1	aaa-aaa-aaa
D. Variations brutes (valeurs individuelles)	D. Panels	mêmes éléments
E. Séries chronologiques à usages multiples	E. Combinaisons, plan à panel fractionné (PPF)	
	F. Bases principales	

ou des coûts. Par ailleurs, le calcul des variations individuelles (variations brutes ou micro-variations) (D) nécessite des panels tandis que le calcul des valeurs cumulatives (B) nécessite une modification des échantillons et est plus rapide lorsqu'il n'y a pas de chevauchement. En ce qui concerne les niveaux courants (A), il est possible de réduire quelque peu les variances à l'aide d'estimateurs qui exploitent des corrélations découlant de chevauchements partiels. Dans le cas des variations nettes (C), l'estimation profite des corrélations découlant de n'importe quel degré de chevauchement et dans la plupart des autres cas, l'estimation profite des corrélations qui découlent de chevauchements complets (Cochran 1977; Kish 1987; Kish 1965). On peut souvent en arriver à des compromis acceptables lorsqu'on peut définir les usages. Cependant, des considérations accessoires peuvent écarter certains plans (par exemple, les chevauchements peuvent être défendus ou permis) et imposer l'utilisation de plans moins efficaces mais tout aussi acceptables.

Les six plans de sondage du tableau 4 se distinguent surtout par le degré (et le genre) de chevauchement entre les périodes. Dans le cas du chevauchement complet, le mode de renouvellement (aaa-aaa) indique que les portions sont les mêmes pour toutes les périodes; dans le cas où il n'y a aucun chevauchement (aaa-bbb), le mode de renouvellement n'indique aucune portion commune et dans le cas du chevauchement partiel (abc-cde-efg), c et e ne représentent qu'un chevauchement d'un tiers entre des périodes successives. Nous nous intéressons ici plus particulièrement aux effets sur différents usages de la variation du degré de chevauchement P dans divers plans (chevauchement complet: $P = 1$; aucun chevauchement: $P = 0$ et chevauchement partiel: $0 < P < 1$). Les usages sont analysés en fonction des variances des moyennes estimées puisque celles-ci (de même que les pourcentages, les taux et les proportions) sont les estimations les plus courantes et les plus simples. Les effets sur les autres estimations ne seront pas radicalement différents mais ils sont trop nombreux, trop variés et trop complexes pour faire l'objet d'une analyse dans la présente étude.

La question des panels (avantages, inconvénients, problèmes et solutions) est traitée plus en détail dans d'autres ouvrages (Duncan et Kalton 1986; Kish 1987). Nous attirons plus particulièrement l'attention du lecteur sur les PPF, ou plans à panel fractionné, dont nous encourageons l'utilisation comme plans de sondage à usages multiples. Le PPF combine un échantillon constant P avec de nouveaux échantillons avec renouvellement de sorte que *Pa-Pb-Pc-Pd* désignent les échantillons périodiques. Les échantillons avec renouvellement a, b, c, d, etc. peuvent être groupés de manière à former de plus gros échantillons. L'échantillon constant P sert principalement à produire des estimations sur les micro-variations (variations brutes individuelles) mais il assure aussi le chevauchement partiel nécessaire à de meilleures estimations des niveaux courants et des macro-variations (variations nettes — moyennes) *pour n'importe quelle paire de périodes.*

9. CALCUL ET PRÉSENTATION DES ERREURS D'ÉCHANTILLONNAGE

Bien que l'examen de cette question dans le cadre de la présente étude semble discutable, nous sommes convaincus qu'il s'agit là d'un problème qui touche les plans de sondage à usages multiples.

À cette titre, les méthodes de calcul et de présentation des erreurs d'échantillonnage méritent d'être mises en évidence parmi les nombreuses statistiques produites habituellement à la suite d'enquêtes. Il ne suffit pas de présenter l'erreur type d'un seul ou de quelques-uns des paramètres statistiques les plus importants: ils sont trop nombreux et trop variées. C'est pourquoi on en est venu à calculer à l'aide des variances d'autres expressions de la variabilité d'échantillonnage, notamment les estimations des "effets du plan" d_g^2 et parfois, à l'aide de $d_g^2 = 1 + \rho_g(\bar{b}_g - 1)$, les estimations des coefficients de corrélation intraclasse synthétique ρ_g.

En résumé, voici les points à retenir: a) calculer des erreurs d'échantillonnage pour de nombreuses variables parce que les variances, les effets du plan (d_g^2) et les coefficients de corrélation intraclasse (ρ_g) varient énormément d'une variable à l'autre; b) vous devrez peut-être faire des moyennes d'erreurs d'échantillonnage car il pourrait être inopportun ou embarrassant de les présenter toutes; c) il pourrait être ni possible ni nécessaire de calculer des erreurs d'échantillonnage pour toutes les sous-classes car on peut souvent obtenir une approximation de ces erreurs à l'aide de modèles acceptables; d) il est nécessaire de présenter les erreurs d'échantillonnage pour les sous-classes et d'autres paramètres statistiques afin de guider les personnes qui prennent connaissance des rapports (Kish 1965; Kish 1987; Verma et coll. 1980). Nous espérons que ce sujet recevra à l'avenir, de la part des théoriciens et des méthodologistes, toute l'attention qu'il mérite.

10. CONCLUSIONS

Dans les sections 4 à 9, nous avons proposé des approches et des solutions très diversifiées pour les dix sources d'incompatibilité qui avaient été présentées dans la section 3. Dans la section 5, le calcul d'une moyenne des résultats de répartition de la taille de l'échantillon entre les domaines semble produire des solutions de compromis étonnamment bonnes. L'utilisation d'un plus grand nombre de critères de stratification, recommandée dans la section 6, peut aussi procurer des avantages appréciables. Dans les sections 4 et 7, des considérations relatives aux estimations de sous-classe ont abouti à des décisions totalement différentes en ce qui concerne les plans de sondage. Dans la section 8, nous avons vu comment adapter le mieux possible les plans de sondage pour enquêtes périodiques aux objectifs de ces enquêtes et comment ces plans constituent le meilleur compromis pour les enquêtes à usages multiples. Nous avons examiné simultanément les divers niveaux d'usages et les diverses sources d'incompatibilité. Dans tout problème, l'essentiel est de savoir poser la bonne question. Nous proposons donc le plan à usages multiples comme un nouveau modèle destiné à remplacer les solutions "optimales" à des questions arbitrairement partielles comme la suivante: Quelle est la répartition optimale pour la moyenne $\bar{y}$ ou le total $\hat{Y}$ de la variable "la plus importante"?

BIBLIOGRAPHIE

BEAN, J.A., et BURMEISTER, L.F. (1978). A review of optimal sample allocation for multipurpose surveys, *Biometrika*, 20, 3-14.

CHATTERJEE, S. (1967). A note on optimum stratification, *Skandinavisk Actuarietidskrift*, 50, 40-44.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e éd.). New York: John Wiley and Sons, Inc.

DALENIUS, T. (1957). *Sampling in Sweden*, Stockholm: Almquist and Wicksell.

DUNCAN, G.J., et KALTON, G. (1986). Issues of design and analysis of surveys across time. *International Statistical Review*, 54.

FELLEGI, I.P., et SUNTER, A.B. (1974). Balance between different sources of survey errors. *Sankhyā*, 36, 119-142.

FOREMAN, E.K. (1983). Integrated programmes of household surveys: design aspects. *Bulletin of the International Statistical Institute*.

KING, A.J., et JESSEN, R.J. (1945). The master sample of agriculture, *Journal of the American Statistical Association*, 38-56.

KISH, L. (1987). *Statistical Design for Research*: New York: Wiley-Interscience.

KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 28, 1-12.

KISH, L. (1980). Design and estimation for domains. *The Statistician*, London, 29, 209-222.

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

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

KISH, L. (1961). Efficient allocation for multipurpose samples. *Econometrica*, 29, 363-385.

KISH, L., et ANDERSON, D.W. (1978). Multivariate and multipurpose stratification, *Journal of the American Statistical Association*, 73, 24-34.

KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, B. 36, 1-37.

KISH, L., et SCOTT, A.J. (1971). Retaining units after changing strata and probabilities, *Journal of the Royal Statistical Society*, 66, 461-470.

KIREGYERA, B., et GACHUKI, P. (1985). Experiences in panel surveys: examples from an integrated sample survey programme in Kenya. *Bulletin of the International Statistical Institute*.

MACURA, M., et CLELAND, J. (1985). *A Celebration of Statistics: the ISI Centenary Volume*, (éds. A.C. Atkinson et S.E. Fienberg), New York: Springer Verlag.

MURTHY, M.N. (1974). Evaluation of multi-subject sample survey systems. *International Statistical Review*, 42.

MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

NATIONS-UNIES (1980). *National Household Survey Capability Programme* (NHSCP). New York: United Nations.

RODRIQUEZ-VERA, A. (1982). *Multipurpose Optimal Sample Allocation Using Mathematical Programming*. Thèse de doctorat, The University of Michigan, Ann Arbor.

VERMA, V., SCOTT, C., et O'MUIRCHEARTAIGH, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, A, 143, 431-473.

YATES, F. (1981). *Sampling Methods for Censuses and Surveys*, (4^{e} éd.). London: Griffin and Co.

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 35–45
Statistique Canada

Sur la stratification de populations asymétriques

PIERRE LAVALLÉE et MICHEL A. HIDIROGLOU[1]

RÉSUMÉ

Pour un degré de précision donné, Hidiroglou (1986) a défini un algorithme permettant de diviser la population en une strate à tirage complet et en une strate à tirage partiel de manière à minimiser la taille de l'échantillon global en supposant un échantillonnage aléatoire simple sans remise dans la strate à tirage partiel. Sethi (1963) a proposé un algorithme permettant un découpage optimal de la population en un certain nombre de strates à tirage partiel. Dans cet article, il est question d'un algorithme itératif qui vise à déterminer les bornes de strates pour une population fortement asymétrique découpée en une strate à tirage complet et en un certain nombre de strates à tirage partiel. Ces bornes de strates sont calculées de manière à minimiser la taille de l'échantillon global étant donné un degré de précision relative, un échantillonnage aléatoire simple sans remise dans les strates à tirage partiel et une répartition à la puissance "p" de l'échantillon entre ces mêmes strates. L'algorithme présenté dans cet article est une combinaison des travaux d'Hidiroglou (1986) et de Sethi (1963).

MOTS CLÉS: Algorithme itératif; bornes optimales; tirage complet; tirage partiel.

1. INTRODUCTION

Dans le cas de populations fortement asymétriques comme celles que l'on retrouve dans les enquêtes entreprises, un échantillonnage efficace n'est possible qui si l'on découpe ces populations en une strate à tirage complet et en un certain nombre de strates à tirage partiel. Toutes les unités de la strate à tirage complet sont prélevées avec une probabilité égale à un (1) tandis que les unités des strates à tirage partiel sont prélevées suivant un modèle probabiliste. Glasser (1962) et Hidiroglou (1986) ont établi des règles approximatives pour le découpage d'une population en une strate à tirage complet et en une strate à tirage partiel. Glasser (1962) a déterminé l'emplacement de la coupure entre les strates en supposant qu'un échantillon de taille déterminée devait être prélevé dans les deux strates et que dans le cas de la strate à tirage partiel, il devait s'agir d'un échantillonnage aléatoire simple sans remise. Pour sa part, Hidiroglou (1986) a déterminé l'emplacement de la coupure entre les strates en supposant qu'il fallait respecter un degré de précision voulu. Ces deux approches sont duales en ce sens que Glasser cherche à minimiser la variance d'échantillonnage pour une taille d'échantillon donnée tandis que Hidiroglou cherche à minimiser la taille d'échantillon pour une variance d'échantillonnage donnée.

Dans cet article, nous proposons un algorithme qui vise à découper une population fortement asymétrique en une strate à tirage complet et en un certain nombre de strates à tirage partiel. Par cet algorithme, nous cherchons à minimiser la taille de l'échantillon global étant donné le coefficient de variation de l'estimateur et le mode de répartition de l'échantillon entre les strates à tirage partiel. Les bornes de strates sont déterminées en fonction d'une variable auxiliaire qui est en corrélation étroite avec les renseignements recueillis dans l'enquête. Si, par exemple, le chiffre des ventes annuelles est une des variables mesurées dans un recensement des détaillants, cette variable auxiliaire peut servir à déterminer les bornes de strates pour une enquête spécialisée qui

[1] Pierre Lavallée, méthodologiste, et Michel A. Hidiroglou, chef, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario) K1A 0T6, Canada. Les auteurs tiennent à remercier France Bilocq de la Division des méthodes d'enquêtes-entreprises de Statistique Canada pour avoir programmé les exemples.

vise à recueillir des données sur les ventes mensuelles. Pour ce qui a trait aux enquêtes polyvalentes, si l'on utilise des variables auxiliaires qui ne sont pas en corrélation étroite avec la variable principale, les bornes de strates ne seront plus aussi optimales. L'algorithme que nous proposons est une version modifiée de la méthode de stratification de Sethi (1963). Les bornes de strates qui découlent de cet algorithme sont optimales et déterminent la taille d'échantillon minimum requise.

Le mode de répartition que nous avons choisi pour illustrer la méthode est la répartition à la puissance p. L'utilisation de ce mode de répartition permet de publier des estimations de strates dont les coefficients de variation ne sont pas très différents les uns des autres. La répartition à la puissance p a été proposée par Carrol (1970), Fellegi (1981) et Bankier (1988). En pratique, elle est un compromis entre la répartition de Neyman et la nécessité d'avoir des coefficients de variation égaux pour chaque strate. Un inconvénient de la répartition de Neyman est qu'elle produit des coefficients de variation très différents d'une strate à l'autre si l'on doit avoir des estimations pour chaque strate. Par ailleurs, un mode de répartition qui produit les mêmes coefficients de variation d'une strate à l'autre peut exiger un échantillon beaucoup plus grand que celui qu'exige la répartition de Neyman. Dans le cadre de notre analyse, la répartition à la puissance p nous permettra de produire des estimations pour des strates d'entreprises de tailles variées (petites, moyennes et grandes) avec des coefficients de variation similaires.

Nous comparerons la méthode élaborée dans cet article, au point de vue des bornes et de la taille d'échantillon, à la règle de la racine de f cumulative de Dalenius — Hodges (1959) ainsi qu'à une combinaison des méthodes de stratification d'Hidiroglou (1986) et de Dalenius — Hodges (1959). L'algorithme, qui est récursif par définition, est facile à programmer et converge rapidement vers les bornes optimales. Il permet également une réduction sensible de la taille d'échantillon pour des critères de fiabilité donnés.

2. LE PROBLÈME

Considérons une population ordonnée finie de N unités:

$$y_{(1)}, y_{(2)}, \ldots, y_{(N)},$$

où $y_{(i)} \leq y_{(i+1)}$ pour $i = 1, 2, \ldots, N-1$. Cette population est découpée en L strates. Le nombre d'unités dans chaque strate est désigné par N_h, $h = 1, 2, \ldots, L$. Selon le plan d'échantillonnage, n_h unités doivent être prélevées dans chaque strate à tirage partiel de taille N_h ($h = 1, 2, \ldots, L-1$) au moyen d'un échantillonnage aléatoire simple sans remise; notons par ailleurs que $n_L = N_L$. La moyenne à estimer est

$$\bar{Y} = \sum_{h=1}^{L} \sum_{j=M_{h-1}+1}^{M_h} y_{(j)}/N, \tag{2.1}$$

où $M_h = \sum_{i=1}^{h} N_i$ pour $h = 1, 2, \ldots, L$ et $M_o = 0$.

Compte tenu de cette définition, l'estimateur de la moyenne de population $\bar{Y}$ est

$$\hat{\bar{Y}} = \left[\sum_{h=1}^{L-1} \frac{N_h}{n_h} \sum_{j=m_{h-1}+1}^{m_h} z_j + \sum_{j=M_{L-1}+1}^{N} y_{(j)} \right] /N \tag{2.2}$$

où $y_{M_{h-1}+1} \leq z_j \leq y_{M_h}$ pour $j = m_{h-1}+1, \ldots, m_h$ $(h = 1, 2, \ldots, L-1)$, $m_h = \sum_{i=1}^{h} n_i$ pour $h = 1, 2, \ldots, L$ et $m_o = 0$.

Supposons que le degré de précision voulu pour la moyenne estimée est défini par c (coefficient de variation) et que la proportion d'unités échantillonnées devant être réparties entre les $L-1$ premières strates est a_h $(h = 1, 2, \ldots, L-1)$ où $\sum_{h=1}^{L-1} a_h = 1$. Le terme "a_h" représente en fait n'importe quel mode de répartition entre les strates. Dans le cas, par exemple, de la répartition proportionnelle à N à la puissance p,

$$a_h = \frac{N_h^P}{\sum_{h=1}^{L-1} N_h^P} \quad (h = 1, 2, \ldots, L-1)$$

et dans le cas de la répartition proportionnelle à Y à la puissance p,

$$a_h = \frac{Y_h^P}{\sum_{h=1}^{L-1} Y_h^P},$$

où $0 < p < \infty$. Étant donné des hypothèses relativement simples et une valeur de p appropriée, la répartition à la puissance p a la propriété d'engendrer des coefficients de variation relativement uniformes pour les strates à tirage partiel sans accroître notablement le coefficient de variation global. Cette uniformité des coefficients de variation est souvent souhaitée par les utilisateurs des données d'enquête.

En pratique, on choisit souvent 1/2 ou 1/3 comme valeur de p. Une faible valeur de p (c.-à-d., une valeur proche de 0) donne habituellement des coefficients de variation comparables d'une strate à l'autre tandis qu'une valeur plus élevée élargit l'écart entre les coefficients mais accroît du même coup la précision des estimations globales.

Soulignons également que la répartition à la puissance p définie ci-dessus équivaut au mode proposé par Bankier (1988) lorsque les coefficients de variation des strates à tirage partiel sont égaux.

La variance de $\hat{\bar{Y}}$ est

$$V(\hat{\bar{Y}}) = \frac{1}{N^2} \sum_{h=1}^{L-1} \frac{N_h}{n_h} (N_h - n_h) S_h^2, \tag{2.3}$$

où S_h^2 désigne la variance de population de chaque strate h. Si l'on tient compte du degré de précision voulu (coefficient de variation c), la variance de Y peut être redéfinie $V(\hat{\bar{Y}}) = c^2 \bar{Y}^2$. En remplaçant n_h et $V(\hat{\bar{Y}})$ dans l'équation (2.3) par $(n - N_L)\, a_h$ et $c^2 \bar{Y}^2$ respectivement et en isolant n, on obtient l'équation

$$n = N_L + \frac{\sum_{h=1}^{L-1} N_h^2 S_h^2 / a_h}{(N c \bar{Y})^2 + \sum_{h=1}^{L-1} N_h S_h^2}. \tag{2.4}$$

Le problème consiste à déterminer des bornes $b_{(1)}, b_{(2)}, \ldots, b_{(L-1)}$ (où $y_{(1)} < b_{(1)} < \ldots < b_{(L-1)} < y_{(N)}$) de manière à minimiser la taille n de l'échantillon global, étant donné le degré de fiabilité c et le mode de répartition spécifié (représenté par a_h).

3. L'ALGORITHME

La méthode utilisée dans le présent article pour calculer des bornes de strates pour un degré de précision voulu a déjà été utilisée par Dalenius (1950) pour calculer des bornes de strates pour une taille d'échantillon donnée. On suppose en premier lieu que l'échantillonnage se fait à partir d'une population dont la distribution de fréquences peut être représentée avec suffisamment de précision par une fonction continue de densité $f(y)$. Ainsi, pour un ensemble donné de limites $b_{(1)}, \ldots, b_{(L-1)}$ nous définissons les quantités suivantes:

$$W_h = \int_{b_{(h-1)}}^{b_{(h)}} f(y)\, dy, \tag{3.1}$$

$$\mu_h = \int_{b_{(h-1)}}^{b_{(h)}} y f(y)\, dy / W_h, \tag{3.2}$$

$$\sigma_h^2 = \int_{b_{(h-1)}}^{b_{(h)}} y^2 f(y)\, dy / W_h - \mu_h^2, \tag{3.3}$$

pour $h = 1, \ldots, L$, où $b_{(o)} = -\infty$, $b_{(L)} = +\infty$.

On peut alors reformuler l'équation (2.4) comme suit:

$$n = NW_L + \frac{N\left(\sum_{h=1}^{L-1} W_h^2 \sigma_h^2 / a_h\right)}{N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2}, \tag{3.4}$$

où

$$\mu = \int_{b_{(o)}}^{b_{(L)}} y f(y)\, dy.$$

Il convient de souligner que même s'il s'agit d'une population de grande taille, le facteur de correction pour population finie (cpf) est toujours présent dans l'équation (3.4) - voir Dalenius-Gurney (1951). Par définition, la strate à tirage complet doit avoir une population finie si l'on veut que la taille d'échantillon soit finie. En outre, le fait d'ignorer le facteur de c.p.f. ne se traduirait pas par une variance nulle pour la strate à tirage complet.

Le terme "a_h" dans l'équation (2.3) peut aussi être exprimé à l'aide des quantités (3.1, (3.2) et (3.3). Dans le cas de la répartition proportionnelle à N à la puissance p, nous avons:

$$a_h = \frac{W_h^p}{\sum_{h=1}^{L-1} W_h^p}, \tag{3.5}$$

pour $h = 1, .., L-1$.

Pour ce qui a trait à la répartition proportionnelle à Y à la puissance p, nous avons l'équation suivante:

$$a_h = \frac{(W_h \mu_h)^p}{\sum_{h=1}^{L-1} (W_h \mu_h)^p}, \tag{3.6}$$

où $0 < p < \infty$.

Dans cet article, nous considérerons surtout le répartition proportionnelle à Y mais les calculs peuvent être effectués aussi bien pour la répartition proportionnelle à N et, de fait, pour n'importe quel mode de répartition représenté par le terme a_h où $\sum_{h=1}^{L-1} a_h = 1$. Si nous remplaçons a_h dans l'équation (3.4) par l'expression définie en (3.6), nous obtenons

$$n = N W_L + \frac{N \left[\sum_{h=1}^{L-1} (W_h \sigma_h)^2 (W_h \mu_h)^{-p} \right] \left[\sum_{h=1}^{L-1} (W_h \mu_h)^p \right]}{N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2}. \tag{3.7}$$

Afin de déterminer les bornes optimales $b_{(1)}, \ldots, b_{(L-1)}$ de manière que la taille n de l'échantillon soit minimale, on calcule les dérivées partielles de l'équation (3.7) par rapport à $b_{(1)}, \ldots, b_{(L-1)}$, respectivement et on les pose égales à zéro. On obtient alors les équations suivantes:

Pour $h = 1, \ldots, L-2$,

$$[F T_h - F T_{h+1}]\ b_{(h)}^2 +$$

$$[F K_h - 2\mu_h F T_h - F K_{h+1} + 2\mu_{h+1} F T_{h+1} + 2\mu_h AB - 2\mu_{h+1} AB]\ b_{(h)} +$$

$$[F T_h \mu_h^2 + F T_h \sigma_h^2 - F T_{h+1} \mu_{h+1}^2 - F T_{h+1} \sigma_{h+1}^2 - AB\mu_h^2 + AB\mu_{h+1}^2] = 0, \tag{3.8}$$

et pour $h = L-1$,

$$[F T_{L-1} - AB]\ b_{(L-1)}^2 +$$

$$[F K_{L-1} - 2\mu_{L-1} F T_{L-1} + 2\mu_{L-1} AB]\ b_{(L-1)} +$$

$$[F T_{L-1} \mu_{L-1}^2 + F T_{L-1} \sigma_{L-1}^2 - AB\mu_{L-1}^2 - F^2] = 0, \tag{3.9}$$

où

$$A = \sum_{h=1}^{L-1} (W_h \mu_h)^p,$$

$$B = \sum_{h=1}^{L-1} (W_h \sigma_h)^2 (W_h \mu_h)^{-p},$$

$$F = N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2,$$

$$K_h = B p (W_h \mu_h)^{p-1} - A p (W_h \sigma_h)^2 (W_h \mu_h)^{-p-1},$$

$$T_h = A W_h (W_h \mu_h)^{-p}.$$

Si nous identifions le coefficient de $b_{(h)}^2$ par α_h, le coefficient de $b_{(h)}$ par β_h et l'ensemble des autres termes par γ_h, nous pouvons représenter les équations (3.8) et (3.9) comme des équations quadratiques de la forme $\alpha_h b_{(h)}^2 + \beta_h b_{(h)} + \gamma_h = 0$. Or, comme le souligne Sethi (1963), les termes α_h, β_h et γ_h sont eux-mêmes des fonctions de $b_{(1)}, \ldots, b_{(L-1)}$ par les intégrales (3.1), (3.2) et (3.3). En adoptant l'approche de Sethi (1963), nous pouvons résoudre facilement les équations (3.8) et (3.9) à l'aide de la méthode itérative suivante:

ÉTAPE 1 : Prendre des bornes arbitraires $b_{(1)}' < \ldots < b_{(L-1)}'$.

ÉTAPE 2 : Calculer les proportions W_h', les moyennes μ_h' et les variances $\sigma_h^{2'}$ (à l'aide des équations (3.1), (3.2) et (3.3), respectivement) fondées sur ces bornes, $h = 1, \ldots, L-1$.

ÉTAPE 3 : Remplacer la série de bornes initiale par $b_{(1)}'', \ldots, b_{(L-1)}''$ où

$$b_{(h)}'' = \frac{-\alpha_h' + \sqrt{\beta_h'^2 - 4\alpha_h'\gamma_h'}}{2\alpha_h'}, \quad h = 1, \ldots, L-1. \tag{3.10}$$

ÉTAPE 4 : Reprendre les étapes 2 et 3 jusqu'à ce qu'on obtienne deux séries consécutives identiques ou très semblables, c'est-à-dire

$$\max_{h=1}^{L-1} | b_{(h)}'' - b_{(h)}' | < \epsilon \text{ pour une valeur } \epsilon > 0. \tag{3.11}$$

Soulignons qu'il est possible de démontrer que le signe précédant la racine carrée ($\sqrt{\ }$) est positif parce que $b_{(h)}'$ se situe entre μ_h' et μ_{h+1}'.

L'inconvénient de cet algorithme est que son utilisation exige que l'on dispose de renseignements sur la fonction de densité approximative $f_{(y)}$. Comme la population considérée est finie, nous pouvons éliminer cet inconvénient en remplaçant les quantités (3.1), (3.2) et (3.3) par les expressions correspondantes fondées sur une population finie. Ainsi, conformément au procédé appliqué dans Cochran (1977), nous pouvons remplacer les paramètres de population infinie donnés par les expressions (3.1), (3.2) et (3.3) par les paramètres correspondants pour population finie, c'est-à-dire:

$$W_h = \frac{N_h}{N}, \tag{3.12}$$

$$\bar{Y}_h = \frac{1}{N_h} \sum_{\substack{j=b_{(h-1)}+1}}^{b_{(h)}} y_{(j)}, \tag{3.13}$$

$$S_h^2 = \frac{1}{N_{h-1}} \sum_{j=b_{(h-1)}+1}^{b_{(h)}} y_{(j)}^2 - N_h \bar{Y}_h^2, \tag{3.14}$$

pour $h = 1, \ldots, L$.

Ces formules peuvent alors nous permettre de déterminer sans difficulté à l'aide de la méthode itérative suivante les bournes de strates $b_{(1)}, \ldots, b_{(L-1)}$ de manière que la taille n de l'échantillon global minimum pour un degré de fiabilité (c) et un mode de répartition données:

ÉTAPE 0 : Classer les éléments de la population $y_1, \ldots, y_N$ par ordre croissant et poser $b_{(0)} = y_{(1)}$ et $b_{(L)} = y_{(N)}$.

ÉTAPE 1 : Choisir des bornes arbitraires de telle sorte que $b_{(0)} < b'_{(1)} < \ldots < b'_{(L-1)} < b_{(L)}$.

ÉTAPE 2 : Calculer la proportion W'_h, la moyenne $\bar{Y}'_h$ et la variance $S_h^{2'}$ (à l'aide des équations (3.12), (3.13) et (3.14) respectivement) fondées sur ces bornes, $h = 1, \ldots, L$-1.

ÉTAPE 3 : Remplacer la série de bornes initiale par $b''_{(1)}, \ldots, b''_{(L-1)}$ où

$$b''_{(h)} = \frac{-\alpha'_h + \sqrt{\beta'^2_h - 4\,\alpha'_h\,\gamma'_h}}{2\,\alpha'_h}, \quad h = 1, \ldots, L-1.$$

ÉTAPE 4: Reprendre les étapes 2 et 3 jusqu'à ce qu'on obtienne deux séries consécutives identiques ou très semblables, c'est-à-dire

$$\max_{h=1}^{L-1} | b''_{(h)} - b'_{(h)} | < \epsilon \text{ pour une valeur } \epsilon < 0.$$

Dans la section qui suit, nous comparons l'application de cet algorithme à celle d'autres méthodes.

4. EXEMPLES

Afin d'illustrer l'algorithme décrit dans la section 3, nous allons nous servir de données tirées des enquêtes annuelles sur le commerce de détail et le commerce de gros réalisées par Statistique Canada. Ces enquêtes permettent d'évaluer le chiffre des ventes des entreprises spécialisées dans le commerce de détail ou le commerce de gros. Nous utilisons trois populations pour illustrer l'algorithme. Il s'agit des grossistes en produits divers du Québec (population 1), des grossistes en produits alimentaires divers du Manitoba (population 2) et des magasins de vente au détail d'appareils ménagers et d'appareils électroniques du Québec (population 3). Le choix des populations s'est fait de manière à obtenir un éventail des tailles de population: grande, moyenne et petite. Les coefficients d'asymétrie pour ces populations sont 24.2 (population 1), 6.5 (population 2) et 13.6 (population 3).

Nous allons comparer les résultats de l'algorithme proposé à ceux de deux autres méthodes. La première de ces méthodes consiste simplement à stratifier la population au moyen de la règle de la racine de f cumulative de Dalenius — Hodges (1959). La seconde méthode consiste à déterminer l'emplacement de la coupure entre la strate à tirage complet et la strate à tirage partiel au

moyen de la formule d'approximation d'Hidiroglou (1986), puis à appliquer la règle de la racine de f cumulative pour découper la population qui n'est pas incluse dans la strate à tirage complet en un certain nombre de strates à tirage partiel. Nous désignerons respectivement ces méthodes comme la règle $f^{1/2}$ CUM et la méthode combinée; quant à l'algorithme proposé, nous le désignerons comme la méthode optimale. Il n'est pas réaliste d'utiliser uniquement la méthode de Dalenius-Hodges (1959) car, dans la pratique, celle-ci ne serait appliquée qu'une fois que la strate à tirage complet aurait été définie à l'aide d'une règle arbitraire donnée. Néanmoins, nous illustrons ici l'application de cette méthode pour mettre en garde le lecteur contre une utilisation inconsidérée de cette méthode pour des populations fortement asymétriques.

Hidiroglou (1986) détermine l'emplacement de la coupure entre les strates au moyen de la formule itérative suivante:

$$b_{TA}'' = \mu_{[N-t']} = \left\{ \frac{N-t'-1}{(N-t')^2} N^2 c^2 \bar{Y}^2 + S^2_{[N-t']} \right\}^{1/2}, \tag{4.1}$$

où

$$\mu_{[N-t']} = \frac{1}{N-t'} \sum_{i=1}^{N-t'} y_{(i)} \tag{4.2}$$

Tableau 1

Effet de la variation du coefficient de variation et de la puissance p sur la taille d'échantillon pour trois méthodes de stratification
(population 1 — taille = 1221)

			Méthode de stratification								
			Règle $f^{1/2}$ Cum			Composée			Optimale		
c	p	Strates	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$
0.05	0.25	1	1196	177*		1017	16		891	11	
		2	20	20	3,715,320	152	14	465,180	290	13	302,912
		3	5	5	14,786,280	52	52	1,131,961	40	40	1,835,930
		Total		202			82			64	
0.05	0.50	1	1196	178*		1017	16		863	10	
		2	20	20	3,715,320	152	13	465,180	318	14	289,422
		3	5	5	17,786,280	52	52	1,131,961	40	40	1,832,038
		Total		203			81			64	
0.01	1.00	1	1196	616*		751	37		687	36	
		2	20	20	3,715,320	215	34	196,840	374	78	162,068
		3	5	5	14,786,280	255	255	383,033	160	160	564,076
		Total		641			326			274	
0.05	1.00	1	1196	180*	3,715,320	1017	16		858	8	
		2	20	20	14,786,280	152	11	465,180	323	16	271,920
		3	5	5		52	52	1,131,961	40	40	1,867,254
		Total		205			79			64	
0.10	1.00	1	1196	56*		1073	7		1007	7	
		2	20	20	3,715,320	109	4	592,900	191	9	442,357
		3	5	5	14,786,280	39	39	1,953,113	23	23	4,032,950
		Total		81			50			39	

*Une répartition disproportionnée est nécessaire pour satisfaire au coefficient de variation.

Tableau 2

Effet de la variation du coefficient de variation et de la puissance p sur la taille d'échantillon pour trois méthodes de stratification
(population 2 — taille = 44)

			Méthode de stratification								
			Règle $f^{1/2}$ Cum			Composée			Optimale		
c	p	Strates	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$
0.05	0.25	1	42	38		32	1		29	1	
		2	1	1*	137,939,900	6	1	4,708,409	11	1	3,029,455
		3	1	1	459,739,000	6	6	10,622,301	4	4	17,461,464
		Total		40			8			6	
0.05	0.50	1	42	38		32	1		28	1	
		2	1	1*	137,939,900	6	1	4,708,409	12	1	2,582,819
		3	1	1	459,739,000	6	6	10,622,301	4	4	17,640,325
		Total		40			8			6	
0.01	1.00	1	42	42		25	1		25	1	
		2	1	1	137,939,900	5	1	1,059,550	10	4	1,153,322
		3	1	1	459,739,000	14	14	3,742,377	9	9	5,969,271
		Total		44			16			14	
0.05	1.00	1	42	38		32	1		26	1	
		2	1	1*	137,939,900	6	1	4,708,409	14	2	1,779,500
		3	1	1	459,739,000	6	6	10,622,301	4	4	17,349,902
		Total		40			8			7	
0.10	1.00	1	42	30		34	1		28	1	
		2	1	1*	137,939,900	6	1	4,848,218	13	1	2,413,800
		3	1	1	459,739,000	4	4	16,749,625	3	3	30,091,449
		Total		32			6			5	

*Une répartition disproportionnée est nécessaire pour satisfaire au coefficient de variation.

et

$$S^2_{[N-t']} = \frac{1}{N-t'-1} \sum_{i=1}^{N-t'} (y_{(i)} - \mu_{[N-t']})^2.$$

Le nombre d'unités à tirage complet obtenues pour chaque étape du processus itératif est t'. Le point de départ de cette approximation est

$$b'_{TA} = \mu_{[N]} + \{N c^2 \bar{Y}^2 + S^2_{[N]}\}^{1/2} . \tag{4.3}$$

Le processus itératif cesse lorsque l'inéquation suivante est satisfaite:

$$0 \leq 1 - n(t'')/n(t') < 0.10 \tag{4.4}$$

Tableau 3

Effet de l'accroissement du nombre de strates sur la taille d'échantillon pour deux méthodes de stratification
$p = 1, c = 0.05$

Population 1 ($N = 1221$) Méthode de stratification	Strates	Nombre de strates: 3 N_h	3 n_h	3 $b_{(h)}$	4 N_h	4 n_h	4 $b_{(h)}$	5 N_h	5 n_h	5 $b_{(h)}$
Composée	1	1017	16		897	6		823	3	
	2	152	11	465,180	194	5	311,117	194	2	245,090
	3	52	52	1,131,961	78	4	641,252	101	2	465,180
	4				52	52	1,131,961	51	2	751,297
	5							52	52	1,131,961
	Total		79			67			61	
Optimale	1	858	8		704	3		655	2	
	2	323	16	271,920	373	7	173,981	358	4	149,327
	3	40	40	1,867,254	112	6	604,869	163	5	453,114
	4				32	32	2,676,449	29	4	1,522,329
	5							16	16	5,810,487
	Total		64			48			31	
Population 3 ($N = 161$)										
Composée	1	106	6		84	2		71	1	
	2	39	6	265,480	38	2	185,320	35	1	155,260
	3	16	16	553,255	23	2	335,620	22	1	265,480
	4				16	16	553,255	17	1	385,720
	5							16	16	553,255
	Total		28			22			20	
Optimale	1	86	4		55	1		34	1	
	2	65	9	199,415	61	3	125,572	51	1	83,594
	3	10	10	680,942	39	5	312,769	42	2	192,215
	4				6	6	826,942	29	3	382,236
	5							5	5	906,894
	Total		23			15			12	

où

$$n(t') = t' + \frac{(N-t')^2 S^2_{[N-t']}}{(N c \bar{Y})^2 + (N-t') S^2_{[N-t']}} . \qquad (4.5)$$

Les tableaux 1 et 2 donnent les résultats pertinents pour une grande population (population 1) et une petite population (population 2) étant donné divers coefficients de variation et diverses valeurs de p (pour la répartition à la puissance p). Le tableau 3 donne les résultats obtenus pour la grande population (population 1) et une population de taille moyenne (population 3) lorsqu'on modifie le nombre de strates. Pour les trois tableaux, la répartition de l'échantillon entre les strates à tirage partiel s'est faite suivant le mode de répartition proportionnelle à Y à la puissance p.

Les tableaux 1 et 2 nous permettent de tirer les conclusions suivantes. La règle de la racine de f cumulative est très inefficace dans les présentes circonstances. La méthode composée réduit

de façon substantielle le taille de l'échantillon. L'application de la méthode optimum entraîne une réduction additionnelle de l'ordre de 20% pour les trois strates figurant dans ces deux tableaux. Pour un coefficient de variation donné, la variation de la puissance "p" influe peu sur la taille de l'échantillon. Comme prévu, la taille d'échantillon augmente lorsque le coefficient de variation (c) diminue (pour une même valeur de p). La méthode optimale définit moins d'unités à tirage complet (strate 3) que la méthode composée ou, en d'autres termes, la borne de la strate à tirage complet est plus élevée pour la première méthode que pour la seconde. La règle de la racine de f cumulative n'est plus aussi efficace lorsqu'il s'agit de déterminer la borne de la strate à tirage complet. Il est facile de voir que la borne calculée à l'aide de cette règle est considérablement plus élevée que celles calculées à l'aide des autres méthodes.

Dans le tableau 3, nous comparons les résultats obtenus par la méthode composée et la méthode optimale pour deux populations lorsque le nombre de strates varie, étant donné un coefficient de variation et une valeur de p (pour la répartition proportionnelle à Y à la puissance p). Des conclusions semblables à celles tirées des deux premiers tableaux se dégagent du tableau 3. L'accroissement du nombre de strates a pour effet de réduire le nombre d'unités échantillonnées, peu importe la méthode utilisée. Toutefois, la réduction est plus prononcée dans le cas de la méthode optimale.

5. CONCLUSION

Le découpage optimal d'une population asymétrique en une strate à tirage complet et en un certain nombre de strates à tirage partiel a permis de réduire sensiblement la taille de l'échantillon global pour un degré de précision relative donné. La méthode proposée peut être adaptée à n'importe quel mode de répartition et à n'importe quel nombre de strates. On peut également faire abstraction de la condition relative au tirage complet.

L'algorithme, qui est récursif par définition, converge rapidement. Il peut être appliqué facilement par ordinateur par l'intermédiaire du SAS, du FORTRAN ou de n'importe quel autre langage évolué.

BIBLIOGRAPHIE

BANKIER, M.D. (1988). Power allocations, determining sample sizes for sub-national areas. À paraître dans *The American Statistician*.

CARROL, J. (1970). Allocation of a sample between States. Note de service non publiée de l'Australian Bureau of Census and Statistics.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e éd.). New York: John Wiley & Sons.

DALENIUS, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, 33, 203-213.

DALENIUS, T., et GURNEY, M. (1951). The problem of optimum stratification.II, *Skandinavisk Aktuarietidskrift*, 34, 133-148.

DALENIUS, T., et HODGES, J.L.Jr. (1959). Minimum variance stratification, *Skandinavisk Aktuarietidskrift*, 54, 88-101.

FELLEGI, I.P. (1981). Should the census counts be adjusted for allocation purposes? — Equity considerations. Dans *Current Topics in Survey Sampling* (éds. D. Krewski, R. Platek et J.N.K. Rao), New York: Academic Press, 47-76.

GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Revue Internationale de Statistique*, 30, 28-32.

HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design, *The American Statistican*, 40, 27-31.

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

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 47–62
Statistique Canada

Domaines de recherche relatifs à la SIPP (enquête sur le revenu et la participation aux programmes)[1]

DANIEL KASPRZYK[2]

RÉSUMÉ

La SIPP (Survey of Income and Program Participation/enquête sur le revenu et la participation aux programmes) est une enquête permanente menée par le U.S. Bureau of the Census auprès d'un échantillon de ménages représentatif de la population à l'échelle nationale. Le but principal de la SIPP est d'améliorer la mesure de l'information sur la situation économique des ménages et des particuliers aux États-Unis. Pour atteindre ce but, des interviews sont menées à plusieurs reprises auprès des individus échantillonnés. Elles sont basées sur un questionnaire contenant des questions d'approfondissement dont la période de référence est courte. Le plan de sondage à plusieurs interviews de la SIPP soulève des questions d'ordre méthodologique et statistique qui concernent toutes les enquêtes par panel menées auprès de familles et de particuliers. Ces questions sont traitées dans le présent article du point de vue de la SIPP. Il s'agit: 1) de la conception du questionnaire; 2) de la collecte des données, notamment des règles de conduite adoptées à l'égard des répondants, du mode de collecte des données, de la durée de la période de référence et des règles de conduite adoptées à l'égard des personnes ayant déménagé; 3) des définitions, du plan de sondage et des estimations; 4) de l'erreur de réponse.

MOTS CLÉS: Enquêtes par panel; conception du questionnaire; plan de sondage; estimations longitudinales; erreur de réponse.

1. INTRODUCTION

La SIPP (Survey of Income and Program Participation/enquête sur le revenu et la participation aux programmes) est une nouvelle enquête permanente menée par le U.S. Bureau of the Census auprès d'un échantillon de ménages représentatif de la population à l'échelle nationale. Elle permet de recueillir une information complète et détaillée au sujet des ressources économiques des ménages américains et de l'incidence que les programmes fiscaux et les programmes de transfert de l'administration publique peuvent avoir sur la situation financière de ces derniers. Les responsables des politiques fédérales se basent sur les données tirées de l'enquête pour déterminer l'efficacité des programmes fiscaux et des programmes de transfert, estimer les coûts futurs des programmes et le nombre de bénéficiaires éventuels ainsi que pour évaluer les répercussions de toute modification envisagée. Le but de la SIPP est d'améliorer la mesure de l'information sur la situation économique des ménages et des particuliers aux États-Unis. C'est l'aboutissement d'un vaste programme d'élaboration, l'ISDP (Income Survey Development Program/programme de mise au point de l'enquête sur le revenu) dans le cadre duquel on s'est penché sur les définitions, les méthodes, les questionnaires, les périodes de référence et d'autres éléments de nature semblable. (Ycas et Lininger 1981).

La SIPP a été créée parce qu'il fallait combler les lacunes du supplément sur le revenu greffé chaque année au mois de mars à la CPS (Current Population Survey/enquête sur la population), lequel était jusqu'alors la principale source de renseignements sur la répartition du revenu des ménages et des particuliers aux États-Unis. Ces lacunes étant liées au plan de sondage, à

[1] Cet article contient les résultats des travaux de recherche effectués par le personnel du Bureau of the Census. Les opinions qui y sont exprimées sont entièrement celles de l'auteur et ne reflètent pas nécessairement la position du Bureau of the Census.

[2] Daniel Kasprzyk, SIPP Research and Coordination Staff, United States Bureau of the Census, Washington D.C. 20233

l'instrument et aux méthodes d'enquête, il était difficile d'y remédier. En 1975, le Département américain de la santé et des services sociaux a donc décidé de mettre sur pied l'ISDP pour corriger les principaux défauts de la CPS, en l'occurrence: 1) la sous-déclaration du revenu de biens et d'autres sources moins fréquentes de revenu, 2) les erreurs de déclaration et de classification ayant trait aux prestations des principaux programmes de sécurité du revenu et aux autres types de renseignements que les gens ont du mal à fournir correctement (par exemple la répartition mensuelle du revenu touché au cours de l'année) et 3) le manque d'informations nécessaires à l'analyse de la participation et de l'admissibilité aux programmes. Les tests pilotes effectués dans le cadre de l'ISDP se distinguent des autres enquêtes, particulièrement la CPS, à plusieurs égards, notamment: 1) les interviews ont été menées à intervalles réguliers dans le courant de l'année; 2) la plupart des catégories de revenu ont été déclarées sur une base mensuelle; 3) le revenu a été déclaré sur une base individuelle; 4) on a suivi les individus tout au long de la période d'enquête afin d'obtenir des données sur l'évolution des revenus et de la composition des familles; 5) on a recueilli des données sur des sujets particuliers comme l'incapacité, la garde des enfants, la fécondité, la valeur nette et les impôts payés afin d'en savoir davantage sur le contexte dans lequel les prestations sont versés ainsi que sur les personnes qui dépendent des programmes pour vivre et le bien-être économique de la population en général. Comme l'ISDP a donné naissance à la SIPP, on retrouve dans cette dernière de nombreuses caractéristiques du premier, dont le plan de sondage ainsi que le contenu et la présentation du questionnaire.

La SIPP a été lancée en octobre 1983 à titre d'enquête permanente comportant un panel de 21,000 ménages choisis pour représenter la population hors établissement institutionnel des États-Unis. Chaque ménage est interviewé une fois tous les quatre mois pendant environ deux ans et demi, la période de référence des principales questions de l'enquête étant les quatre mois qui précèdent l'interview. Chaque ménage est donc interviewé huit fois en tout. Un nouveau panel est introduit chaque année. Ce plan de sondage permet de produire des estimations transversales à partir des réponses de deux panels. Pour plus de renseignements à propos du plan de sondage, des opérations et du contenu relatifs à la SIPP, voir Nelson, McMillen et Kasprzyk (1985).

Le présent article traite de certains aspects du programme (aspects méthodologiques, plan de sondage, aspects statistiques) qui sont des sujets de préoccupation. Nous parlerons 1) de la conception du questionnaire; 2) de la collecte des données, notamment des règles de conduite adoptées à l'égard des répondants, du mode de collecte des données, de la durée de la période de référence et des règles de conduite adoptées à l'égard des personnes ayant déménagé; 3) des définitions, du plan de sondage et des estimations; 4) de l'erreur de réponse.

2. CONCEPTION DU QUESTIONNAIRE

L'ISDP visait principalement à résoudre des problèmes reliés à la sous-déclaration et à la mauvaise classification du revenu dans le supplément de mars de la CPS. Au cours d'un test pilote de l'ISDP, on a eu recours à deux versions du questionnaire que nous désignerons, pour simplifier, par les expressions "questionnaire abrégé" et "questionnaire long".

Avec le questionnaire abrégé, on cherchait à recueillir directement des données sur le revenu tout en limitant le fardeau de réponse. Il s'agissait d'interroger chaque membre du ménage sur certaines catégories de revenu qu'il aurait pu percevoir. Si, pour une catégorie donnée, il avait reçu de l'argent, le montant touché au cours de la période de référence était déterminé avant de passer à la catégorie suivante.

Par contre, avec le questionnaire long, la stratégie adoptée consistait à cerner les événements, les expériences et autres détails associés à certaines catégories de revenu perçu. Le formulaire contenait un ensemble assez complet de questions d'approfondissement concernant le revenu touché ainsi que de longues questions visant à déterminer les montants en cause. On ne

cherchait pas à obtenir le montant reçu relativement à une catégorie de revenu tant que toutes les catégories pour lesquelles un revenu avait été touché n'avaient pas été déterminées.

L'hypothèse que l'on cherchait à vérifier était que les renseignements sur le revenu fournis au moyen du questionnaire long seraient plus complets et plus précis; Olson (1980) résume l'analyse qui a été faite sur les deux versions du questionnaire. Cette analyse a pris plusieurs formes, dont il est question dans le résumé:1) observation des intervieweurs pendant la formation et les interviews; 2) compte rendu des intervieweurs et des observateurs; 3) examen de chaque questionnaire complété; 4) analyse des taux de participation à l'enquête et des taux de réponse aux questions; et enfin 5) analyse de la qualité des données et des questionnaires rejetés à la vérification, particulièrement ceux dont le rejet était dû au fait que l'interviewer n'avait pas suivi les instructions indiquant de "passer à". C'est une version modifiée du questionnaire long qui a été retenue pour la suite des essais et finalement pour la SIPP. Les intervieweurs et les répondants ont en effet trouvé que le questionnaire long était plus facile à remplir et, grâce à lui, on a obtenu des taux de déclaration du revenu plus élevés.

Toujours dans le cadre de l'ISDP, on a fait une expérience mettant en jeu différents types de questionnaires. Cette expérience opposait deux approches:l'une axée sur le ménage et l'autre axée sur l'individu, cette dernière faisant appel à un questionnaire dont la forme et le contenu étaient le fruit de tests pilotes réalisés antérieurement. Dans l'approche axée sur le ménage, on s'est servi d'une version remaniée d'un questionnaire utilisé pour l'essai du supplément sur le revenu de la CPS réalisé en avril 1978. On cherchait, au moyen de ce questionnaire, à réduire le fardeau de réponse en demandant à un seul membre du ménage s'il y avait dans le ménage quelqu'un qui avait touché une certaine catégorie de revenu au cours de la période de référence. Chaque réponse affirmative était suivie d'une question pour déterminer quel membre du ménage avait touché le revenu en question. Ce n'est qu'une fois qu'on avait obtenu des renseignements au sujet de toutes les catégories de revenu perçu par l'ensemble des membres du ménage qu'on cherchait à connaître le montant reçu dans chaque cas. On pensait que cette approche aurait pour effet de réduire la durée de l'interview sans faire baisser la qualité des données.

Dans l'approche axée sur l'individu, on posait des questions sur toutes les sources de revenu à un premier membre du ménage, puis à un second, et ainsi de suite. On remplissait un questionnaire distinct pour chaque adulte faisant partie du ménage échantillonné, mais on avait recours autant que possible à des instructions indiquant de "passer à" et à des cases qu'il suffisait de cocher afin de diminuer le nombre de questions posées à chaque répondant.

Du point de vue de la qualité des données et de la durée de l'interview, les différences observées entre les deux approches se sont avérées minimes. Elles n'étaient pas grandes non plus en ce qui a trait aux estimations du pourcentage de personnes ayant touché diverses catégories de revenu ni du nombre de cas de non-réponse dus à l'impossibilité ou au refus de fournir les renseignements demandés. On s'attendait à ce que la durée de l'interview soit beaucoup moins grande lorsque l'approche axée sur le ménage était utilisée plutôt que l'approche axée sur l'individu, or elle n'a été inférieure que d'environ cinq minutes par ménage et d'environ trois minutes par personne. Comme l'approche axée sur le ménage n'a pas donné de résultats sensiblement meilleurs que l'approche axée sur l'individu, c'est cette dernière qui, légèrement améliorée et raffinée, a été retenue pour la SIPP.

Le type de questionnaire à utiliser et la marche à suivre pour la collecte des données font l'objet de discussions qui se poursuivent dans le cadre du programme de la SIPP. La principale question consiste à savoir si l'on obtient de meilleures estimations au moyen d'interviews ne faisant pas appel aux réponses données antérieurement (interviews dites indépendantes) ou au moyen d'interviews au cours desquelles on rappelle aux répondants les renseignements qu'ils ont déjà fournis (interviews dites avec rétro-information). Dans la SIPP, on a recours à l'approche avec rétroinformation pour mettre à jour les données à chaque interview et ainsi connaître les dernières tendances relatives au revenu, mais cette approche n'a pas été évaluée.

Il est possible d'adopter l'approche avec rétro-information pour la collecte des données sur la valeur nette personnelle. Dans la SIPP, ces données sont des données ponctuelles recueillies

à un an d'intervalle. Plus particulièrement, les données sur l'actif et le passif, recueillies au cours du cycle 4 auprès du panel de 1984, ont été fournies à la moitié des répondants lors de l'interview du cycle 7. Ces données n'ont été communiquées qu'à la moitié des répondants afin de pouvoir comparer l'approche avec rétro-information à l'approche indépendante.

L'argument invoqué pour justifier l'approche avec rétro-information est que les répondants donnent de meilleures estimations de l'évolution de leur revenu si on leurs rappelle auparavant le montant qu'ils avaient déclarés l'année précédente. Si les répondants savent de combien leur actif a changé et si on leur rappelle le montant de départ, on peut supposer que le montant déclaré pour l'année en cours reflète le changement réel pendant la période de référence. Lamas et McNeil (1987) analysent ces données, mais ils n'affirment rien de définitif sur l'incidence de l'approche avec rétro-information car ils ne disposent pas de données de référence. Ils disent bien, cependant, que l'interview avec rétro-information ne change pas les estimations transversales et que les différences observées entre les sousgroupes pour ce qui est de la valeur nette sont celles auxquelles on s'attendait. L'étude de l'évolution de la valeur nette au microniveau dans les ménages ayant fourni toutes les données sur leurs ressources tend à montrer que l'interview avec rétro-information fait baisser les estimations des changements de la valeur nette.

La question des résultats obtenus selon qu'on procédait à une interview indépendante ou à une interview avec rétro-information s'est également posée relativement à la mesure de la branche d'activité et de la profession. Auprès des panels de 1984 et de 1985, ces données ont été recueillies indépendamment à chaque interview même si le répondant n'avait pas changé d'employeur. Cette façon de procéder tient compte du fait que les fonctions d'un travailleur peuvent varier à l'occasion et permet de noter cette évolution. En outre, les fonctions peuvent changer au point que le poste occupé par le travailleur en question n'est plus classé dans la même catégorie d'une interview à l'autre même si celui-ci est au service du même employeur.

La collecte indépendante des données sur la branche d'activité et la profession pose cependant des problèmes. Une légère variation dans la manière dont le répondant décrit ses fonctions ou dans l'interprétation de cette description par les personnes chargées d'attribuer les codes de classification peut entraîner un changement de catégorie professionnelle injustifié.

L'étude de ce problème a permit de produire des estimations du nombre de fois que la profession et la branche d'activité des répondants qui étaient restés au service du *même* employeur avaient été classifiées dans une catégorie différente d'une interview à l'autre . Chez les personnes faisant partie du panel de 1984 qui avaient déclaré travailler pour le même employeur au cours des douze premiers mois de l'enquête, on a observé un changement du code à trois chiffres d'une interview à l'autre dans environ 40% des cas pour ce qui est de la profession et dans 20% des cas pour ce qui est de la branche d'activité (Kalton, McMillen et Kasprzyk 1986).

Pour réduire le nombre de changements de codes de profession et de branche d'activité découlant d'une erreur de réponse aléatoire ou d'une mauvaise interprétation des réponses par les codeurs et pour abréger les interviews, on a décidé de modifier la SIPP à partir du panel de 1986. En l'occurrence, on a introduit une question de sélection par laquelle on demande aux répondants si leurs activités ou fonctions ont changé au cours des huit derniers mois. Si la réponse est négative, les questions détaillées sur la profession et la branche d'activité ne leur sont pas posées. Les codes attribuables à ces deux variables sont alors assignés à partir des réponses fournies à l'interview précédente.

Il importe de signaler le fait suivant:bien que ce changement ait été introduit au moment où le panel de 1986 a été créé, les données sur la profession et la branche d'activité se rapportant au panel de 1985 qui ont été recueillies au cours de la même période que celles se rapportant au panel de 1986 l'ont été selon l'interview indépendante. Ces deux façons de procéder utilisées simultanément constituent donc une expérience en quelque sorte spontanée dont les résultats n'ont pas encore été analysés.

3. COLLECTE DES DONNÉES

Quatre aspects de la collecte des données pour la SIPP seront traités ici: 1) les règles de conduite adoptées à l'égard des répondants; 2) le mode de collecte des données; 3) la durée de la période de référence; 4) les règles de conduite adoptées à l'égard des répondants ayant déménagé.

Règles de conduite adoptées à l'égard des répondants

Lorsqu'on a à mener des interviews dans des ménages composés de plusieurs membres, il faut décider si les réponses par personne interposée sont acceptables. Comme il peut arriver que les membres du ménage ne soient pas tous présents au moment de l'interview, on peut gagner du temps et de l'argent en posant aux personnes qui sont là les questions concernant celles qui n'y sont pas. Toutefois, pour certaines questions, les données fournies par un répondant substitut risquent d'être moins précises que celles qui auraient été communiquées par le répondant lui-même. Kalton, Kasprzyk et McMillen (1988) traitent de cette question dans le contexte des enquêtes par panel.

Une expérience portant sur les règles de conduites adoptées à l'égard des répondants a été menée dans le cadre de l'ISDP. On a comparé la qualité des données obtenues dans un groupe témoin où une déclaration par personne interposée a été acceptée si un des membres du ménage se sentait en mesure de fournir les renseignements concernant la personne absente, aux données obtenues dans un autre groupe où les déclarations par personne interposée n'ont été acceptées que dans des situations exceptionnelles (par exemple si le répondant était physiquement ou mentalement incapable de répondre aux questions, s'il ne parlait pas l'anglais ou s'il allait être absent pendant toute la durée de l'enquête). Environ 85% des adultes interrogés dans les ménages où l'on a exigé l'autodéclaration parlaient en leur propre nom comparativement à 65% dans les ménages où l'on a accepté des déclarations par personne interposée. On a donc obtenu environ 20% de déclarations en nom propre de plus dans le groupe où ce genre de déclaration a été exigé que dans l'autre groupe (Coder 1980).

Les taux de refus étaient légèrement plus élevés dans le groupe où l'on a exigé l'autodéclaration tandis que le pourcentage de ménages interviewés était un peu plus fort dans le groupe où l'on a accepté les déclarations par personne interposée. L'écart était cependant trop petit dans un cas comme dans l'autre pour qu'on puisse en déduire laquelle des deux règles de conduite était la plus avantageuse. La proportion de personnes non-interviewées dans les ménages où au moins un adulte avait répondu aux questions était plus élevée dans le groupe où l'on a exigé l'autodéclaration. Pour ce qui est des pourcentages de personnes ayant touché diverses catégories de revenu, la différence observée entre les deux groupes était légère également, et les taux de non-réponse due à l'impossibilité ou au refus de communiquer les renseignements demandés aux questions sur les montants touchés dans chaque catégorie de revenu n'étaient pas systématiquement plus faibles dans le groupe où l'autodéclaration a été exigée.

Dans le groupe où l'on a exigé l'autodéclaration, les personnes interrogées ont plus souvent consulté leurs dossiers avant de répondre aux questions concernant la rémunération, et les taux de réponse à la question sur la rémunération horaire étaient plus élevés, mais en général l'écart entre les résultats obtenus dans les deux groupes n'étaient pas concluants. C'est pour cette raison et parce qu'on a estimé qu'il coûtait de 4 à 6% plus cher d'exiger l'autodéclaration de même que parce qu'il a fallu adopter une procédure de rappel pour obtenir des renseignements essentiels qui n'étaient pas connus au moment de l'interview que les déclarations par personne interposée sont maintenant acceptées dans la SIPP.

Il a également fallu se demander quelle règle de conduite adopter à l'égard des étudiants universitaires. En général, on considère qu'un étudiant fait partie du ménage de ses parents tant qu'il ne s'est pas établi ailleurs de façon permanente. On traite donc habituellement les étudiants qui n'habitent pas chez leurs parents pendant la durée de leurs études comme des membres du ménage temporairement absents, et on demande à un autre membre du ménage de leurs parents de répondre aux questions à leur place. Lors d'un test pilote effectué dans le

cadre de l'ISDP, on a tenté de déterminer si les renseignements relatifs à des étudiants absents de chez leurs parents obtenus auprès de répondants substituts étaient exacts. Dans un premier temps, on a mené l'interview par personne interposée auprès d'un membre du ménage des parents et, dans un deuxième temps, on a mené l'interview auprès de l'étudiant lui-même à l'endroit où il demeurait pendant la période de ses études. Les résultats de cette étude sont décrits par Roman et O'Brien (1984). L'analyse est limitée à cause d'erreurs commises sur le plan de l'organisation et de la réalisation du test. Les auteurs signalent cependant que les répondants substituts ne sont souvent pas en mesure de dire si un étudiant touche une certaine catégorie de revenu et que, même s'ils sont en mesure de le faire, lorsqu'on leur demande plus de détails, ils sont portés à répondre qu'ils ne les connaissent pas. Les auteurs notent également que plus le montant du revenu ou de la dépense est élevé, plus la réponse fournie par un répondant substitut est bonne.

Mode de collecte des données

La plupart des interviews (environ 95%) menées pour la SIPP sont des interviews sur place (Kalton, McMillen et Kaprzyk 1986). Comme ce genre d'interview coûte de plus en plus cher, le Census Bureau étudie actuellement la possibilité d'augmenter substantiellement le nombre d'interviews téléphoniques réalisées pour cette enquête.

On a donc effectué au mois de juin 1985 un essai préliminaire visant à déterminer la faisabilité d'interviews téléphoniques "préparées", c'est-à-dire d'interviews téléphoniques menées auprès de ménages qui avaient accordé des interviews sur place lors d'un cycle antérieur. Cet essai préliminaire a été réalisé par deux bureaux régionaux du Census Bureau auprès d'un échantillon de 280 ménages. Le taux de refus (environ 2.5%) et le taux de non-contact (environ 11%) étaient conformes aux attentes du personnel. En ce qui concerne les taux de non-réponse aux diverses questions, ils n'étaient pas plus élevés que prévu non plus (U.S. Bureau of the Census, 1986).

Étant donné les résultats obtenus, on a procédé à un essai national de l'interview téléphonique pour la SIPP (SIPP National Telephone Test) qui s'est déroulé du mois d'août au mois de novembre 1986 et du mois de février au mois d'avril 1987. Il s'agissait d'évaluer l'utilisation sur une grande échelle des interviews téléphoniques "préparées" et de déterminer si les gens sont prêts à fournir des données par téléphone au cours de deux interviews consécutives. On a décidé que les ménages situés dans 50% des segments feraient l'objet d'un nombre maximum d'interviews téléphoniques et les autres, d'un nombre maximum d'interviews sur place. Les interviewers ont effectué presque toutes les interviews téléphoniques à domicile. Gbur et Durant (1987) donnent les résultats préliminaires de la première étapes de l'expérience. Selon eux, l'interview téléphonique ne semble pas avoir eu de conséquences graves sur les taux de réponse des ménages, et les taux de réponse des personnes étaient semblables quel que soit le mode utilisé. Quant aux taux de non-réponse aux questions, ils n'ont été que légèrement influencés par le recours à l'interview téléphonique. Le compte rendu des autres résultats suivra.

Durée de la période de référence

L'ISDP ayant porté principalement sur les techniques de collecte des données susceptibles d'améliorer la déclaration des revenus monétaires et non monétaires, la durée de la période de référence pour la plupart des questions posées au cours de l'enquête est un sujet auquel on a accordé beaucoup d'importance lors de la détermination du plan de sondage.

On a abordé ce sujet de deux points de vue pendant le programme. Premièrement, on a comparé les résultats obtenus au moyen d'une interview dont la période de référence était de six mois et ceux obtenus au moyen de deux interviews consécutives dont la période de référence était de trois mois chacune. Deuxièmement, on a comparé les résultats obtenus si l'on demandait aux répondants de déclarer les revenus de biens qu'ils avaient touchés au cours des trois mois antérieurs et ceux obtenus si on leur demandait de déclarer les revenus touchés au cours des six mois antérieurs.

Olson (1980) décrit certains résultats de la première expérience. Fait non surprenant, lorsque la période visée par les questions est de six mois, la proportion du revenu déclarée pour le début de la période est inférieure à ce qu'elle devrait être, et ce pour plusieurs sources de revenu comme la rémunération, l'aide aux familles à faible revenu avec enfants à charge et les prestations d'assurance-chômage. Cette observation, bien que non définitive, confirme l'hypothèse selon laquelle les chances d'omission par défaut de mémoire augmentent avec la durée de la période de référence. Il est ressorti d'une autre analyse que le nombre de sources de revenu déclarées par ménage pour les trois premiers mois de la période de référence de six mois était inférieur au nombre obtenu pour les mêmes mois lorsque la période de référence était de trois mois. Les résultats de la deuxième expérience n'ont pas été analysés à cause du retrait des subventions accordées au programme de mise au point de l'enquête.

Les résultats de la première expérience et certains faits observés à d'autres étapes de l'ISDP ont amené les planificateurs de la SIPP à adopter une période de référence de quatre mois pour cette enquête; cette décision permet de maintenir les coûts dans les limites permises par le budget et la qualité des données à un niveau satisfaisant.

Règle de conduite adoptée à l'égard des répondants ayant déménagé

Un aspect important du plan de sondage utilisé dans l'ISDP et à présent dans la SIPP est le fait que toutes les personnes faisant partie d'un ménage de l'échantillon au moment de la première interview continuent à faire partie de l'échantillon pendant les deux années et demie que le panel est conservé, même si une personne ou plus déménage. Pour des raisons d'ordre financier et opérationnel, il a été décidé que les interviews sur place seraient menées à la nouvelle adresse uniquement si certaines contraintes géographiques étaient respectées; dans l'ISDP, il fallait que la nouvelle adresse se situe dans un rayon de 50 miles d'une unité primaire d'échantillonnage et, dans la SIPP, dans un rayon de 100 miles.

Lorsqu'on crée un panel, on tire un échantillon d'adresses, et ce n'est qu'à la première interview que les personnes habitant à chaque adresse sont identifiées. Après cela, l'échantillon n'est plus un échantillon d'adresses mais un échantillon de personnes, et il est constitué de tous les individus énumérés lors de la première interview. On demande à ces derniers et à quiconque partage le même logement de se prêter aux interviews ultérieures.

Pendant la durée de l'ISDP, deux questions relatives aux répondants ayant déménagé ont été considérées importantes: 1) la production d'estimations transversales à chaque interview et 2) les coûts supportés lorsqu'on suit ces répondants. Huang (1984) propose plusieurs poids de base non biaisés pour les estimations transversales relatives à la population hors établissement institutionnel à utiliser si l'échantillon contient des répondants ayant déménagé. Il établit le lien entre les observations effectuées à un moment donné et les probabilités de sélection des ménages de l'échantillon original. Deux approches sont décrites dans son étude:la première, fondée sur la multiplicité, est liée au nombre de façons dont un nouveau ménage peut se retrouver dans l'échantillon et la seconde, fondée sur la notion de "part égale", suppose que tous les membres d'un ménage ont le même poids. Pour la SIPP, comme pour l'ISDP, on a adopté l'approche fondée sur la notion de "part égale".

La question des coûts a été traitée dans une étude consacrée à ce sujet ("Mover's Cost Study"). Cette étude avait pour but d'évaluer les coûts, au chapitre de la collecte des données, reliés au fait de suivre les répondants après qu'ils ont déménagé. White et Huang (1982) font un compte rendu de l'étude et présentent certains résultats fondés sur la ligne de conduite adoptée, au cours du test pilote, à l'égard de ces répondants. Ils ont trouvé que lorsqu'on a suivi ces derniers pendant un an, le nombre de ménages devant être interviewés a augmenté de 8.8%; en outre, au bout de quinze mois, les répondants ayant déménagé représentaient 22% de l'échantillon total, le nombre d'heures d'interviews s'était accru de 7% et le nombre de miles déclarés par les intervieweurs, de 11.4%.

Jean et McArthur (1984) se sont penchés sur la question de la collecte des données relatives aux répondants ayant déménagé dans le contexte de la SIPP et ils recommandent certains

changements susceptibles d'améliorer le champ d'observation de la SIPP à l'avenir. Kalton et Lepkowski (1985) traitent également de la règle de conduite adoptée à l'égard des répondants de la SIPP ayant déménagé et ils proposent un programme de recherche visant à mesurer le sous-recouvrement pour certaines sources et à déterminer quels sous-groupes sont les plus affectés. Plus récemment, Jean et McArthur (1987), dans une étude des données obtenues pendant cinq cycles d'interviews, ont constaté que parmi les personnes ayant déménagé après la première interview, c'est-à-dire entre le cycle 2 et le cycle 5, 69% se sont prêtées à chacune des cinq interviews, 23% n'ont pas passé la cinquième et 9% ont passé la cinquième mais en ont manqué au moins une auparavant.

4. DÉFINITIONS, PLAN DE SONDAGE ET ESTIMATION

Dans le cadre de l'ISDP et encore dans celui du programme de la SIPP, il s'est fait des travaux de recherche importants dans le domaine de la définition d'unités annuelles d'analyse à l'aide de données infra-annuelles et de l'évaluation de ces définitions d'un point de vue statistique. Le traitement de la non-réponse dans les enquêtes par panel a également suscité énormément d'intérêt parmi les chercheurs. Enfin, on a étudié et on étudie toujours les techniques d'estimation susceptibles de réduire l'erreur d'échantillonnage de même que les méthodes d'échantillonnage des sous-groupes les mieux adaptées à l'enquête.

Définitions qui tiennent compte du facteur temps

Les statistiques annuelles sur les familles et les ménages sont des indicateurs importants du bien-être économique de la nation. La SIPP permet de recueillir des données infra-annuelles, en fait des données mensuelles, qui témoignent de l'évolution de la composition des ménages. Grâce à ces données, il est possible de produire des statistiques annuelles sur les ménages qui rendent bien compte des changements que subissent ces derniers dans le courant d'une année, contrairement aux statistiques actuelles qui ne font tout simplement pas cas de cette réalité. La construction d'unités annuelles d'analyse, qu'ils s'agisse de ménages, de familles ou d'unités de programme, pose le problème des poids longitudinaux et des techniques d'imputation à adopter. Le principal problème, cependant, est conceptuel. Étant donné que la composition des ménages varie entre le début et la fin d'une année, quand est-il indiqué pour les mesures annuelles de rendre compte de cette évolution? Autrement dit, comment faudrait-il définir les ménages et les familles de manière à pouvoir tenir compte des données recueillies à deux moments ou plus et à ne pas être en contradiction avec les estimations transversales habituellement produites sur les ménages et les familles?

Les analystes du Census Bureau se sont longuement penchés sur la question d'une définition des ménages et des familles qui tiendrait compte du facteur temps (McMillen et Herriot 1985; Citro 1985). Citro, Hernandez et Herriot (1986) ainsi que Citro, Hernandez et Moorman (1986) ont fait état des travaux de recherche empirique dans lesquels plusieurs définitions de ménages de ce genre et diverses mesures du revenu annuel et des types de famille ont été examinées. Dans ces travaux, quatre définitions possibles ont été mises de l'avant:on considère qu'un ménage est le même au bout d'une certaine période 1) si la personne de référence est la même; 2) si le membre principal du ménage est le même (la différence entre cette définition et la précédente se manifeste dans le traitement des ménages comportant un couple marié, où la personne de référence peut être soit la femme soit le mari mais où le membre principal est toujours la femme); 3) si la personne de référence est la même et si la famille qui le constitue appartient au même type; 4) si la personne de référence est la même, la famille qui le constitue appartient au même type et le nombre de membres n'a pas changé.

Il ressort à première vue de ces travaux que le choix d'une définition n'influe pas sensiblement sur les statistiques annuelles relatives aux ménages à faible revenu ni sur celles relatives

aux ménages selon les types auxquels ceux-ci appartiennent. Si cette constatation est confirmé par ailleurs, ce seront d'autres critères comme la facilité d'application et la simplicité des opérations qui détermineront le choix d'une définition du ménage qui tienne compte du facteur temps.

Estimation statistique d'après des définitions tenant compte du facteur temps

Les travaux de recherche sur la production d'estimations en fonction de définitions qui tiennent compte du facteur temps se sont faits dans deux directions:estimations longitudinales concernant les personnes et estimations longitudinales concernant les ménages (unité familiale ou unité de programme). Les travaux sur les estimations concernant les personnes comprennent le calcul des probabilités de sélection permettant d'obtenir des estimations longitudinales non biaisées des caractéristiques individuelles et l'utilisation de totaux de contrôle à d'autres étapes de l'estimation (Judkins et coll. 1984). Kobilarcik et Singh (1986) ont raffiné ces opérations et décrit une méthode qu'ils proposent et qui permettrait de produire des poids longitudinaux pour l'analyse des données sur les personnes obtenues au cours des trois premières interviews de la SIPP.

Kobilarcik et Singh définissent l'univers longitudinal comme étant constitué de la population hors établissement institutionnel (y compris les casernes) au 1er décembre 1983, soit au milieu de la période d'interview du cycle 1. L'échantillon tiré de l'univers longitudinal compte, quant à lui, les personnes admissibles qui habitaient dans le logement sélectionné au moment de la première interview. Pour les besoins de cette méthode d'estimation, les personnes "interviewées" sont celles qui s'étaient prêtées aux trois premières interviews de la SIPP et qui, au moment de la première interview, faisaient partie d'un ménage dont tous les membres admissibles avaient fait l'objet d'une interview, de même que celles qui faisaient partie d'un ménage interviewé pendant le cycle 1 mais qui étaient décédées ou avaient déménagé à l'extérieur des limites géographiques de l'enquête dans le courant du deuxième ou du troisième cycle.

Pour la méthode d'estimation, les personnes non interviewées sont donc celles qui, au moment de la première interview, faisaient partie d'un ménage dont au moins un des membres ne s'était pas prêté à l'interview et celles qui faisaient partie d'un ménage interviewé pendant le cycle 1 mais qui ne s'étaient pas prêtées soit à la deuxième soit à la troisième interview, soit aux deux. On a attribué un poids positif aux personnes classées dans la catégorie des personnes interviewées. Pour cet univers, les poids ont été obtenus de la manière habituelle:en prenant l'inverse de la probabilité de sélection, en calculant un facteur de correction pour les non-interviews et en ajustant les chiffres au moyen des totaux de contrôle de la population. L'ajustement pour la non-réponse s'est faite en deux étapes, soit en ajustant d'abord pour la non-réponse des ménages et ensuite pour celle des personnes, la seconde étape faisant appel aux renseignements recueillis à la première interview.

La question des estimations longitudinales concernant les ménages (unité familiale ou unité de programme) est également à l'étude. Plusieurs façons d'aborder cette question ont été décrites par Ernst, Hubble et Judkins (1984) et plus récemment par Ernst (1988). Ce dernier démontre pourquoi la pondération au moyen de l'inverse de la probabilité de sélection ne produit pas, en général, de bonnes estimations longitudinales relatives aux ménages et aux familles et il propose une classe de méthodes qui le font. Il décrit en outre les difficultés qu'on peut rencontrer si, au moment d'appliquer ces méthodes, on n'a pas accès aux données nécessaires pour établir les poids. Il définit alors les conditions que les définitions doivent satisfaire pour permettre de contourner ces problèmes. Enfin, il parle de méthodes visant à corriger les estimations longitudinales pour tenir compenser la non-réponse ainsi qu'à ajuster les variables démographiques aux estimations indépendantes.

Non-réponse et imputation

Dans le cas d'enquêtes longitudinales comme la SIPP et celles qui ont été menées au cours de l'ISDP, les problèmes de refus et de non-réponse sélective s'ajoutent au problème cumulatif

de la non-réponse à l'un ou l'autre des cycles du panel. Un aspect important des travaux réalisés jusqu'à présent a donc été l'étude de méthodes de compensation de la non-réponse. Kalton (1983) passe en revue les méthodes actuellement utilisées dans la recherche sur les techniques d'enquête, tandis que Kalton et Kasprzyk (1982, 1986) se penchent sur les méthodes d'imputation et résument les caractéristiques du biais et de la variance dans plusieurs classes de méthodes.

Les données tirées de la SIPP peuvent être traitées soit comme des données longitudinales, soit comme des données transversales. Chapman, Bailey et Kaprzyk (1986) décrivent les méthodes utilisées pour compenser la non-réponse des unités pour la SIPP ainsi que pour d'autres enquêtes du Census Bureau. Ils parlent des problèmes que cela pose lorqu'on a affaire à une enquête à plusieurs interviews. Dans une enquête par panel, cependant, la non-réponse peut prendre deux formes:non-réponse aux questions, qui se produit lorsqu'une unité participe à l'enquête mais ne répond pas à toutes les questions, et non-réponse aux cycles, qui se produit lorsqu'une unité participent à certaines interviews, mais pas à toutes.

Heeringa et Lepkowski (1986) décrivent des classes générales de méthodes d'imputation longitudinale qui pourraient être utilisées à la place de la méthode du "hot deck", qui est une méthode d'imputation transversale. Ils comparent les résultats obtenus au moyen d'une méthode simple d'imputation longitudinale, la substitution directe longitudinale (où l'on met à la place des réponses manquantes à une interview les réponses fournies aux mêmes questions à une autre interview) aux résultats obtenus par la méthode du "hot deck". Ils montrent, et cela ne surprend personne, que la méthode de substitution directe entraîne une sous-estimation des changements. Ils concluent cependant que cette sous-estimation est peut-être préférable à la surestimation marquée qui découle de l'utilisation de la méthode du "hot deck".

Les enquêtes par panel soulèvent un problème de données manquantes que les autres types d'enquêtes ne posent pas, et c'est le problème de la non-réponse aux cycles. En général, la non-réponse aux cycles donne lieu à un plus grand nombre de données manquantes pour un individu en particulier que la non-réponse aux questions. Par ailleurs, les données obtenues auprès de répondants qui se prêtent à tous les cycles d'interviews fournissent plus de détails sur l'unité non répondante que les données obtenues auprès de répondants ayant sauté certains cycles. Pour compenser la non-réponse, il faut donc avoir recours à la pondération ou à l'imputation ou encore à une combinaison des deux. Kalton, Lepkowski et Lin (1985) parlent de ce problèmes et des résultats obtenus à la suite d'essais empiriques effectués dans le contexte de l'ISDP. Il ressort clairement de leur étude que le choix à faire entre la pondération et l'imputation pour régler ce genre de problème de données manquantes est loin d'être évident. Kalton (1986) de même que Kalton et Miller (1986) poussent plus loin l'analyse du problème et en viennent à la conclusion que l'imputation fausse certaines estimations et que la pondération constitue éventuellement une meilleure solution pour de grandes sous-catégories si une diminution de la taille réelle de l'échantillon ne porte pas à conséquence. Ils ajoutent toutefois que l'imputation peut être préférable lorqu'il s'agit d'estimations produites pour des petites sous-catégories et que la diminution de la taille d'échantillon est considérable. Lorsqu'on travaille avec un fichier longitudinal de la SIPP contenant les données tirées de trois interviews, la différence dans la taille de l'échantillon selon qu'on procède à la pondération ou à l'imputation n'étant pas considérable, la pondération constitue la solution générale la plus sage. Enfin, Lepkowski (1988) conclut après d'autres travaux de recherche empiriques que pour régler le problème de la non-réponse aux cycles il faut absolument tenir compte de facteurs comme les principaux buts visés par le plan de sondage, le plan d'échantillonnage du panel et la répartition de la non-réponse aux cycles. Il propose des critères à retenir au moment de mettre au point une méthode de traitement de la nonréponse et termine en disant que la pondération semble être la meilleure méthode de compensation pour la non-réponse aux cycles.

Réduction de l'erreur d'échantillonnage au moyen de techniques d'estimation

Deux méthodes visant à réduire l'erreur d'échantillonnage au moyen de techniques d'estimation sont à l'étude:l'utilisation d'un estimateur composite et celle de dossiers administratifs.

L'estimation composite est une technique qui combine les estimations produites pour la période en cours et celles produites pour les périodes antérieures afin d'améliorer les estimations en mettant à profit les corrélations qui existent entre les réponses fournies par les mêmes unités d'analyse à deux moments différents. Ce genre d'estimation est particulièrement efficace lorsque la corrélation est forte, ce qui est le cas pour un bon nombre d'éléments d'information importants recueillis au moyen de la SIPP. Chakrabarty (1986) a fait une première étude des types d'estimateurs composites susceptibles de convenir à la structure des données de la SIPP. Toutefois, le contenu de cette enquête n'a pas été suffisamment stable pendant les premières années de son existence pour qu'on puisse sérieusement penser à adopter un estimateur composite.

Une autre solution envisagée en vue de réduire la variance est l'utilisation de données de sources administratives pour une stratification a posteriori. Les méthodes d'estimation transversale utilisées présentement reposent sur un rajustement du second degré visant à accroître la précision des estimations. A cette fin, on procède à un ajustement par quotient des estimations pour les mois de collecte et les mois de référence en fonction des estimations de population. Cependant, le Census Bureau a accès à certains fichiers de l'Internal Revenue Service et de la Social Security Administration qui peuvent être utilisés pour la production de distributions détaillées du revenu brut ajusté selon l'âge, la race et le sexe. Nous venons tout juste de commencer à étudier de quelle façon ces données administratives pourraient servir à une stratification a posteriori visant à améliorer les estimations du revenu moyen et médian des particuliers et des ménages ainsi que les estimations des déciles de la distribution du revenu des particuliers et des ménages. La question primordiale actuellement à l'étude est de savoir de combien on peut réduire de la variance des estimations avec une telle méthode. Fay et Huggins (1988) en donneront quelques indications.

Échantillonnage de sous-populations spéciales

On dit souvent que certains sous-groupes de la population sont plus touchés que les autres par la politique sociale adoptée par le gouvernement. On pense notamment aux pauvres, aux personnes âgées, aux noirs, aux citoyens d'origine latino-américaine et aux bénéficiaires des programmes fédéraux de sécurité du revenu. Dès le départ, dans l'ISDP, on voulait choisir un plan de sondage qui permettrait de rendre les estimations relatives aux sous-populations plus fiables. C'est pourquoi on a mis l'accent sur le prélèvement des échantillons dans les fichiers administratifs. On a donc souvent tiré ces derniers des listes de bénéficiaires de programmes administrés par le gouvernement fédéral ou par les États (Kasprzyk 1983; Bowie et Kasprzyk 1987).

Un groupe de travail du Census Bureau a analysé des propositions de souséchantillonnage visant à surreprésenter certaines populations spéciales. Il s'agissait de comparer la fiabilité des estimations produites lorsque divers plans de sous-échantillonnage étaient utilisés. Le groupe a mis en évidence certaines caractéristiques de sous-échantillonnage en fonction de variables démographiques et de variables relatives au revenu et il a évalué la fiabilité selon différents taux et caractéristiques de sous-échantillonnage (U.S. Bureau of the Census 1985).

Le groupe de travail en est venu à la conclusion que, pour une enquête générale sur le revenu comme la SIPP, les propositions de sous-échantillonnage ne permettaient d'améliorer que légèrement les estimations relatives aux personnes à faible revenu et que ce faible avantage ne l'emportait pas sur les inconvénients, notamment une plus grande complexité des opérations, le fait que l'échantillon ne soit plus autopondéré et une baisse de précision en ce qui a trait aux personnes à revenu moyen.

5. ERREUR DE RÉPONSE

L'erreur de réponse constitue un aspect d'un problème plus général, l'erreur non due à l'échantillonnage dont parlent Kalton, Kasprzyk et McMillen (1988). Les erreurs de réponse sont commises lorsque des données incorrectes sont inscrites dans les questionnaires, et ce pour diverses raisons notamment parce que le questionnaire est mal fait, la mémoire du répondant fait défaut ou le répondant n'est pas le bon. Pour permettre de mieux comprendre l'erreur de réponse en général, nous allons brièvement examiner ce problème dans le contexte des données de la SIPP sur les flux bruts et de l'étude de vérification des enregistrements.

Données de la SIPP sur les flux bruts

L'analyse des données relatives aux programmes effectuée mois par mois dans le cadre de l'ISDP a révélé que le renouvellement des bénéficiaires avait tendance à être déclaré plus souvent entre les cycles d'interview qu'au cours des cycles (Moore et Kasprzyk 1984). L'analyse des données tirées de la SIPP (Burkhead et Coder 1985) qui a porté sur la variation mensuelle du montant des prestations reçues pendant une période de douze mois a pour sa part accordé une attention particulière aux changements survenus entre le dernier mois d'une période de référence et le premier mois de la période de référence suivante.

Les résultats obtenus au moyen des données de l'ISDP et des données de la SIPP sont semblables, en ce sens que les changements observés sont irréguliers et sont clairement liés au plan d'interview. Les changements bruts sont nettement plus prononcés entre le dernier mois d'une période de référence et le premier mois de la période suivante. Hill (1987) s'est servi des données mensuelles recueillies au cours des cycles d'interviews de 1984 et de 1985 de la PSID (Panel Study of Income Dynamics/ étude par panel de la dynamique du revenu) pour mesurer les changements excessifs observés entre les cycles comparativement à ceux observés au cours des cycles et essayer d'en déterminer la cause. Il a trouvé que même si l'ordre des questions et la durée des période de référence étaient différents, les changements entre les cycles étaient plus nombreux que les changements pendant les cycles, comme pour la SIPP. Les causes principales du problème ne sont pas connues, mais la formulation des questions et la présentation du questionnaire de même que les défaillances de mémoire des répondants et l'interaction entre ces deux facteurs semblent être des causes probables.

Weidman (1986) a effectué une analyse empirique dans le but de déterminer s'il existait un lien évident entre certaines caractéristiques des répondants et les changements observés quant à un certain nombre de catégories de revenu perçu. Il n'en a trouvé aucun, sur une période de plusieurs mois consécutifs, entre les répartitions de changements bruts, le fait que les déclarations aient été faites en nom propre ou par personne interposée et neuf variables démographiques (l'âge, la race, le sexe, le niveau d'instruction, l'état matrimonial, la taille du ménage, le mode d'occupation du logement, le lien avec la personne de référence et la taille de la région métropolitaine), mais il a remarqué qu'un plus grand nombre de changements survenaient lorsque certaines donnés étaient imputées. L'absence de lien notable indique qu'il faut essayer de trouver d'autres façons d'expliquer ce phénomène.

Les estimations de flux bruts continuent de susciter beaucoup d'intérêt. Hubble et Judkins (1986) ont mis au point un modèle permettant d'estimer le biais des estimations de flux bruts résultant des erreurs de réponse, modèle dont les paramètres sont estimés au moyen des taux d'erreur de réponse de la SIPP et du rapport des estimations de flux bruts à l'intérieur des cycles et d'un cycle à l'autre. Il est nécessaire de poser plusieurs hypothèses et de mettre sur pied un programme de réinterview permettant de produire des données précises sur les flux bruts pendant la période visée. Weidman (1987) propose des modèles linéaires qui tentent de représenter les liens existant entre les changements observés et les changements réels. Il est vrai que les modèles, qui utilisent uniquement les données d'enquête, sont simplifiés à l'extrême; toutefois, ils illustrent la nécessité d'obtenir plus de renseignements sur la structure des erreurs dont sont entachées les déclarations des prestations que touchent les répondants de la SIPP bénéficiant des programmes de transfert.

Étude de vérification des enregistrements de la SIPP

Pour mettre en évidence la structure des erreurs relevées dans les résultats de la SIPP relativement à la nature des prestations touchées et aux montants en cause, on peut, entre autres, réaliser des études de validation des éléments communs à la fois aux données d'enquête et aux données administratives. Une étude de ce genre a été amorcée dans le cadre du programme de la SIPP dans le but d'en apprendre davantage sur la question de la qualité des réponses.

Cette étude devrait permettre de mieux juger de la qualité des données tirées de la SIPP et, éventuellement, de produire des estimations précises des erreurs de réponse et de non-réponse, afin que l'on puisse corriger les données d'enquête ou modifier les méthodes de collecte pour améliorer la qualité de ces données. L'étude vise notamment à évaluer les aspects suivants: 1) la qualité des données déclarées par les répondants ayant indiqué qu'ils sont bénéficiaires de divers programmes de transfert administrés par leur État ou le gouvernement fédéral; 2) la qualité des données recueillies au sujet du montant des prestations versées en vertu de ces programmes; 3) les corrélations démographiques pouvant être établies pour mesurer la qualité des données; 4) l'importance des erreurs de classification; 5) l'incidence qu'a sur la qualité des données le fait que les déclarations aient été faites par le répondant lui-même ou par un répondant substitut; 6) les effets dus au renouvellement des bénéficiaires entre les cycles d'interview. Quatre programmes à l'étude sont administrés par les États et les six autres, par le gouvernement fédéral. Moore et Marquis (1987) présentent des résultats très provisoires dont il ressort que les problèmes de déclaration sont différents pour l'aide aux familles à faible revenu avec enfants à charge et pour les bons alimentaires:dans le cas de l'aide aux familles, il y a un problème certain de sous-déclaration ainsi qu'une difficulté à situer le versement des prestations dans le temps, et dans le cas des bons alimentaires, seule la seconde difficulté a été observée.

6. CONCLUSION

Comme pour toutes les enquêtes permanentes de grande envergure, il faut poursuivre la recherche si l'on veut mieux saisir l'incidence des techniques d'enquête sur la qualité des données recueillies. Une enquête telle que la SIPP, dont la mise en oeuvre est complexe, exige que l'on s'efforce d'essayer de comprendre le processus d'évaluation. Le large éventail de sujets soulevés dans cet article – la collecte, l'estimation longitudinale, les définitions tenant compte du facteur temps et l'erreur de réponse – renseignent le lecteur sur les questions qui ont retenu l'attention tout au long du programme de mise au point de l'enquête et qui continuent de le faire depuis que l'enquête ellemême a été lancée.

REMERCIEMENTS

L'auteur remercie sincèrement les reviseurs pour leurs commentaires utiles ainsi que Mme Hazel Beaton pour les travaux de secrétariat qu'elle a assurés pendant la préparation de ce manuscrit.

BIBLIOGRAPHIE

BOWIE, C., et KASPRZYK, D. (1987). A review of the use of administrative records in the Survey of Income and Program Participation. *SIPP Working Paper Series No. 8721*, Washington, D.C.: U.S. Bureau of the Census.

BURKHEAD, D., et CODER, J. (1985). Gross changes in income recipiency from the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 351-356.

CHAKRABARTY, R. (1986). Composite estimation for SIPP: A preliminary report. *SIPP Working Paper Series No. 8610.* Washington, D.C.: U.S. Bureau of the Census.

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

CITRO, C. (1985). Alternative definitions of longitudinal households in the Income Survey Development Program: Implications for annual statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 381-386.

CITRO, C., HERNANDEZ, D., et HERRIOT, R. (1986). Longitudinal household concepts in SIPP: Preliminary results. *Proceedings of the U.S. Bureau of the Census Second Annual Research Conference*, 598-619.

CITRO, C., HERNANDEZ, D., et MOORMAN, J. (1986). Longitudinal household concepts in SIPP. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington, D.C., 361-366.

CODER, J. (1980). Some results from the 1979 Income Survey Development Program research panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 540-545.

ERNST, L. (1988). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys* (éds. D. Kasprzyk, G. Duncan, G. Kalton, et M.P. Singh), New York: John Wiley, à paraître.

ERNST, L., HUBBLE, D., et JUDKINS, D. (1984). Longitudinal family and household estimation in SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 682-687.

FAY, R.E., et HUGGINS, V.J. (1988). Use of administrative data in SIPP longitudinal estimation. Article à présenter à la Section on Survey Research Methods, American Statistical Association, August 1988.

GBUR, P., et DURANT, S. (1987). Testing telephone interviewing in the Survey of Income and Program Participation and some early results. Communication présenté au International Symposium on Telephone Survey Methodology, 1987, Charlotte, North Carolina.

HEERINGA, S., et LEPKOWSKI, J. (1986). Longitudinal imputation for the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 206-210.

HILL, D. (1987). Response errors around the seam: Analysis of change in a panel with overlapping reference records. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D. C., 210-215.

HUANG, H. (1984). Obtaining a cross-sectional estimate from a longitudinal survey: Experience of the ISDP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 670-675.

HUBBLE, D., et JUDKINS, D. (1986). Measuring the bias in gross flow estimates in the presence of auto-correlated response errors. *Proceedings of the Section on Survey Research Methods, American American Statistical Association*, Washington, D.C., 237-242.

JEAN, A., et McARTHUR, E. (1987.) Tracking Persons Over Time. *SIPP Working Paper Series No. 8701*, Washington, D.C.: U.S. Bureau of the Census.

JEAN, A., et McARTHUR, E. (1984). Some data collection issues for panel surveys with application to the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 745-750.

JUDKINS, D., HUBBLE, D., DORSCH, J.R., McMILLEN, D., et ERNST, L. (1984). Weighting of persons for SIPP longitudinal tabulations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 676-681.

KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.

KALTON, G. (1983). *Compensating for missing survey data*. Survey Research Center, University of Michigan.

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

KALTON, G., et KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 22-31.

KALTON, G., KASPRZYK, D., et McMILLEN, D.B. (1988). Nonsampling errors in panel surveys. In *Panel Surveys* (éds. D. Kasprzyk, G. Duncan, G. Kalton, et M.P. Singh), New York: John Wiley, à paraître.

KALTON, G., et LEPKOWSKI, J. (1985). Following rules in SIPP. *Journal of Economic and Social Measurement*, 13, 319-329.

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

KALTON, G., McMILLEN, D., et KASPRZYK, D. (1986). A review of nonsampling error issues in the Survey of Income and Program Participation. *Proceedings of the U.S. Bureau of the Census Second Annual Research Conference*, 147-165.

KALTON, G., et MILLER, M. (1986). Effects of adjustments for wave nonresponse on panel survey estimates. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 194-199.

KASPRZYK, D. (1983). Social Security number reporting, the use of administrative records, and the multiple frame design in the Income Survey Development Program. Dans *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program* (ISDP), (éd. M. David), Washington, D.C.: Social Science Research Council, 171-198.

KOBILARCIK, E., et SINGH, R. (1986). SIPP longitudinal estimation for persons' characteristics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 214-219.

LAMAS, E.J., et McNEIL, J.M. (1987). An analysis of the SIPP asset and liability feedback experiment. *Proceedings of the Social Statistics Section, American Statistical Association*, 194-199.

LEPKOWSKI, J. (1988). The treatment of wave nonresponse in panel surveys. Dans *Panel Surveys*, (éds. D. Kasprzyk, G. Duncan, G. Kalton, et M.P. Singh). New York: John Wiley, à paraître.

McMILLEN, D.B., et HERRIOT, R. (1985). Toward a longitudinal definition of households. *Journal of Economic and Social Measurement*, 13, 349-360.

MOORE, J.C., et KASPRZYK, D. (1984). Month-to-month recipiency turnover in the ISDP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 726-731.

MOORE, J.C., et MARQUIS, K. (1987). Utilisation des données de dossiers administratifs pour l'évaluation de la qualité des estimations d'enquêtes. Communication présenté au Symposium international sur l'utilisation statistique de données administratives, 1987, Ottawa, Canada.

NELSON, D., McMILLEN, D., et KASPRZYK, D. (1985). An overview of the Survey of Income and Program Participation: Update 1. *SIPP Working Paper Series No. 8401*. Washington, D.C.: U.S. Bureau of the Census.

OLSON, J. (1980). *Reports from the Site Research Test*, (éd. J. Olson). Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services, United States.

ROMAN, A.M., et O'BRIEN, D.V. (1984). The student follow-up investigation of the 1979 ISDP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 732-737.

U.S. Bureau of the Census (1986). Additional results from the SIPP telephone test. Note de service de J. Coder à A. Norton, avril 9, 1986.

U.S. Bureau of the Census (1985). SIPP research on SIPP oversampling/subsampling. Note de service de R. Singh à G. Shapiro et D. Kasprzyk, août 12, 1985.

WEIDMAN, L. (1987). Examination of relationships between actual and reported changes in the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 216-220.

WEIDMAN, L. (1986). Investigation of gross changes in income recipiency from the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 231-236.

WHITE, G., et HUANG, H. (1982). Mover follow-up costs for the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 376-381.

YCAS, M., et LININGER, C. (1981). The Income Survey Development Program: Design features and initial findings. *Social Security Bulletin*. Baltimore, Md.: Social Security Administration, 44.

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 63–73
Statistique Canada

Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes

DAN SCHNELL, WILLIAM J. KENNEDY, GARY SULLIVAN, HEON JIN PARK et WAYNE A. FULLER[1]

RÉSUMÉ

Le présent article contient la description d'un programme d'ordinateur personnel servant à l'estimation de la variance pour de grandes enquêtes. Ce programme, connu sous le nom de PC CARP, permet de calculer des estimations pour des agrégats, des rapports, des moyennes, des quantiles et des coefficients de régression et d'estimer les variances correspondantes.

MOTS CLÉS: Échantillon d'enquête; estimation de la variance; logiciel d'enquête.

1. INTRODUCTION

L'analyse de données d'enquête comporte habituellement un grand nombre d'observations et des calculs relativement complexes en ce qui a trait à la variance. Grâce aux derniers perfectionnements qui leur ont été appliqués, les ordinateurs personnels peuvent maintenant servir à traiter des données d'enquêtes complexes. Nous nous proposons dans cet article de décrire un programme d'ordinateur personnel qui est destiné à l'analyse de données d'enquête et qui a été mis au point à l'université Iowa State.

La création d'un logiciel d'ordinateur personnel pour l'estimation de la variance a été le résultat de la collaboration entre l'université Iowa State et le International Statistical Programs Center du U.S. Census Bureau. Par ce projet, le Census Bureau visait à offrir aux pays en voie de développement un logiciel qui pourrait être utilisé sur place pour traiter des données recueillies sur place. Le projet de l'université Iowa State s'inscrivait dans un programme du Census Bureau qui avait pour but de créer des logiciels pour la gestion des enquêtes et le contrôle et la totalisation des données.

Au début des années 1970, l'université Iowa State, se fondant sur les travaux d'Hidiroglou (1974) et Fuller (1975), élaborait un programme ayant pour but de calculer des coefficients de régression et la matrice estimée des covariances de ces coefficients pour des données d'enquête. Ce programme, connu sous le nom de SUPER CARP, devait plus tard être modifié de manière à inclure l'estimation d'agrégats et de rapports, les statistiques de sous-population, les tableaux à double entrée et les échantillons à deux degrés. La dernière révision de SUPER CARP remonte à 1980. Ce programme a ouvert la voie à la création de logiciels pour ordinateurs personnels. Pour cette raison, le programme d'ordinateur personnel a été baptisé PC CARP.

2. CARACTÉRISTIQUES DU PROGRAMME

PC CARP est conçu pour les modèles IBM PC, IBM PC/XT, IBM PC/AT et les ordinateurs compatibles. Il exige une mémoire d'au moins 410 kilo-octets ainsi qu'un coprocesseur "mathématique".

[1] Dan Schnell, Centers for Disease Control, 1600 Clifton Road, NE, Atlanta, (Georgie) 30333. William J. Kennedy, Gary Sullivan, Heon Jin Park et Wayne A. Fuller, Département de statistique, Iowa State University, Ames, Iowa 50011, États-Unis.

PC CARP traite avec autant de facilité et d'efficacité les longues et les petites séries de données. Il ne pose pas de limite pour le nombre de strates ou de grappes qui peuvent être incluses dans une série de données et peut accepter jusqu'à 50 paramètres à la fois. Il accepte, par ailleurs, les fichiers-disques à format de bloc fixe ou à structure libre.

Le programme peut servir à calculer les variances pour des échantillons à un ou à deux degrés avec correction pour population finie. Dans le cas des échantillons à plus de deux degrés, les termes de correction pour population finie n'existent qu'à deux niveaux. En ce qui concerne les échantillons à deux degrés, le programme calcule les taux de sondage à l'intérieur des grappes à partir des taux de sondage des strates et des poids des observations.

En règle générale, on connaîtra la strate et la grappe (unité primaire d'échantillonnage) auxquelles appartient chaque observation du fichier et on connaîtra aussi le poids de ces observations, qui correspond à l'inverse de la probabilité de sélection. L'utilisateur est libre d'introduire les taux de sondage du premier degré. Les plans de sondage simples, comme l'échantillonnage aléatoire simple, n'exigent pas tous ces renseignements. Dans de tels cas, il est possible d'avoir moins de paramètres.

Dans le cas d'un échantillonnage stratifié, toutes les observations qui sont contenues dans la même strate doivent être regroupées. S'il s'agit d'un échantillonnage par grappes, toutes les observations contenues dans la même grappe doivent être regroupées.

Le tableau 1 contient une liste des paramètres statistiques qui peuvent être calculés à l'aide du PC CARP. Outre les fonctions énumérées dans le tableau 1, il existe des cartes supplémentaires

Tableau 1

Fonctions d'analyse disponibles dans PC CARP

Analyse	Coefficient de variation	Matrice des covariances	Effet du plan	Remarques
Analyses de populations				
Estimation d'agrégats	X	X	X	50 variables au maximum
Estimation de rapports	X	X	X	50 variables au maximum sans les covariances, 15 avec covariances
Différence de rapports			X	15 variables au maximum
Analyses de strates				
Agrégats	X		X	50 variables au maximum
Moyennes	X		X	50 variables au maximum
Proportions	X		X	50 variables au maximum
Analyses de sous-populations				
Agrégats	X		X	Classement recoupé Variables multiples
Moyennes	X		X	Classement recoupé Variables multiples
Proportions	X		X	Classement recoupé Variables multiples
Rapports	X		X	Classement recoupé Variables multiples
Autres analyses				
Tableau à double entrée		X		50 cases au maximum, test de proportionnalité
Régression		X		50 variables au maximum Tests à plusieurs degrés de liberté, valeurs prédites résidus
Analyse unidimensionnelle			X	Variables multiples, fonctions de distribution cumulatives empiriques, quantiles

pour l'estimation de la fonction logistique et la stratification a posteriori. Ces fonctions supplémentaires sont décrites dans les sections 4.4 et 4.5. Un "X" sous la rubrique "matrice des covariances" signifie que l'on peut calculer la matrice des covariances d'un vecteur d'estimations du genre de celles indiquées dans la colonne de gauche. L'erreur type peut être calculée pour tous les paramètres statistiques mais la matrice des covariances d'un vecteur ne peut l'être que pour un nombre limité de paramètres. En outre, le coefficient de variation peut être calculé pour un grand nombre de paramètres. L'effet du plan, désigné par DEFF, peut être calculé sur demande pour un grand nombre de paramètres. Voir Kish (1965) pour une description de l'effet du plan.

Les analyses de population (agrégat, rapport et différence de rapports) et les analyses de strate s'effectuent d'une manière simple. Pour ce qui est des autres analyses (analyse de sous-population, de tableau à double entrée, de régression et analyse unidimensionnelle), on trouvera les détails pertinents à la section V.

Les analyses de sous-population offrent à l'utilisateur la possibilité de croiser des variables de classification. Grâce à cela, l'utilisateur peut créer de nouvelles structures de classification à l'aide de deux variables de classification ou plus. Supposons, par exemple, que les données d'entrée comprennent les variables de classification "âgé" (6 niveaux), "sexe" (2 niveaux) et "niveau d'instruction" (5 niveaux). Si nous combinons ces trois variables, nous allons obtenir une nouvelle structure de classification à soixante niveaux. Suivant cette structure, l'utilisateur peut établir des estimations pour un nombre indéterminé de variables dépendantes.

L'analyse des tableaux à double entrée est définie par deux variables de classification et une variable dépendante. Il est possible de définir plus d'une variable dépendante pour une paire de variables de classification. Pour chaque variable dépendante, le programme calcule les tables de fréquences de cases ainsi que les tables de proportions fondées sur les totaux de ligne, les totaux de colonne et le grand total. Il calcule aussi des erreurs types pour tous les estimateurs et produit une fonction des observations visant à tester l'hypothèse de la proportionnalité. Cette fonction est fondée sur une approximation de Satterthwaite de la distribution du critère chi-carré de Pearson. Voir aussi Rao et Scott (1984).

L'analyse de régression par les moindres carrés pondérés permet de calculer des estimations de coefficients ainsi qu'une matrice estimée des variances-covariances qui tient compte du plan de sondage. Ces calculs sont présentés dans Fuller (1975) et repris en bref dans Hidiroglou et coll. (1980). Des tests F à plusieurs degrés de liberté pour des séries de coefficients et les variables t habituelles sont également prévus. L'utilisateur a aussi la possibilité de calculer des résidus et des valeurs estimatives.

L'analyse unidimensionnelle produit des statistiques qui décrivent la distribution d'une variable. L'utilisateur définit la variable d'intérêt et détermine une sous-population en choisissant une catégorie d'une variable de classification. Ainsi, l'utilisateur pourrait choisir de recueillir des données sur le revenu personnel des membres de la catégorie des professions libérales dans la classification des professions. Le programme produit en l'occurrence des estimations de la moyenne de sous-population, de la variance, de la fonction de distribution, de quantiles et de l'écart interquartile.

3. DÉTAILS SUR LE PROGRAMME

PC CARP est écrit presque entièrement en FORTRAN, le langage de programmation scientifique le plus connu, et c'est le compilateur Professional FORTRAN d'IBM qui a été choisi pour le projet. Une petite partie du code — quelques sections de l'interface avec l'utilisateur — est rédigé en langage assembleur d'IBM.

Durant l'élaboration du programme, on a surtout cherché à créer une interface conviviale et à réduire au minimum le nombre de fois que les données doivent être parcourues. On a réussi à créer une interface conviviale en mettant en place un système de réponse interactif à écran. Par ailleurs, un algorithme monopasse pour l'estimation de la variance de paramètres statistiques

élémentaires a permis de réduire au minimum la fréquence de lecture des fichiers. Une seule lecture des données suffit pour calculer la plupart des estimateurs et leur variance.

Le programme peut calculer des estimateurs pour la population totale, pour chaque strate ou pour des sous-populations déterminées. Ces estimateurs sont pour la plupart des fonctions de totaux d'échantillon pondérés. Si l'on veut calculer, par exemple, l'estimateur des rapports $R_1 = Y_1/X_1$ et $R_2 = Y_2/X_2$, on recueille les totaux pour Y_1, X_1 Y_2, et X_2. Si l'estimation concerne la population totale, on forme ces totaux par une seule lecture des données. Dans le cas des estimations de strate, les totaux peuvent être accumulés, et combinés si nécessaire, puis restitués strate par strate. Comme les données sont groupées par strate, une lecture suffit pour déterminer les totaux de strate pour n'importe quelle quantité de strates. Pour ce qui a trait aux estimations relatives à une sous-population, une seule lecture des données pourrait ne pas suffire si le nombre de catégories définies par la structure de classification est élevé. L'analyse de régression et l'analyse unidimensionnelle exigent deux lectures de données.

Tous les estimateurs, à l'exception des estimateurs d'agrégats, sont des fonctions non linéaires des moments pondérés de l'échantillon. Il est donc nécessaire d'utiliser une méthode conforme à une fonction non linéaire pour estimer la variance de la distribution approximative de ce genre d'estimateurs. Voir Wolter (1985) pour une analyse de l'estimation de la variance pour des enquêtes complexes. La méthode de Taylor (méthode des écarts statistiques) est la méthode d'estimation de la variance utilisée dans PC CARP. Il a été démontré que la méthode de Taylor était en règle générale aussi efficace sinon plus efficace que les autres méthodes d'estimation de la variance des paramètres statistiques à l'étude (p. ex.: les rapports). Voir, par exemple, Frankel (1971). La méthode de Taylor pour l'estimateur de rapports est décrite dans des ouvrages aussi courants que celui de Cochran (1977) et la même méthode pour l'estimateur de coefficients de régression est décrite dans Fuller (1975).

Dans la plupart des cas, l'estimateur et sa variance estimée peuvent être calculés par la même lecture de données parce que l'approximation de Taylor du premier ordre de la variance peut être exprimée en fonction des variances des totaux. Par exemple, l'approximation de Taylor du premier ordre de $\hat{R} = \hat{Y}/\hat{X}$ est

$$\hat{R} \doteq R + X^{-1}(\hat{Y} - R\hat{X}),$$

où $R = Y/X$ est le rapport des totaux réels. Par conséquent, la variance estimée d'un rapport $\hat{R} = \hat{Y}/\hat{X}$ peut être calculée à l'aide de la variance estimée des totaux de Y, X, et $(Y - X)$. De même, la matrice estimée des covariances pour $\hat{R}_1 = \hat{Y}_1/\hat{X}_1$ et $\hat{R}_2 = \hat{Y}_2/\hat{X}_2$, peut être calculée à l'aide des variances estimées des totaux de Y_1, X_1, $(Y_1 - X_1)$, Y_2, X_2, $(Y_2 - X_2)$, $(Y_1 - Y_2)$, $(Y_1 - X_2)$, $(Y_2 - X_1)$, et $(X_1 - X_2)$.

L'algorithme servant au calcul de la moyenne pondérée, des sommes des carrés pondérées et des matrices de produits croisés est décrit dans Herraman (1968). Pour les valeurs d'échantillon $\{X_i\}$ et les poids correspondants $\{W_i\}$, la suite des moyennes pondérées $\bar{X}_K$, et des sommes des carrés corrigées pondérées S_K est calculée à l'aide des formules suivantes

$$\bar{X}_K = \bar{X}_{K-1} + a_K d_K \text{ et } S_K = S_{K-1} + D_K - D_K a_K,$$

où $d_k = X_K - \bar{X}_{K-1}$, $a_K = W_K(\Sigma_{i=1}^{K} W_i)^{-1}$, et $D_K = d_k^2 W_K$.

Il est possible de calculer simultanément trois composantes de variance pour n'importe quel estimateur donné. Il s'agit de la composante de variance du premier degré, de la composante de variance optionnelle du second degré et de la variance optionnelle d'un échantillonnage aléatoire simple utilisée dans le calcul de l'effet du plan. Calculer toutes ces composantes par une seule

lecture de données exige beaucoup d'espace mémoire. Cependant, lorsqu'on travaille avec de grands échantillons, la suppression de lectures entières compense l'utilisation d'espace mémoire additionnel.

Le programme exécute régulièrement des vérifications visant à déceler des erreurs de calcul comme un diviseur nul. Si, par exemple, l'utilisateur introduit une série de données avec une strate à une seule grappe, le programme attribuera une variance nulle à la strate, complètera les calculs et fera imprimer un message d'erreur qui signalera la strate en question.

Le système de traitement d'erreurs a été conçu de manière à prévenir les arrêts de programme causés par des erreurs de l'utilisateur qui peuvent être corrigées facilement. Le programme renferme des fonctions de vérification pour l'absence de réponse, l'inexactitude de noms de fichiers et l'inexactitude de définitions de variables d'analyse. Si une erreur est décelée, le PC CARP permet à l'utilisateur de réintroduire les données ou de sortir du programme.

On a évalué la précision du programme en construisant des exemples d'application puis en comparant les résultats à ceux obtenus à l'aide du programme principal SUPER CARP. La série de données de Longley (1967) a servi à évaluer la précision du programme de régression. Des vérifications supplémentaires ont été faites à l'aide de PROC MATRIX du progiciel SAS. Voir Barr et coll. (1979). On a constaté que PC CARP avait un niveau de précision numérique équivalent à celui des progiciels principaux. On a aussi vérifié la cohérence interne de PC CARP en calculant des estimateurs équivalents à l'aide d'options différentes (p. ex., calcul d'une moyenne de sous-population au moyen de l'option "analyse de sous-population" et de l'option "analyse de rapports").

Lorsque PC CARP demande de l'information à l'utilisateur, une série de questions à réponse brève apparaît à l'écran avec des instructions détaillées. La première série d'affichages demande à l'utilisateur de fournir des renseignements sur l'organisation et l'emplacement des données. Des options "Help" et "Go Back" peuvent être utilisées à plusieurs endroits.

La seconde phase de l'exécution du programme a trait à la définition de l'analyse. Dans cette phase, l'utilisateur choisit le genre d'analyse, les options pertinentes et les variables d'analyse. Un nombre indéfini d'analyses peuvent être exécutées à l'aide des données définies dans la phase 1.

4. CHARACTÉRISTIQUES SPECIALES

4.1 Tableau à double entrée

Comme nous l'avons vu dans la Section 2, cette option fournit automatiquement à l'utilisateur quatre tableaux dont les éléments sont déterminés par le genre de liaisons qui ont été définies dans la construction du tableau à propos des fréquences par case.

Nous exposons ci-dessous la méthode qui a servi à construire le tableau des proportions de case et la matrice estimée des covariances des proportions. Supposons que le tableau est formé de R lignes et de C colonnes et posons $\hat{Y}_{rc}$ comme la fréquence estimée pour la case rc. Soit $\hat{Y}$ le vecteur colonne $RC \times 1$ des fréquences de case, créé en reproduisant l'une au-dessous de l'autre, dans l'ordre normal, les colonnes du tableau $R \times C$. Soit

$$\hat{Y}_{..} = \sum_{r=1}^{R} \sum_{c=1}^{C} \hat{Y}_{rc},$$

$$\hat{P}_{rc} = \hat{Y}_{..}^{-1} \hat{Y}_{rc}$$

le total estimé de la population et la proportion de case estimée pour la case rc respectivement.

Soit $\hat{P}$ le vecteur colonne $RC \times 1$, analogue à $\hat{Y}$, composé des RC valeurs $\hat{P}_{rc}$ disposées en colonnes. La matrice estimée des covariances pour $\hat{P}$ est

$$\hat{V}_{PP} = \hat{Y}_{..}^{-2}[I_{RC} - (\hat{P} \otimes J'_{RC})]\, \hat{V}_{YY}[I_{RC} - (\hat{P} \otimes J'_{RC})]',$$

où $\hat{V}_{YY}$ est la matrice estimée des covariances du vecteur des fréquences de case $\hat{Y}$, I_{RC} est la matrice unité de dimension $RC \times RC$ et J_{RC} est un vecteur colonne de dimension $RC \times 1$, constitué de uns.

La matrice $\hat{V}_{PP}$ sert à calculer la variable à tester pour l'hypothèse de la proportionnalité. L'hypothèse nulle, en l'occurrence, est que les éléments intérieurs du tableau de la population sont le produit des proportions marginales. Voir Rao et Scott (1984) pour une analyse des tests relatifs à ce genre d'hypothèses. Le test dans PC CARP est fondé sur une approximation de Satterthwaite de la distribution du critère chi-carré de Pearson, construite comme s'il s'agissait de proportions multinomiales. L'approximation est valide pour n'importe quelle variable d'analyse.

4.2 Estimation de quantiles

L'option "analyse unidimensionnelle" permet notamment de calculer des estimations de quantiles ainsi qu'un estimateur de l'erreur type des quantiles. Lorsqu'on veut calculer des quantiles, il faut tout d'abord construire un estimateur de la fonction de distribution cumulative. Une première lecture des données permet de déterminer l'intervalle des observations de même que la moyenne et l'écart type de l'échantillon. Par la même occasion, on relève les trois plus grandes et les trois plus petites observations.

La fonction de distribution cumulative estimée est définie par

$$\hat{F}_Y(x) = \left(\sum_{t=1}^{m} w_t Z_{St} \right)^{-1} \sum_{t=1}^{m} w_t Z_{St} I_Y(x),$$

où la sommation est étendue aux m éléments de l'échantillon, w_t est le poids d'échantillon, Z_{St} est une fonction indicatrice qui prend la valeur 1 si l'observation est comprise dans la sous-population d'intérêt et la valeur 0 dans le cas contraire, et $I_Y(x)$ est égale à 1 si $Y < x$ et est égale à zéro dans le cas contraire. Le domaine de la variable est divisé en 100 intervalles et la fonction de distribution cumulative est estimée aux 101 points définis par cette division.

Le programme estime ensuite la matrice des covariances pour la fonction de distribution estimée évaluée à 25 points, $j = 1, 5, \ldots, 96$. Les erreurs types estimées sont lissées à l'aide d'une moyenne mobile de trois points et une interpolation permet d'établir une erreur type estimée pour chacun des 101 points de la fonction de distribution estimée. Une interpolation linéaire permet de créer une fonction de distribution estimée qui est croissante au sens large. À l'aide des erreurs types lissées, le programme détermine une borne supérieure monotone croissante et une borne inférieure monotone croissante, qui délimitent un intervalle de confiance ponctuel de 95% pour la fonction de distribution. Ces bornes sont ensuite inversées pour former des intervalles de confiance de 95% pour les quantiles. L'écart interquartile et l'erreur type correspondante sont aussi estimés.

L'estimation de quantiles repose sur une théorie qui suppose l'existence d'une fonction de distribution de superpopulation sous-jacente avec une densité positive. Voir Francisco (1987) pour les aspects théoriques et Park (1987) pour les aspects relatifs aux calculs.

4.3 Estimation par régression

Les estimations des coefficients d'un modèle de régression linéaire sont calculées par la méthode des moindres carrés pondérés. Suivant la méthode décrite dans Fuller (1975), le programme calcule un estimateur de la matrice des covariances du vecteur des coefficients en tenant compte du plan de sondage.

Le vecteur des coefficients est estimé par

$$\hat{b} = (X'WX)^{-1}X'WY,$$

où X est la matrice $n \times p$ des valeurs de la variable indépendante, $\hat{Y}$ est le vecteur à n dimensions des valeurs de la variable dépendante, W est une matrice dont la diagonale est formée des poids des observations et le reste est formé de zéros, et n est le nombre total des observations. La variance de $\hat{b}$ est estimée par

$$\hat{V}(\hat{b}) = (X'WX)^{-1}\hat{G}_W(X'WX)^{-1}.$$

La matrice $\hat{G}_W$ est

$$\hat{G}_W = C \sum_{i=1}^{L} h_i \sum_{j=1}^{n_i} (\hat{d}_{ij} - \bar{d}_{i.})(\hat{d}_{ij} - \bar{d}_{i.})',$$

où

$$\hat{d}_{ij} = \sum_{k=1}^{m_{ij}} X_{ijk}\hat{v}_{ijk}W_{ijk},$$

$$\hat{v}_{ijk} = Y_{ijk} - \hat{b}'X_{ijk},$$

$$\bar{d}_{i.} = n_i^{-1}\sum_{j=1}^{n_i} \hat{d}_{ij},$$

$h_i = (n_i - 1)^{-1}n_i$, $C = (n - p)^{-1}(n-1)$, m_{ij} est le nombre d'éléments contenus dans la grappe j de la strate i, n_i est le nombre de grappes dans la strate i, n est le nombre total d'observations, L est le nombre de strates et p est le nombre de coefficients estimés. L'estimateur de la variance calculé par cette option diffère de l'estimateur habituel calculé par les moindres carrés pondérés en ce que la matrice $\hat{G}_W$ est substituée à $(X'WX)s^2$.

Le programme calcule une fonction des observations R^2 multiple pour les modèles qui ont une ordonnée ou une abscisse à l'origine. De plus, il effectue toujours un test F pour la régression générale et offre à l'utilisateur la possibilité de tester des sous-ensembles de coefficients.

4.4 Régression logistique

Cette option permet d'obtenir des estimations du modèle logistique à plusieurs variables. Les algorithmes destinés à la régression logistique ont été élaborés après que la version initiale de PC CARP a vu le jour. Comme la fonction de moyenne pour le modèle logistique est non linéaire dans les paramètres, les estimateurs sont calculés à l'aide d'un algorithme d'itération fondé sur les moindres carrés pondérés. Les variances des estimateurs sont calculées au moyen des méthodes décrites dans Fuller (1975) et appliquées à l'estimation non linéaire. Voir aussi Binder (1983). Au point de vue des modalités d'application, l'option "régression logistique" est en tout point semblable à l'option "régression". Par exemple, les variables dépendantes et indépendantes y sont définies de la même façon.

4.5 Stratification a posteriori

Après que l'on a élaboré la version originale du programme PC CARP, une option "stratification a posteriori" a été mise au point pour un bon nombre des estimateurs. La stratification a posteriori est censée être une opération par laquelle on corrige les poids de manière à obtenir des estimations pour certaines catégories, qui concordent avec des totaux de population connus. C'est ce que Fuller et Sullivan (1987) appellent la stratification a posteriori gamma.

Le programme calcule la variance de l'estimateur de stratification a posteriori à partir d'une formule dans laquelle l'estimateur est exprimé comme la somme d'estimateurs de rapports.

4.6 Regroupement de strates

Aux fins du calcul de la variance, l'utilisateur peut recourir à cette option afin d'éliminer les strates à une grappe. Lorsque cette option est appliquée, chaque strate à une grappe est assimilée à la strate suivante dans la série de données. On modifie en conséquence le numéro de strate et le numéro de grappe des enregistrements touchés par cette fusion. S'il existe des taux de sondage de strates, les nouveaux taux sont définis par

$$f_i^* = (n_i f_i^{-1} + n_{i+1} f_{i+1}^{-1})^{-1}(n_i + n_{i+1}),$$

où la strate i, avec $n_i = 1$, a été combinée avec la strate $i + 1$. Ces nouveaux taux sont enregistrés dans un fichier annexe pour usage ultérieur. L'ordre de classement des strates influera sur le mode de regroupement des séries de données et les taux de sondage de strates. Le programme exige une lecture de données additionnelle lorsque l'utilisateur recourt à l'option "regroupement de strates" ou à l'option "à deux degrés".

4.7 Imputation par la méthode "hot deck"

PC CARP exige des séries de données complètes pour l'analyse. Pour le contrôle et l'imputation des données, beaucoup d'utilisateurs choisiront d'écrire un programme spécial ou encore utiliseront l'un des programmes offerts à cette fin dans la gamme PC.

Un programme d'imputation par "hot deck", appelé PRE CARP, peut être fourni avec le PC CARP pour ceux qui en font la demande. La méthode "hot deck" consiste à remplacer une valeur manquante par la valeur de l'article correspondant de l'enregistrement qui vient immédiatement avant dans le fichier de données. PRE CARP permet à l'utilisateur de définir une variable de classification qui peut renfermer jusqu'à 10 catégories de telle sorte que la valeur manquante est remplacée par la valeur de l'enregistrement précédent dans la même catégorie. PRE CARP créera aussi une variable indicatrice pour chaque variable pour laquelle il y a des valeurs manquantes. Cette variable indicatrice peut être ensuite utilisée avec l'option "analyse de sous-population" pour calculer des moyennes en fonction des observations originales.

5. EXEMPLES

Dans cette section, nous effectuons plusieurs analyses à l'aide d'une série de données composée et présentons les durées d'exécution. Le but des passages d'essai n'est pas d'étudier toutes les combinaisons possibles de facteurs qui peuvent influer sur la durée de l'exécution mais d'estimer le temps requis pour exécuter quelques-unes des analyses offertes dans le programme.

Les données d'essai ont été constituées à partir d'un sous-ensemble de la seconde édition de la National Health and Nutrition Examination Survey (NHANES II). La série ainsi constituée comprend 2400 observations réparties en 32 strates. Chaque strate compte deux unités primaires d'échantillonnage et ces unités d'échantillonnage sont de taille variée. De plus, chaque observation est affectée d'un poids de sondage non nul.

Les variables contenues dans la série de données sont les suivantes:

1. Sexe	1 = homme, 2 = femme
2. Origine raciale	1 = blanc, 2 = noir, 3 = autre
3. Caractère agricole	1 = ménage non agricole, 2 = ménage agricole
4. Revenu	revenu du ménage (en milliers de dollars)
5. Âge	âge (en années).

Figure 1. Résultats de l'analyse C; âge moyen selon le sexe et l'origine raciale.

MOYENNES DE SOUS-POPULATION

Variable dépendante: âge

Catégorie	Estimation	E.T.	C.V.	DEFF (effet du plan)
Sexe = 1.0000	Origine raciale = 1.0000			
	3.06811D+01	6.19678D-01	2.0197D-02	1.3967D+00
Sexe = 1.0000	Origine raciale = 2.0000			
	3.13016D+01	7.88580D-01	2.5193D-02	9.6384D-01
Sexe = 1.0000	Origine raciale = 3.0000			
	3.41579D+01	2.24111D+00	6.5610D-02	2.0965D+00
Sexe = 2.0000	Origine raciale = 1.0000			
	1.33742D+01	3.18904D-01	2.3845D-02	1.2588D+00
Sexe = 2.0000	Origine raciale = 2.0000			
	*********			
Sexe = 2.0000	Origine raciale = 3.0000			
	1.71957D+01	9.53816D-01	5.5468D-02	1.1389D+00

Une variable qui prend la valeur 1 pour chaque observation (variable constante) a été créée par le programme. Les analyses exécutées sont les suivantes:

A. Revenu moyen de la population échantillonnée
B. Revenu moyen par strate
C. Âge moyen pour la classification selon le sexe et l'origine raciale
D. Fonctions de distribution de l'échantillon selon l'âge pour les groupes agricoles et non agricoles.

L'analyse A (estimation du revenu moyen) a été exécutée au moyen de l'option "analyse de rapports", où le revenu servait de numérateur et la variable constante de dénominateur. Les estimations du revenu moyen par strate (analyse B) ont été calculées directement au moyen de l'option "moyenne de strate". L'analyse C a été exécutée au moyen de l'option "moyenne de sous-population" en croisant les variables de classification "sexe" et "origine raciale" et en définissant l'"âge" comme la variable dépendante. Les résultats de l'analyse C sont reproduits dans la figure 1. Les symboles "*****" qui paraissent sous la classification "sexe = 2 origine raciale = 2" indiquent qu'il n'y avait pas d'observations relatives à cette catégorie. Les valeurs de l'effet du plan mettent en relief l'importance de tenir compte du plan de sondage dans l'estimation de la variance. Par exemple, l'effet du plan pour l'estimation correspondant à la catégorie "sexe = 1 et origine raciale = 3" est environ 2. Cela signifie que la variance estimée de la moyenne de l'échantillon pour un échantillonnage aléatoire simple équivaut à la moitié de la variance estimée pour un échantillonnage en grappes stratifié. Les caractéristiques de la distribution de l'âge pour chacun des deux états de la variable "caractère agricole" ont été estimées au moyen de l'option "analyse à une variable". La figure 2 donne les résultats de l'analyse à une variable pour les ménages non agricoles. Toutes les variances et les estimations d'erreurs types indiquées dans cette figure tiennent compte du plan de sondage.

La durée d'exécution des analyses A, B, C et D pour les 2,400 observations a été de 70, 135, 120 et 360 secondes respectivement. L'exécution s'est faite sur IBM PC AT, les données étant stockées sur le disque dur et lues en structure non imposée. Les taux de sondage des strates n'ont pas été introduits dans le programme. Les résultats ont été acheminés au récepteur de contrôle et à un fichier-disque. Les effets du plan pour les estimations ont dû être considérés dans les quatre analyses. Les trois premières analyses ont exigé chacune une seule lecture des données. Un plus

Figure 2. Résultats de l'analyse unidimensionnelle pour les ménages non agricoles (Exemple D).

ANALYSE UNIDIMENSIONNELLE 1

Variable de classification: caractère agricole état: 1

Nombre d'éléments de l'échantillon dans la sous-population = 111

Variable dépendante: âge

Variance de la sous-population = 4.25645D+02

C.V. de la sous-population = 7.14101D-01

Moyenne de la sous-population

Estimation	E.T.	C.V.	DEFF
2.8891089D+01	2.2543875D+00	7.80305D-02	9.86336D-01

Valeurs extrêmes des éléments de l'échantillon dans la sous-population

		Identification de la première observation		
Valeurs les plus faibles	Nombre de valeurs observées	strate	grappe	poids
1.000D+00	1	32	1	3.000D+00
2.000D+00	1	15	1	2.000D+00
3.000D+00	3	10	2	3.000D+00

		Identification de la première observation		
Valeurs les plus élevées	Nombre de valeurs observées	strate	grappe	poids
7.400D+01	2	29	1	3.000D+00
7.100D+01	2	7	1	2.000D+00
7.000D+01	4	7	1	2.000D+00

Quantiles

	Estimation	E.T.	Intervalle de confiance de 95%
0.01	2.2690811D+00	7.3167585D-01	(8.05729D-01, 3.73243D+00)
0.05	4.2364814D+00	1.1977759D+00	(3.34942D+00, 8.14053D+00)
0.10	7.7750203D+00	1.3563759D+00	(5.36691D+00, 1.07924D+01)
0.25	1.3652930D+01	1.4225576D+00	(9.99238D+00, 1.56826D+01)
0.50	1.9449315D+01	2.2740912D+00	(1.57192D+01, 2.48156D+01)
0.75	4.5698071D+01	4.7577709D+00	(3.58936D+01, 5.49247D+01)
0.90	6.2787426D+01	2.3472775D+00	(5.61417D+01, 6.45308D+01)
0.95	6.5923423D+01	1.2228344D+00	(6.43837D+01, 6.92750D+01)
0.99	7.1714993D+01	1.1425033D+00	(7.05401D+01, 7.40000D+01)

Écart interquartile

Estimation	E.T.
3.2045141D+01	4.2434890D+00

grand nombre de paramètres statistiques ont été calculés pour les analyses B et C que pour l'analyse A. Quant à l'analyse D, elle a nécessité quatre lectures de données, soit deux pour chaque analyse à une variable.

REMERCIEMENTS

Outre le U.S. Census Bureau, le Soil Conservation Service du Département de l'agriculture de É.-U. a apporté son soutien au projet PC CARP. Nous tenons à remercier les arbitres et les éditeurs pour leurs commentaires utiles. Les personnes dans les pays en voie de développement qui s'intéressent au programme PC CARP sont priées de s'adresser à: PC CARP, International Statistical Programs Center, U.S. Bureau of the Census, Washington (D.C.), É.-U. 20233. Les autres personnes sont priées de s'adresser à: PC CARP, Statistical Laboratory, Iowa State University, Ames (Iowa), É.-U. 50011. Des exemplaires du programme ont été offerts à des organismes répartis dans plus de 20 pays.

BIBLIOGRAPHIE

BARR, A.J., GOODNIGHT, J.H., SALL, J.P., BLAIR, W.H., et CHILKO, D.M. (1979). SAS User's Guide. SAS Institute, Raleigh, North Carolina.

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

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

DISKIN, B.A. (1985). Microcomputers in developing country statistical offices: current use and a look to the future. Actes de la 45^{e} assemblée de l'Institut international de statistique, Amsterdam.

FRANCISCO, C.A. (1987). Estimation of quantiles and the interquartile range in complex surveys. Thèse de doctorat non publiée, Iowa State University, Ames, Iowa.

FRANKEL, M.R. (1971). *Inference from Survey Samples*. Institute of Social Research, University of Michigan, Ann Arbor.

FULLER, W.A. (1975). Regression analysis for sample survey. Sankhyā C 37, 117-132.

FULLER, W.A., et SULLIVAN, G. (1987). Gamma Post Stratification. Report to the U.S. Bureau of the Census. Département de statistique, Iowa State University, Ames, (Iowa).

HERRAMAN, C. (1968). Sums of squares and products matrix. *Applied Statistics*, 17, 289-292.

HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Thèse de doctorat non publiée, Iowa State University.

HIDIROGLOU, M.A., FULLER, W.A., et HICKMAN, R.D. (1980). *SUPER CARP*, Département de statistique, Iowa State University, Ames, Iowa.

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

LONGLEY, J.W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association*, 62, 819-841.

PARK, H.J. (1987). Univariate Analysis in PC CARP. Creative Component pour le M.S. non publiée, Iowa State University, Ames, Iowa.

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.

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

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 75–92
Statistique Canada

Test des opérations de redressement de 1986 dans le Central Los Angeles County

GREGG DIFFENDAL[1]

RÉSUMÉ

En vue du recensement décennal de 1990, le Census Bureau a étudié la possibilité de redresser les chiffres du recensement pour tenir compte du taux de sous-dénombrement estimé. À cette fin, il a exécuté un recensement d'essai dans un secteur à majorité hispanique du Central Los Angeles County afin de vérifier le calendrier et les caractéristiques des opérations de redressement réalisées au moyen d'une enquête post-censitaire (EP). Cet article vise à exposer les méthodes qui ont été utilisées pour produire des données de recensement qui tiennent compte de la population non dénombrée; on y trouve également les résultats de l'application de ces méthodes. Les méthodes utilisées pour le redressement des données de recensement comprenaient l'élaboration d'un plan de sondage, l'estimation "bivalente" et le calcul d'estimations régionales. Le plan de sondage prévoyait un échantillon d'îlots stratifiés selon l'origine ethnique. L'appariement s'est fait par ordinateur tandis que les opérations de contrôle et de résolution ont été exécutées manuellement. L'estimateur "bivalent", appelé aussi estimateur de Petersen ou saisie-resaisie, a servi à estimer la population. À cause de la nature des recensements, les chiffres du recensement ont été redressés avant de servir au calcul de l'estimateur bivalent. Avant de corriger les estimations régionales, on a ajusté un modèle de régression au facteur de redressement (estimateur bivalent divisé par le chiffre du recensement) afin de réduire les effets de la variabilité d'échantillonnage. Un estimateur synthétique a permis d'effectuer le redressement jusqu'au niveau de l'îlot. Les résultats de l'estimation bivalente sont présentés pour la région d'essai selon les trois principaux groupes ethniques (hispanique, asiatique, autre), le mode d'occupation, l'âge et le sexe. Enfin, nous présentons en bref les résultats du redressement des estimations régionales du recensement par îlot et nous en faisons l'analyse.

MOTS CLÉS: Sous-dénombrement dans le recensement; estimation bivalente; estimation synthétique; enquête post-censitaire.

1. INTRODUCTION

Depuis le premier recensement de la population des É.-U. en 1790, on a toujours eu de la difficulté à dénombrer toutes les personnes qui devaient l'être. Grâce aux progrès réalisés en démographie et en statistique, on a pu établir à partir de 1950 des estimations du taux de couverture du recensement. Ces estimations ont servi à évaluer les faiblesses du recensement et à définir les points qui devaient être améliorés en vue des recensements subséquents. Les estimations du taux de couverture du recensement ont connu une amélioration soutenue depuis la publication des premières estimations du genre en 1950. Une série d'estimations indique que le taux de sous-dénombrement pour l'ensemble du pays était de 4.4% en 1950, de 3.3% en 1960, de 2.8% en 1970 et de 1% en 1980. Malgré cette baisse soutenue du taux de sous-dénombrement, les estimations demeurent élevées pour certains groupes. Dans le cas des personnes de race noire, par exemple, le taux de sous-dénombrement se maintient à environ 5 points de pourcentage au-dessus de la moyenne nationale.

Les résultats indiquent aussi des taux de sous-dénombrement élevés pour d'autres groupes ethniques, notamment la population d'origine hispanique. De même, les taux de sous-

[1] Gregg Diffendal, Undercount Research Staff, Statistical Research Division, Bureau of the Census, Washington (D.C.) 20233, É.-U. Cet article est un compte rendu des recherches qui ont été faites par un membre du personnel du Census Bureau. Les opinions qui y sont exprimées sont celles de l'auteur et ne reflètent pas nécessairement la position du Census Bureau.

dénombrement sont plus élevés dans les grandes villes que dans les régions rurales et plus élevés chez les hommes que chez les femmes. Les membres du groupe d'âge 20-45 présentent aussi un fort taux de sous-dénombrement.

Depuis 1950, on mesure le taux de sous-dénombrement aux États-Unis au moyen d'une enquête post-censitaire (EP) et d'une analyse démographique. Le Census Bureau a laissé savoir que ces deux méthodes seraient les principaux moyens d'estimer le taux de sous-dénombrement pour le recensement de 1990. Une EP utilise un échantillon indépendant de personnes qui sont appariées aux enregistrements du recensement afin d'estimer la population totale. Marks (1978) et le U.S. Census Bureau (1979) décrivent les expériences qui ont déjà été faites au moyen d'une EP pour mesurer le taux de couverture du recensement. Par ailleurs, l'analyse démographique fait appel aux registres des naissances et des décès et à d'autres fichiers administratifs pour estimer la population totale des États-Unis. Fay et coll. (1988) présentent les estimations des taux de sous-dénombrement de 1980 qui ont été établies à l'aide de l'analyse démographique et du programme post-censitaire (PP).

En 1980, une analyse plus approfondie des chiffres du recensement a donné lieu à un certain nombre de poursuites judiciaires par lesquelles on exigeait un redressement de ces chiffres. Ces poursuites avaient été engagées pour des raisons comme la différence de taux de sous-dénombrement entre les noirs et les non-noirs, la création dans les années 1970 d'un mécanisme de partage des revenus qui liait directement des fonds publics aux chiffres de population et la baisse de population dans des villes et des États pour lesquels le taux de sous-dénombrement avait toujours été supérieur à la moyenne. Le U.S. Census Bureau s'est refusé au redressement des chiffres de 1980 en faisant valoir que le calcul du taux de sous-dénombrement était exposé aux erreurs et qu'un redressement des chiffres du recensement ne serait pas vraiment utile.

Le Census Bureau s'est néanmoins engagé après 1980 dans un programme de recherche visant à étudier diverses façons d'améliorer le calcul des taux de sous-dénombrement (Mulry et coll. 1981 et Hogan 1984). À ce sujet, Hogan (1984) a proposé une série de tests à effectuer concurremment avec les recensements d'essai. Cela a commencé par une EP à Tampa (Floride) en 1985 dans le but de tester et d'évaluer l'appariement par ordinateur. Ce test a confirmé la faisabilité de ce mode d'appariement (Jaro et Childers 1986). Des recensements d'essai et des EP ont aussi été réalisés en 1986 à Los Angeles et au Mississipi. Dans le premier cas, l'EP visait à tester le calendrier et les caractéristiques des opérations de redressement des chiffres du recensement. Dans le second cas, l'EP visait à évaluer les opérations d'une telle enquête dans une région d'essai rurale (Anolik 1988).

Une enquête précensitaire (EPR) a aussi été réalisée à Los Angeles en 1986 pour déterminer s'il était possible de réduire le délai des opérations de redressement en exécutant une partie du travail sur le terrain avant le recensement plutôt qu'après comme dans une EP (Wolfgang 1987). En 1987, on a réalisé une enquête précensitaire dans le Dakota du Nord afin d'évaluer les opérations d'une telle enquête dans des régions rurales; dans ce cas toutefois, on a procédé à des interviews sur place au lieu d'envoyer des questionnaires par la poste comme cela s'était fait dans les autres régions d'essai. Enfin, on est à préparer la répétition générale de 1988. Cette répétition générale permettra de tester toutes les opérations de recensement avant l'exécution du recensement décennal de 1990.

Cet article porte plus spécialement sur l'EP de 1986 pour le Central Los Angeles County, appelée Test des opérations de redressement (TOR) et réalisée concurremment avec le recensement d'essai. La région d'essai comprenait trois grands groupes ethniques: les personnes d'origine hispanique (qui constituaient environ 75% de la population totale), les personnes d'origine asiatique (environ 15% de la population totale) et les autres personnes, en majorité de race blanche (environ 10% de la population totale). L'EP a révélé un taux de sous-dénombrement estimé de 9%. Pour les trois principaux groupes ethniques de la région d'essai, les taux de sous-dénombrement estimés étaient de 9.8% pour les personnes d'origine hispanique, de 7.3% pour les personnes d'origine asiatique et de 6.2% pour les autres. Dans cet article, nous décrivons la manière dont nous estimons ces taux.

Dans la Section 2, nous exposons la méthode utilisée en 1986 pour estimer les taux de sous-dénombrement et indiquons comment ces estimations servent au redressement des chiffres du recensement. Dans la Section 3, nous analysons le calendrier des opérations prévu pour le TOR, y compris les opérations sur le terrain et l'appariement. La Section 4 contient un résumé des estimations des taux de sous-dénombrement pour les strates formées a posteriori et les îlots. Enfin dans la Section 5, nous faisons un résumé des principaux résultats de notre analyse et tirons quelques conclusions.

2. MÉTHODOLOGIE

2.1 Aperçu des échantillons utilisés dans l'estimation

Pour estimer la population, on a utilisé deux échantillons dans l'EP: l'échantillon P (pour population) et l'échantillon D (pour dénombrement). Le premier sert à mesurer les omissions du recensement tandis que le second sert à mesurer les enregistrements erronés du recensement.

L'échantillon P est un échantillon d'îlots qui s'accompagne d'un listage indépendant d'unités de logement et d'interviews sur place tandis que les données de l'échantillon D sont les enregistrements (chiffres) du recensement qui se rapportent aux îlots échantillonnés. L'échantillon P a servi à recueillir les données nécessaires à l'appariement et à l'estimation, y compris le domicile au jour du recensement. Pour les besoins du plan de sondage, il a fallu définir les éléments de l'échantillon P. L'échantillon P était constitué de toutes les personnes qui demeuraient aux adresses échantillonnées au moment de l'EP. Nous aurions pu choisir d'interviewer les personnes qui demeuraient à ces adresses le jour du recensement. Nous avons éliminé cette possibilité car tous les cas de déménagement impliquent une interview par personne interposée (c'est-à-dire une interview réalisée auprès d'une personne qui n'est pas membre du ménage). Selon la méthode que nous avons adoptée, des personnes ayant déménagé pouvaient se trouver à l'adresse échantillonnée et on pouvait les interviewer sans avoir affaire à un enquêté-substitut. Cependant, toutes les personnes qui ont quitté la région d'essai entre le jour du recensement et le jour de l'EP n'avaient aucune chance de faire partie de l'échantillon P. Par ailleurs, toutes les personnes de l'échantillon P qui demeuraient à l'extérieur de la région d'essai le jour du recensement étaient considérées comme exclues du champ de l'enquête.

Après l'interview, toutes les personnes de l'échantillon P ont été appariées à des enregistrements du recensement, l'appariement se faisant par ordinateur avec une vérification manuelle. Il a aussi fallu définir pour les besoins du plan de sondage l'étendue du secteur de recherche pour l'appariement. Selon l'EP, une personne de l'échantillon P était considérée comme appariée si elle avait été enregistrée lors du recensement n'importe où dans la région d'essai. On aurait pu définir un secteur de recherche moins étendu comme l'îlot servant à l'EP et les îlots avoisinants. Ainsi, il y a concordance pour une personne de l'échantillon P seulement lorsque si la personne correspondante dans le recensement se trouve dans ce secteur de recherche. À ce propos, l'EP de 1990 utilisera un secteur de recherche restreint pour l'appariement.

Tous les cas qui n'ont pu être appariés ont dû faire l'objet d'un suivi qui visait à recueillir des renseignements additionnels pour l'appariement. La tâche des personnes affectées au suivi de l'échantillon P a été sensiblement allégée du fait que toutes les questions nécessaires à l'appariement avaient été posées au moment de l'interview initiale. Par conséquent, seuls les cas où il manquait de l'information sur les caractéristiques personnelles, les cas où l'adresse de personnes ayant déménagé était incomplète et les cas d'appariement incertains ont dû faire l'objet d'un suivi. Les cas non appariés de l'échantillon P ont été jugés résolus et n'ont plus été renvoyés à l'équipe chargée du suivi. De nombreuses personnes de l'échantillon D ont pu être appariées à des personnes de l'échantillon P et ces cas ont été résolus sans devoir réaliser une autre interview. Tous les cas de l'échantillon D qui n'ont pu être résolus à l'aide de l'interview de l'échantillon P ont dû faire l'objet d'une interview de rappel qui visait à déterminer leur code de dénombrement. Les opérations sont analysées plus en détail à la section suivante. Les types d'enregistrements erronés que sert à mesurer l'échantillon D comprennent l'erreur de

géocodage, la répétition, l'information fausse, les personnes nées après le jour du recensement, les personnes décédées avant le jour du recensement et les cas non appariables. On dit qu'il y a erreur de géocodage lorsqu'un enregistrement du recensement existe à l'extérieur du secteur de recherche ou de la région d'essai. Les cas non appariables sont des enregistrements sans nom. Ils entraînent une surestimation du nombre d'enregistrements erronés mais sont traités comme ceux-ci dans le calcul de l'estimateur.

2.2 Estimation bivalente

Afin d'estimer la population totale, on utilise un estimateur bivalent qui réunit les données des échantillons P et D. Wolter (1986) décrit divers estimateurs bivalents et les hypothèses sur lesquelles ils reposent. L'estimateur bivalent utilisé dans le TOR est défini

$$\text{EB} = \frac{N_p(\text{REN-SUB-EE})}{\text{M}}, \tag{1}$$

où N_p est l'estimateur de la population totale de l'EP, REN est le chiffre du recensement non redressé, SUB est le nombre de substitutions de personnes dans le recensement, EE est l'estimateur du nombre d'enregistrements erronés et de personnes non appariables dans le recensement selon l'échantillon D et M est l'estimateur du nombre de personnes qui font partie à la fois de la population du recensement et de celle de l'EP. On parle de substitution de personne lorsqu'une personne quelconque a été enregistrée dans le recensement avec moins de deux caractéristiques démographiques. Afin de mieux illustrer quelques-unes des caractéristiques uniques de l'estimateur bivalent, nous exposons dans le tableau ci-dessous la classification des membres de la population.

Les effectifs du tableau 1 sont estimés à l'aide des composantes de l'estimateur bivalent: $\hat{N}_{11} = M$, $\hat{N}_{+1} = N_p$, $\hat{N}_{1+} = \text{REN-SUB-EE}$. La valeur de N_{22} est inobservable par définition mais on peut l'estimer en supposant que le recensement et l'échantillon P de l'EP sont indépendants l'un de l'autre. L'estimation de N_{22} est définie par l'équation suivante:

$$\hat{N}_{22} = \hat{N}_{12}\hat{N}_{21}/\hat{N}_{11}. \tag{2}$$

À l'aide des estimateurs définis ci-dessus, nous estimons la population totale par $\hat{N}_{++} = \text{EB}$.

À cause des problèmes que soulève l'appariement des données de recensement, il faut effectuer des opérations particulières pour ne pas surestimer la population. L'estimateur bivalent suppose que chaque personne est classée dans une seule case du tableau 1. Ainsi, au lieu d'utiliser simplement les chiffres du recensement, nous soustrayons de ces chiffres l'estimation du nombre d'enregistrements erronés pour obtenir une estimation du nombre **réel** de personnes dénombrées dans le recensement. En outre, l'estimateur bivalent suppose que chaque personne peut être désignée comme une concordance ou une non-concordance. Comme les enregistrements du recensement qui ne renferment pas assez d'information pour les besoins de l'appariement (par exemple, absence de nom ou moins de deux caractéristiques démographiques) ne

Tableau 1
Classification bivalente

		Population cible de l'échantillon P		
		Inclus	Exclus	Total
Recensement	Inclus	N_{11}	N_{12}	N_{1+}
Énumeration	Exclus	N_{21}	N_{22}	N_{2+}
	Total	N_{+1}	N_{+2}	N_{++}

Tableau 2
Strates d'échantillonnage et répartition des îlots échantillonnés

Strates d'échantillonnage	Nombre d'îlots échantillonnés
1. Îlots à majorité hispanique constitués de grands immeubles à logements multiples	8
2. Îlots à majorité hispanique constitués de petits immeubles à logements multiples	49
3. Îlots à majorité hispanique constitués de maisons unifamiliales	39
4. Îlots à majorité asiatique	35
5. Autres îlots	38
6. Îlots comptant deux unités de logement ou moins	21

peuvent être classés avec certitude parmi les concordances ou les non-concordances, nous avons aussi soustrait des chiffres du recensement le nombre de personnes non appariables. Toutes les personnes correspondantes de l'échantillon P ont été désignées comme des non-concordances et classées dans la case N_{21}.

2.3 Plan de sondage

Le plan de sondage prévoyait un échantillonnage stratifié où l'îlot servait d'unité d'échantillonnage. Deux catégories de données ont servi à stratifier la région d'essai: le nombre d'unités de logement par îlot, tiré du fichier d'adresses du recensement de 1986, et les résultats de l'application des données du recensement de 1980 sur l'origine ethnique aux unités géographiques du recensement de 1986. Comme cette opération ne pouvait s'effectuer qu'au niveau du secteur de recensement, qui peut contenir de un à six îlots, la répartition des groupes ethniques s'est faite à ce niveau. Tous les îlots d'un secteur de recensement ont été identifiés au même groupe ethnique et se sont trouvés par conséquent dans la même strate.

La région d'essai a été divisée en six strates d'échantillonnage comme l'indique le tableau 2.

Tous les îlots qui comprenaient des lieux particuliers (le plus souvent des foyers collectifs) ont été classés dans une strate d'échantillonnage distincte. Ces îlots étaient considérés comme exclus du champ de l'enquête et n'ont pas été échantillonnés. Les petits îlots ont été classés dans une strate distincte afin de réduire la variance d'échantillonnage. Tous les îlots des secteurs de recensement qui comptaient au moins 18% d'Asiatiques ont servi à former la strate à majorité asiatique. Tous les îlots à majorité non asiatique des secteurs de recensement qui comptaient au moins 40% de personnes d'origine hispanique ont servi à former les trois strates à majorité hispanique. Les îlots qui n'étaient pas inclus dans l'une ou l'autre des strates précédentes ont servi à former la strate réservée aux personnes d'autres origines.

Les données de 1986 sur le logement renfermaient aussi de l'information sur les maisons unifamiliales et les immeubles à logements multiples. Ces données ont servi à partager les strates à majorité hispanique en trois catégories: maisons unifamiliales, petits immeubles à logements multiples et grands immeubles à logements multiples. La strate de la troisième catégorie était composée des îlots à majorité hispanique où au moins la moitié des unités de logement se trouvaient dans des immeubles de dix logements ou plus. La strate de la première catégorie était formée des îlots à majorité hispanique qui comptaient plus de 50% de maisons unifamiliales. Enfin, la strate de la seconde catégorie comprenait tous les autres îlots à majorité hispanique.

Des îlots ont été prélevés dans chacune des strates au moyen d'un échantillonnage systématique avec probabilités égales. L'échantillon était composé de 190 îlots renfermant environ

6000 unités de logement. Le tableau 2 donne la répartition des îlots échantillonnés selon les strates. Les grands îlots qui comptent 70 unités de logement ou plus ont fait l'objet d'un sous-échantillonnage afin de réduire la tâche des intervieweurs. Le sous-échantillonnage consistait à diviser l'îlot en grappes de 35 à 50 unités de logement à l'aide de tranches d'adresses ou de côtés d'îlot. On choisissait une grappe au hasard dans le but d'interviewer des membres de l'échantillon P. Quant à l'échantillon D, il comprenait toutes les personnes qui faisaient partie de cette grappe le jour du recensement.

2.4 Stratification a posteriori

L'estimateur bivalent est biaisé et ce biais peut être élevé s'il y a de fortes différences de taux de sous-dénombrement entre les sous-groupes de la population (Wolter 1986). Pour réduire au minimum ce biais, on a stratifié a posteriori la région d'essai selon des critères précis de manière à ce que les taux de sous-dénombrement soient comparables dans toutes les strates. On a ensuite calculé des estimations bivalentes dans chaque strate ainsi formée.

Les strates ont été choisies par suite d'un examen de la structure de la région d'essai et d'une analyse des données de l'EP de 1980. On a observé que l'origine ethnique était la variable qui influait le plus sur le taux de sous-dénombrement. On a donc défini trois groupes ethniques: les personnes d'origine hispanique, les personnes d'origine asiatique et les autres. Il n'a pas été possible de créer une strate pour les personnes de race noire puisque la région d'essai comptait très peu de noirs. Le fait d'être locataire constituait une variable explicative importante dans notre étude précédente (Isaki et coll. 1987). Par conséquent, le mode d'occupation a aussi été considéré dans la construction des strates. Par ailleurs, comme on croyait que les personnes d'origine hispanique vivant dans des îlots qui comptaient moins de 50% de personnes d'origine hispanique (appelés îlots à minorité hispanique) n'avaient pas le même taux de sous-dénombrement que les autres personnes de même origine, on les a incluses dans une strate à

Tableau 3

Catégories "origine ethnique — mode d'occupation" utilisées dans la stratification a posteriori, y compris la taille des échantillons

Catégories "origine ethnique — mode d'occupation"	Échantillon P	Échantillon D
Locataires d'origine hispanique vivant dans des îlots à majorité hispanique	8,182	8,739
Propriétaires d'origine hispanique vivant dans des îlots à majorité hispanique	5,688	5,867
Personnes d'origine hispanique vivant dans des îlots à minorité hispanique	896	1,005
Locataires d'origine asiatique	666	911
Propriétaires d'origine asiatique	1,144	1,230
Autres locataires	1,135	1,316
Autres propriétaires	1,841	1,908
Total	19,552	20,976

part. Le tableau 3 donne les sept catégories "origine ethnique — mode d'occupation" qui ont été subdivisées selon l'âge (0-14, 15-29, 30-44, 45-64, 65 +) et le sexe pour former les 70 strates qui ont servi à l'estimation.

Le tableau 3 indique également la taille des échantillons P et D. La taille de l'échantillon P est moins élevée que celle de l'échantillon D à cause notamment des membres de l'échantillon P qui sont venus s'établir dans la région d'essai après le recensement et qui sont exclus par conséquent du champ de l'enquête.

2.5 Traitement des données manquantes

Le calcul des estimations bivalentes exige un fichier de données complet. Or, comme dans le cas de n'importe quelle enquête par sondage, le test de 1986 n'a pas permis de recueillir toutes les données voulues. Schenker (1988) décrit les méthodes utilisées pour résoudre le problème des données manquantes et analyse notamment les effets de diverses hypothèses concernant les données manquantes sur les estimations bivalentes. Afin d'avoir une idée d'ensemble de la question, nous décrivons brièvement ici ces méthodes.

Dans le test de 1986, il manquait des données sur les caractéristiques des personnes et du ménage, le code d'appariement (apparié/non-apparié) pour les personnes de l'échantillon P et le code de dénombrement (exact/erroné) pour les personnes de l'échantillon D. Pour les cas de non-interview de l'échantillon P, on a eu recours à un redressement par pondération. L'imputation des caractéristiques manquantes s'est faite à l'aide d'une méthode "hot-deck".En ce qui concerne les codes d'appariement, on s'est servi d'un modèle de régression logistique pour estimer la probabilité d'appariement. Au lieu d'attribuer un code d'appariement définitif à chaque cas non résolu, on s'est servi des probabilités estimées dans le calcul des estimations bivalentes. On a procédé de façon analogue pour les codes de dénombrement manquants de l'échantillon D.

2.6 Estimation relative aux petites régions

Afin que les utilisateurs disposent de données redressées à tous les niveaux d'agrégation, on calcule des estimations du taux de sous-dénombrement jusqu'au niveau de l'îlot, qui est la plus petite unité géographique. Avant d'effectuer ces calculs toutefois, on utilise un modèle de régression pour "lisser" les effets de l'erreur d'échantillonnage. Dans ce modèle, des facteurs de redressement servent de variable dépendante. Un facteur de redressement est défini comme le quotient de l'estimateur bivalent par le chiffre du recensement:

$$Y = \text{EB/REN}, \tag{3}$$

où REN et EB sont définis comme auparavant.

Le modèle de régression s'exprime par le formule suivante:

$$Y_i = B_o + B_1 X_{i1} + \ldots + B_p X_{ip} + S_i + E_i, \tag{4}$$

où Y_i est le facteur de redressement pour la i-ième strate formée a postériori ($i = 1, \ldots, 70$), X_{ij} = est la variable indépendante ($j = 1, \ldots, p$), B_j est le coefficient de régression à estimer, $S_i$est l'erreur d'échantillonnage du facteur de redressement, E_i est l'erreur du modèle et les valeurs de S_i et de E_i sont indépendantes et distribuées selon une loi normale avec une moyenne nulle et des variances égales à σ_i^2et à ϵ^2 respectivement. Les valeurs d'ϵ^2 et de B_j sont estimées à l'aide des méthodes du maximum de vraisemblance (Ericksen et Kadane 1985). Les valeurs de σ_i^2 sont estimées directement à l'aide de l'échantillon. On fait la moyenne du facteur de

redressement fondé sur l'échantillon et du facteur de redressement fondé sur le modèle pour obtenir le facteur de redressement prévu

$$AF_i = \left(Y_i/\sigma_i^2 + \sum_j X_{ij}\hat{B}_j/\epsilon^2\right)\left(\sigma_i^{-2} + \epsilon^{-2}\right)^{-1}, \tag{5}$$

qui sert à corriger les données de recensement relatives à l'îlot. La variance de AF_i peut être déduite des résultats publiés dans Freedman et Navidi (1986).

L'estimation synthétique a servi à reproduire le redressement effectué dans chaque strate au niveau de l'îlot. L'estimateur synthétique est défini

$$ADJ_{ij} = AF_i \times CEBN_{ij}, \tag{6}$$

où i et j désignent respectivement la strate formée a posteriori et l'îlot et ADJ est le chiffre de population redressé pour l'îlot.

Le chiffre de population redressé ADJ_{ij} est en règle générale un nombre fractionnaire. Or, le recensement sert à dénombrer des unités. Par conséquent, si nous voulons établir une compatibilité entre les opérations de redressement et le recensement, il est nécessaire de convertir les nombres fractionnaires en nombres entiers. Cette conversion (appelée arrondissement contrôlé) permet d'arrondir vers le haut ou vers le bas tous les nombres fractionnaires (Causey et coll. 1985).

Après avoir arrondi les estimations redressées des îlots, on a produit des chiffres selon l'âge, l'origine ethnique et le sexe qui indiquaient le nombre de personnes qu'il fallait ajouter ou retrancher dans chaque îlot. Dans le cas d'un sous-dénombrement, on prélevait aléatoirement dans l'îlot un enregistrement du recensement qui présentait la même série de caractéristiques que la personne oubliée et on en tirait un nouvel enregistrement identique. On a utilisé une catégorie n'ayant pas trait aux ménages pour ajouter des personnes dans les listes du recensement de manière qu'il ne soit pas nécessaire de tenir compte des rapports entre les membres du ménage ou de créer de nouveaux ménages. Zaslavasky (1988) présente une autre méthode, fondée sur la pondération, pour ajouter des personnes et des ménages dans les îlots de recensement. Dans le cas d'un surdénombrement, on signalerait les enregistrements qui présentent les caractéristiques voulues et on les incluerait dans les totalisations corrigées du recensement.

3. OPÉRATIONS ET CALENDRIER DES OPÉRATIONS

Le recensement d'essai avait principalement pour but d'analyser le calendrier et les caractéristiques des opérations de redressement des données du recensement. Les EP qu'avait réalisées le Census Bureau dans les années antérieures s'étaient étendues sur au moins deux ans. Par exemple, l'EP de 1980 a produit des estimations du taux de sous-dénombrement à l'automne de 1981 puis une série d'estimations finales au début de 1982.

Le tableau 4 énumère les principales opérations du recensement et de l'EP ainsi que les dates qui marquent le début et la fin de chacune de ces opérations. On remarquera qu'il s'écoule un certain temps entre la fin d'une opération et le début d'une autre car les opérations du recensement et de l'EP ne sont pas toutes indiquées. Certaines périodes se chevauchent car les opérations correspondantes se déroulaient simultanément. On a amorcé l'EP une fois que les principales opérations régionales touchant le recensement ont été terminées, cela ayant pour but d'éviter que le personnel des bureaux régionaux n'ait à exécuter simultanément les opérations des deux enquêtes.

Tableau 4
Calendrier des opérations pour le TOR de 1986

Opération	Début	Fin
Jour du recensement	16 mars	16 mars
Suivi des cas de non-réponse	9 avril	8 mai
Saisie des noms du recensement	23 mai	10 juin
Fichier du recensement pour l'appariement	8 août	15 août
Liste d'adresses pour l'EP	17 juin	21 juin
Sous-échantillonnage pour l'EP	25 juin	1 juillet
Période d'interview pour l'EP	25 juin	8 août
Saisie des données de l'EP	21 juillet	19 août
Appariement informatisé	28 août	9 septembre
Appariement informatisé (deuxième phase)	9 septembre	3 octobre
Appariement manuel	15 septembre	31 octobre
Suivi sur place	23 septembre	6 novembre
Appariement des résultats du suivi	29 septembre	6 novembre
Saisie des résultats de l'appariement	21 octobre	10 novembre
Préparation des fichiers des échantillons P et D	11 novembre	2 janvier
Imputation	5 janvier	11 janvier
Version finale du fichier du recensement	—	5 janvier
Estimation relative aux strates formées à posteriori	12 janvier	11 février
Estimation relative aux petites régions	12 février	22 février

Le recensement a été réalisé à l'aide d'un questionnaire qui a été posté à chaque unité de logement connue et qu'un membre du ménage devait remplir le jour du recensement (16 mars). Les ménages qui négligeaient de retourner leur questionnaire recevaient la visite d'un agent recenseur qui procédait à une interview sur place. C'est ce qu'on appelle le suivi des cas de non-réponse. Une fois remplis, les questionnaires étaient acheminés au service de traitement pour la saisie des données, qui comprenaient en l'occurrence tous les noms du recensement.

La première étape de l'EP consistait à dresser une liste indépendante de toutes les adresses figurant dans les îlots échantillonnés. On a confronté la liste en question avec une liste administrative pour s'assurer de son exactitude et de son intégralité. Cette mesure de contrôle de la qualité a permis de constater que 127 îlots (67%) ne requéraient aucune modification à la liste d'adresses. Les 63 autres îlots (33%) ont dû faire l'objet de modifications par l'intermédiaire du contrôle de qualité puis les adresses correspondantes ont été relistées. Cette opération a permis d'ajouter des adresses dans 37 îlots, d'en corriger dans 39 îlots et d'en soustraire dans 9 îlots. (Comme certains îlots ont fait l'objet de plus d'un genre de modifications, la somme des nombres ci-dessus ne correspond pas au nombre total d'îlots qui ont dû faire l'objet de modifications.)La mesure de contrôle de la qualité n'a entraîné que des corrections mineures pour les autres adresses. À l'issue de cette opération de contrôle, tous les îlots comptant 70 unités de logement ou plus ont été divisés par sous-échantillonnage à l'aide des côtés d'îlots ou des tranches d'adresses.

L'interview de l'EP a été réalisée sur place. Elle visait à recueillir des renseignements sur les caractéristiques démographiques de toutes les personnes qui demeuraient à l'endroit visité le jour de l'enquête. L'interview comportait des questions particulières comme celles portant sur le domicile le jour du recensement, l'adresse postale, d'autres lieux d'habitation comme les foyers d'étudiants, et les autres personnes qui pouvaient avoir demeuré à l'endroit visité le jour du recensement. Une mesure de contrôle de la qualité du questionnaire de l'EP a permis de vérifier la liste des noms. Quatre-vingt-seize pour cent des questionnaires vérifiés ont passé le test. Pour ce qui est des autres (4%), on a dû procéder à une nouvelle interview et corriger les questionnaires en conséquence.

À la fin de la période d'interview, 5,714 (93.2%) unités de logement avaient fait l'objet d'une interview complète avec un membre du ménage, 193 (3.1%) avaient été classées parmi les logements inoccupés et 189 (3.1%) avaient fait l'objet d'une interview complète avec une personne qui n'était pas membre du ménage (par exemple, un voisin). Seulement 32 (0.5%) unités de logement ont été classées parmi les cas de non-interview. Ce taux extrêmement faible s'explique par le fait que l'on avait alloué cinq semaines pour les interviews.

Une fois remplis, les questionnaires de l'EP ont été préparés en vue de l'appariement informatisé avec le fichier du recensement. L'appariement informatisé a été divisé en deux phases: la première visait à apparier les données de l'EP avec celles de l'échantillon D et la seconde visait à apparier tous les cas de l'échantillon P qui n'avaient pu l'être dans la première phase avec les autres données du recensement. La seconde phase de l'appariement informatisé servait aussi à apparier les personnes ayant déménagé entre le jour du recensement et le jour de l'interview pour l'EP avec les erreurs de géocodage (c.-à-d., lorsque l'unité de logement est associée au mauvais îlot). La première phase de l'appariement informatisé a permis de déterminer une concordance pour 14,700 cas (73.5%) de l'échantillon P et une concordance probable pour 2,550 autres cas (12.0%). La seconde phase de l'appariement informatisé a permis de déterminer une concordance pour 130 personnes (0.7%) et une concordance probable pour 570 autres personnes (2.9%). Étant donné le faible pourcentage de cas de l'échantillon P pour lesquels la seconde phase de l'appariement a permis de déterminer une concordance, nous en avons conclu que le géocodage dans Los Angeles comportait peu d'erreurs.

L'appariement manuel a servi à revoir les résultats de l'appariement informatisé. Il a servi aussi à identifier les cas qui n'avaient pu être appariés à cause d'un manque de données (pour ces cas, il faut recourir à l'imputation). Enfin, c'est durant l'appariement manuel qu'ont été rédigés les questionnaires de suivi pour les cas non résolus des échantillons P et D.

Le suivi sur place visait 1,551 unités de logement, dont 1,511 (97.4%) avaient fait l'objet d'une interview complète. Cette opération a été suivie d'un appariement final. Les résultats définitifs de l'échantillon P indiquent que 17,018 (85.2%) personnes ont pu être appariées à un enregistrement du recensement tandis que 2,373 (11.9%) n'ont pu l'être. En outre, 426 (2.1%) personnes ont été considérées comme exclues du champ de l'enquête (il s'agissait pour la plupart de personnes qui demeuraient à l'extérieur de la région d'essai le jour du recensement) et 161 (0.8%) personnes représentaient des cas non résolus (auxquels on a imputé ultérieurement un code d'appariement). Les résultats définitifs de l'échantillon D indiquent que 19,637 (93.6%) personnes ont été enregistrées correctement dans le recensement tandis que 360 (1.7%) l'ont été incorrectement. De plus, 976 (4.7%) personnes représentaient des cas non résolus auxquels on a imputé un code de dénombrement.

Après l'appariement final, toutes les données manquantes, y compris les codes d'appariement pour l'échantillon P et les codes de dénombrement pour l'échantillon D, ont été imputées. Les résultats ont servi à calculer les estimations bivalentes. Les estimations ont été lissées puis reproduites au niveau de l'îlot pour créer un fichier du recensement redressé. La réduction du délai de production des estimations du taux de sous-dénombrement a été surtout attribuable aux opérations d'appariement. Pour le TOR, l'appariement informatisé et l'appariement manuel se sont étendus sur environ trois mois tandis que les opérations d'appariement pour l'EP de 1980 s'étaient déroulées sur plus d'un an. Une meilleure planification des opérations ainsi qu'une facilité accrue pour consulter les documents du recensement sont deux facteurs qui ont aussi contribué à réduire les délais de production.

4. ESTIMATIONS

4.1 Estimations relatives aux strates formées a posteriori

Cette sous-section présente les estimations du taux de sous-dénombrement pour divers groupes de strates formées a posteriori. Le tableau 5 donne le taux de sous-dénombrement (100(1-REN/EB)), le pourcentage de cas non appariés (100(1-M/N_p)), le pourcentage de cas

Tableau 5

Taux de sous-dénombrement (en pourcentage) et composantes de l'estimateur bivalent pour les strates formées a posteriori

Strates formées a posteriori	Taux de sous-dénombrement[a]	Pourcentage de cas de l'échantillon P non appariés	Pourcentage de cas de l'échantillon D enregistrés incorrectement[b]	Pourcentage de cas ayant fait l'objet d'une substitution dans le recensement
Locataires d'origine hispanique vivant dans des îlots à majorité hispanique	13.7	17.1	2.6	1.7
Propriétaires d'origine hispanique vivant dans des îlots à majorité hispanique	5.5	8.1	1.2	1.5
Personnes d'origine hispanique vivant dans des îlots à minorité hispanique	7.5	9.7	1.4	1.3
Locataires d'origine asiatique	11.1	13.4	2.1	1.2
Propriétaires d'origine asiatique	4.6	6.8	1.2	1.5
Autres locataires	9.9	12.9	2.4	1.7
Autres propriétaires	3.8	5.8	1.3	0.9
0-14	8.8	11.9	2.2	1.6
15-29	13.6	16.2	2.1	1.6
30-44	8.6	10.8	1.4	1.4
45-64	4.5	6.6	1.3	1.4
65+	3.3	5.9	1.7	1.3
Homme	9.7	12.1	1.7	1.5
Femme	8.3	10.8	1.9	1.5
Total	9.0	11.4	1.8	1.5

[a] La sommation de toutes les estimations est étendue à toutes les autres catégories.
[b] Les cas enregistrés incorrectement comprennent les cas non appariables qui n'ont pas fait l'objet d'une substitution.

enregistrés incorrectement (100(EE/REN)) et le pourcentage de cas ayant fait l'objet d'une substitution (100(SUB/REN)).

Une caractéristique de l'estimateur bivalent veut que la somme des estimations pour plusieurs catégories n'égale pas l'estimation directe de la somme pour l'ensemble de ces catégories. Ainsi, pour assurer la cohérence des estimations indiquées dans le tableau 5, on a étendu la sommation de toutes les estimations aux autres catégories pertinentes.

Si nous examinons les taux de sous-dénombrement indiqués dans le tableau 5 pour les catégories "origine ethnique-mode d'occupation", nous en concluons que le mode d'occupation est une bonne variable de stratification, les locataires ayant un taux de sous-dénombrement plus élevé que les propriétaires; l'origine ethnique influe aussi sur le taux de sous-dénombrement, les personnes d'origine hispanique ayant un taux plus élevé que les personnes d'origine

asiatique, qui à leur tour ont un taux plus élevé que les autres. En ce qui concerne les enregistrements erronés, le pourcentage de cas est plus élevé pour les locataires que pour les propriétaires alors que l'origine ethnique n'est pas une variable déterminante sur ce plan. Pour ce qui a trait aux cas ayant fait l'objet d'une substitution, le pourcentage est plus élevé pour les locataires d'origine hispanique et les autres locataires que pour les propriétaires d'origine hispanique et les autres propriétaires. En revanche, on observe la situation inverse dans le cas des Asiatiques.

Si nous examinons maintenant les taux de sous-dénombrement selon l'âge et le sexe, nous constatons que le groupe d'âge 15-29 a le taux le plus élevé et que les hommes ont un taux plus élevé que les femmes. Les groupes d'âge 0-14 et 30-44 présentent des taux de sous-dénombrement comparables, lesquels sont légèrement au-dessous de la moyenne pour la région d'essai. Les groupes d'âge 45-64 et 65+ présentent aussi des taux comparables, qui sont bien inférieurs à ceux des autres groupes d'âge. Au point de vue de la distribution, ces résultats correspondent assez bien à ceux observés antérieurement.

Le pourcentage de cas enregistrés incorrectement est le plus élevé pour les deux premiers groupes d'âge (0-14 et 15-29). Les deux groupes suivants (30-44 et 45-64) présentent les pourcentages les plus faibles. Curieusement, le groupe d'âge 65+ présente un pourcentage qui se situe entre les pourcentages des groupes précités. Le pourcentage de cas enregistrés incorrectement varie peu d'un sexe à l'autre. Enfin, pour ce qui a trait aux cas ayant fait l'objet d'une substitution, les différences de pourcentage entre les groupes d'âge et entre les sexes sont faibles.

4.2 Estimations relatives aux petites régions

Comme il a été mentionné précédemment, avant d'opérer un redressement des données au niveau de l'îlot, nous avons ajusté un modèle de régression dans le but de "lisser" les données et de réduire les effets de la variabilité d'échantillonnage. Le modèle de régression a été ajusté aux 70 facteurs de redressement qui avaient été définis par les strates formées a posteriori. Ce genre de modèle sert à déterminer un mode de calcul du taux de sous-dénombrement uniforme pour les régions ayant les mêmes caractéristiques. Dans un deuxième temps, on rajuste les facteurs de redressement fondés sur l'échantillon de manière à les rendre comparables aux estimations du modèle. Ce procédé s'inspire de l'estimateur de James-Stein et des estimateurs empiriques de Bayes. Les variables indépendantes qui pouvaient être utilisées en l'occurrence dans le modèle étaient des variables indicatrices pour les catégories "origine ethnique-mode d'occupation", les groupes d'âge et les sexes. Il n'était pas permis d'inclure des termes d'interaction dans le modèle. Le modèle qui pouvait être ajusté aux données et qui présentait des coefficients significatifs (selon un modèle de régression non pondéré) était le suivant:

$$\hat{Y} = 1.038 + .090(\text{HR}) + .044(\text{AR}) + .013(\text{OR}) + .058(\text{A15-29}) - .009(\text{A45-64})$$

où $\hat{Y}$ = facteur de redressement axé sur le modèle

LH = 1 si locataire d'origine hispanique vivant dans un îlot à majorité hispanique
= 0 dans les autres cas

LAS = 1 si locataire d'origine asiatique vivant dans un îlot quelconque
= 0 dans les autres cas

AL = 1 si autre locataire vivant dans un îlot quelconque
= 0 dans les autres cas

A15-29 = 1 si groupe d'âge 15-29
= 0 dans le cas contraire

A45-64 = 1 si groupe d'âge 45-64
= 0 dans le cas contraire.

Tableau 6

Résultats du lissage des facteurs de redressement du TOR

Strates formées a posteriori	Sexe/Âge	Estimé		Prévu	
		Facteur de redressement (Y)	Erreur type	Facteur de redressement (AF)	Erreur type
LH dans IH	H' 0-14	1.131	0.020	1.130	0.016
LH dans IH	H' 15-29	1.247	0.030	1.211	0.021
LH dans IH	H' 30-44	1.165	0.029	1.144	0.020
LH dans IH	H' 45-64	1.099	0.043	1.114	0.024
LH dans IH	H' 65+	1.055	0.044	1.110	0.023
LH dans IH	F 0-14	1.124	0.023	1.126	0.018
LH dans IH	F 15-29	1.234	0.032	1.203	0.022
LH dans IH	F 30-44	1.084	0.017	1.098	0.015
LH dans IH	F 45-64	1.125	0.040	1.121	0.024
LH dans IH	F 65+	1.099	0.045	1.122	0.024
PH dans IH	H' 0-14	1.056	0.018	1.050	0.015
PH dans IH	H' 15-29	1.078	0.018	1.084	0.015
PH dans IH	H' 30-44	1.087	0.016	1.072	0.014
PH dans IH	H' 45-64	1.031	0.012	1.031	0.011
PH dans IH	H' 65+	1.073	0.028	1.054	0.019
PH dans IH	F 0-14	1.059	0.020	1.051	0.016
PH dans IH	F 15-29	1.088	0.016	1.090	0.014
PH dans IH	F 30-44	1.033	0.012	1.034	0.011
PH dans IH	F 45-64	1.020	0.012	1.022	0.011
PH dans IH	F 65+	1.033	0.019	1.035	0.015
H dans INH	H' 0-14	1.105	0.052	1.051	0.023
H dans INH	H' 15-29	1.154	0.054	1.106	0.025
H dans INH	H' 30-44	1.131	0.065	1.050	0.024
H dans INH	H' 45-64	1.063	0.050	1.036	0.023
H dans INH	H' 65+	0.991	0.000	0.991	0.000
H dans INH	F 0-14	1.137	0.047	1.059	0.023
H dans INH	F 15-29	1.033	0.022	1.060	0.017
H dans INH	F 30-44	1.079	0.037	1.051	0.021
H dans INH	F 45-64	1.033	0.028	1.031	0.019
H dans INH	F 65+	0.947	0.040	1.013	0.022
LAS dans tous I	H' 0-14	1.059	0.041	1.076	0.026
LAS dans tous I	H' 15-29	1.127	0.044	1.137	0.028
LAS dans tous I	H' 30-44	1.195	0.077	1.093	0.031
LAS dans tous I	H' 45-64	1.004	0.057	1.063	0.030
LAS dans tous I	H' 65+	0.982	0.000	0.982	0.000
LAS dans tous I	F 0-14	1.067	0.047	1.079	0.028
LAS dans tous I	F 15-29	1.215	0.055	1.153	0.029
LAS dans tous I	F 30-44	1.173	0.105	1.087	0.032
LAS dans tous I	F 45-64	1.012	0.061	1.065	0.030
LAS dans tous I	F 65+	1.212	0.127	1.087	0.032

Tableau 6

Résultats du lissage des facteurs de redressement du TOR — Fin

Strates formées a posteriori	Sexe/Âge	Estimé		Prévu	
		Facteur de redressement (*Y*)	Erreur type	Facteur de redressement (*AF*)	Erreur type
PAS dans tous I	H' 0-14	1.045	0.030	1.041	0.019
PAS dans tous I	H' 15-29	1.059	0.038	1.085	0.022
PAS dans tous I	H' 30-44	1.091	0.040	1.053	0.022
PAS dans tous I	H' 45-64	1.035	0.020	1.033	0.016
PAS dans tous I	H' 65+	1.031	0.051	1.037	0.023
PAS dans tous I	F 0-14	1.040	0.041	1.039	0.022
PAS dans tous I	F 15-29	1.052	0.046	1.086	0.024
PAS dans tous I	F 30-44	1.035	0.036	1.037	0.021
PAS dans tous I	F 45-64	1.038	0.019	1.035	0.015
PAS dans tous I	F 65+	1.051	0.045	1.041	0.022
AL dans tous I	H' 0-14	1.037	0.059	1.049	0.027
AL dans tous I	H' 15-29	1.252	0.114	1.115	0.031
AL dans tous I	H' 30-44	1.144	0.066	1.062	0.028
AL dans tous I	H' 45-64	1.055	0.031	1.047	0.022
AL dans tous I	H' 65+	1.068	0.056	1.054	0.027
AL dans tous I	F 0-14	1.148	0.062	1.064	0.027
AL dans tous I	F 15-29	1.126	0.054	1.112	0.028
AL dans tous I	F 30-44	1.134	0.057	1.064	0.027
AL dans tous I	F 45-64	1.068	0.041	1.049	0.025
AL dans tous I	F 65+	0.948	0.021	0.992	0.018
AP dans tous I	H' 0-14	1.044	0.037	1.040	0.021
AP dans tous I	H' 15-29	1.148	0.064	1.103	0.025
AP dans tous I	H' 30-44	1.006	0.048	1.032	0.023
AP dans tous I	H' 45-64	1.036	0.017	1.034	0.014
AP dans tous I	H' 65+	1.017	0.019	1.025	0.016
AP dans tous I	F 0-14	1.159	0.068	1.052	0.024
AP dans tous I	F 15-29	1.081	0.042	1.092	0.023
AP dans tous I	F 30-44	0.997	0.017	1.011	0.014
AP dans tous I	F 45-64	1.025	0.012	1.026	0.011
AP dans tous I	F 65+	0.997	0.012	1.004	0.011

Nota: H: hispanique, L: locataire, I: îlot, H': Homme, F: femme, P: propriétaire, A: autre, NH: non hispanique, AS: asiatique.

(Exemple: LH dans IH: locataire d'origine hispanique vivant dans un îlot à majorité hispanique)

Le modèle de régression indique des taux de sous-dénombrement plus élevés pour les locataires que pour les propriétaires, peu importe l'origine ethnique. De plus, le groupe d'âge 15-29 présente un taux de sous-dénombrement beaucoup plus élevé que les autres groupes. Inversement, le groupe d'âge 45-64 est celui qui présente le taux le plus faible. La variable "sexe" n'étant pas statistiquement significative, elle n'a pas été incluse dans le modèle. Deux facteurs de redressement (hommes d'origine hispanique âgés de 65 et plus et vivant dans des îlots non hispaniques et locataires asiatiques de sexe masculin âgés de 65 et plus) avaient une variance estimée nulle et n'ont donc pas été inclus dans le modèle. Pour ces deux facteurs de redressement, le facteur fondé sur l'échantillon a tenu lieu de facteur prévu.

Le tableau 6 renferme les facteurs de redressement fondés sur l'échantillon (facteurs estimés) et les facteurs de redressement prévus pour les 70 strates formées a posteriori. En règle générale, les facteurs de redressement prévus sont inférieurs aux facteurs estimés les plus élevés mais supérieurs aux facteurs estimés les plus faibles. Les facteurs de redressement prévus ont une moins grande variabilité que les facteurs de redressement fondés sur l'échantillon. Les effets du modèle de régression sont les plus visibles dans le cas des locataires asiatiques de sexe féminin âgées de 65 et plus. Le facteur de redressement prévu pour cette strate est de 1.087 comparativement à un facteur de redressement estimé de 1.212. Dans ce cas, le facteur de redressement prévu est plus conforme à la représentation que l'on se fait du taux de sous-dénombrement pour les personnes de 65 ans et plus, c'est-à-dire un taux moins élevé que pour les autres groupes d'âge.

On a multiplié les chiffres du recensement par les facteurs de redressement prévus pour les 2 405 îlots de la région d'essai. On a ensuite arrondi les chiffres redressés pour obtenir des nombres entiers. Malgré que trois facteurs de redressement prévus aient été inférieurs à 1 (surdénombrement estimé), l'arrondissement n'a pas produit d'estimations gonflées.

Les opérations de redressement ont permis d'ajouter 32,843 noms à la liste des personnes recensées (soit un taux de sous-dénombrement de 8.2%). Si l'on avait utilisé les facteurs de redressement fondés sur l'échantillon, c'est 36,454 noms que l'on aurait ajouté à la liste des personnes recensées (soit un taux de sous-dénombrement de 9.0%). Le lissage a donc réduit de près de 10% le niveau de sous-dénombrement estimé. Cette forte réduction s'explique par le fait que les groupes qui présentaient les estimations de sous-dénombrement les plus élevées avant le lissage étaient ceux qui comptaient le plus grand nombre de personnes.

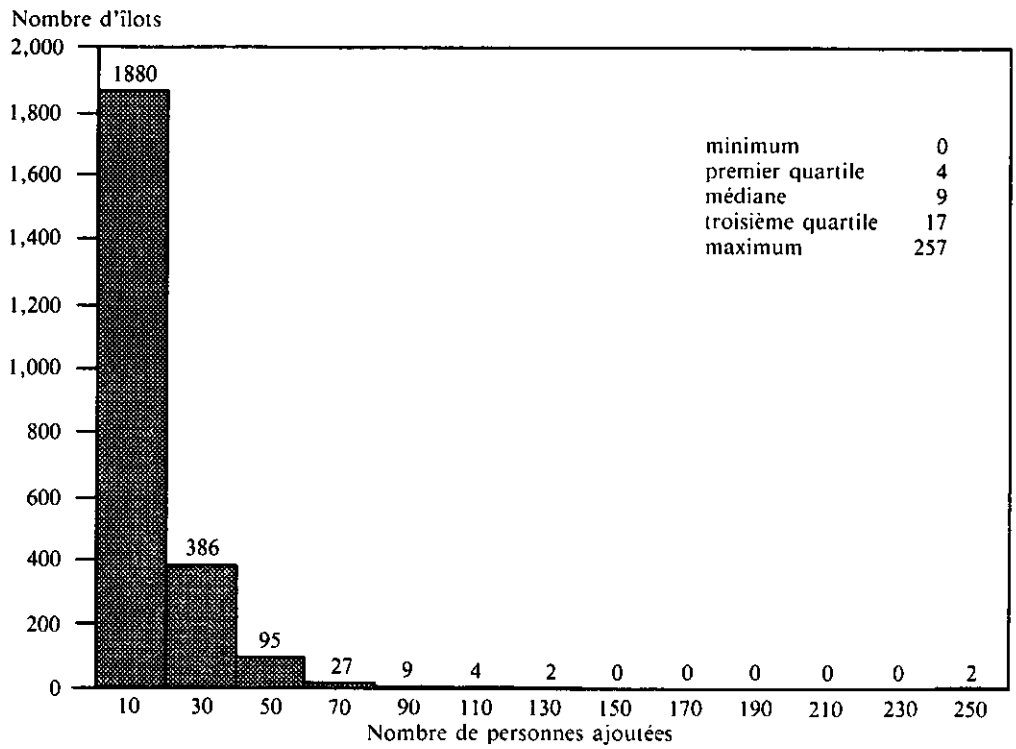

Figure 1. Histogramme et données sommaires concernant le nombre de personnes ajoutées par îlot

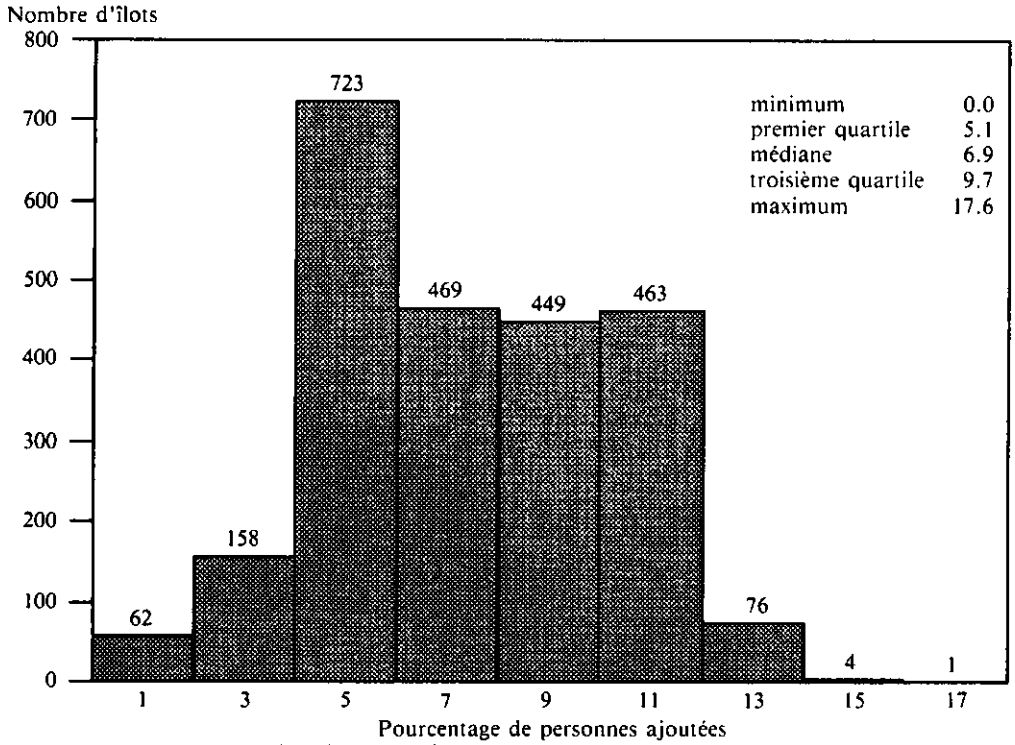

Figure 2. Histogramme et données sommaires concernant le pourcentage de personnes ajoutées par îlot

Pour résumer les opérations de redressement au niveau de l'îlot, nous représentons dans la figure 1 le nombre de personnes qui ont été ajoutées selon le nombre d'îlots concernés et dans la figure 2, le pourcentage de personnes ajoutées selon le nombre d'îlots concernés. Près de 80% des îlots ont reçu moins de 20 personnes. Seulement deux îlots ont reçu plus de 150 personnes. Ces deux îlots étaient assez denses puisqu'ils comptaient chacun environ 2,000 personnes. Plus de 80% des îlots avaient un taux de sous-dénombrement estimé variant de 4% à 12%. Nombre de petits îlots ont reçu un faible pourcentage des personnes additionnelles parce que les estimations avaient été arrondies vers le bas et que cela s'était traduit par une forte variation du pourcentage. Les îlots qui ont reçu le plus fort pourcentage de personnes additionnelles étaient ceux qui comptaient surtout des locataires d'origine hispanique; du reste, ces îlots étaient ceux pour lesquels les facteurs de redressement prévus étaient les plus élevés.

5. CONCLUSION

Cet article nous a permis d'étudier les méthodes, les opérations et les résultats du Test des opérations de redressement (TOR). Ce test visait à vérifier le calendrier et les caractéristiques des opérations de redressement des données du recensement eu égard à la population non dénombrée.

Les résultats du TOR prouvent qu'il est possible de produire des estimations du taux de sous-dénombrement dans des délais raisonnables. En outre, le TOR a été réalisé plus rapidement que n'importe quelle enquête post-censitaire antérieure.

Les résultats du TOR indiquent un taux de sous-dénombrement de 9% pour le recensement d'essai du Central Los Angeles County. Nous avons établi des estimations bivalentes pour 70 catégories formées en fonction de l'origine ethnique et du mode d'occupation, de l'âge et du

sexe. Nous avons lissé ces estimations en leur ajustant un modèle de régression puis nous avons reproduit les estimations ainsi obtenues au niveau de l'îlot. L'utilisation d'un taux de sous-dénombrement estimé pour l'îlot permet l'agrégation de données à n'importe quel niveau au-dessus de l'îlot.

Schenker (1988) et Hogan et Wolter (1988) font une évaluation des opérations et définissent les hypothèses sur lesquelles reposent les estimateurs. Ces deux ouvrages ainsi que le présent article constituent une analyse approfondie des chiffres du recensement et des taux de sous-dénombrement estimés pour le recensement d'essai.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance aux personnes qui ont collaboré à cette étude: Dan Childers, Howard Hogan, Cary Isaki, Matthew Jaro, Jan Jaworski, Charisse Jeffries, Robert O'Brien, Arona Pistiner, Nathaniel Schenker, Maria Urrutia, Debbie Wagner et Kirk Wolter. L'auteur tient aussi à remercier l'arbitre, dont les commentaires ont permis d'améliorer le contenu de cet article.

BIBLIOGRAPHIE

ANOLIK, IRWIN (1988). The 1986 rural post-enumeration survey in East Central Mississippi. Statistical Research Division Report Series. CENSUS/SRD/RR-87/09.

CAUSEY, B.D., COX, L.H., et ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

CITRO, CONSTANCE F., et COHEN, MICHAEL L. (1985). *The Bicentennial Census — New Directions for Methodology in 1990*. Washington: National Academy Press.

ERICKSEN, EUGENE P., et KADANE, JOSEPH B. (1985). Estimating the population in a census year — 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-109.

FAY, R.E., PASSEL, J.S., et ROBINSON, J.G. (1988). The coverage of population in the 1980 census. 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.

FREEDMAN, D. et NAVIDI, W. (1986). Models for adjusting the census. *Statistical Science*, 1, 3-11.

HOGAN, HOWARD R. (1984). Research plan on adjustment. *Proceedings of the Social Statistics Section American Statistical Association*, 452-457.

HOGAN, HOWARD R., et WOLTER, KIRK M. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14.

ISAKI, CARY T., SCHULTZ, LINDA K., SMITH, PHILIP J., et DIFFENDAL, GREGG J. (1987). Small area estimation research for census undercount-progress report. Dans *Small Area Statistics: An International Symposium*, 219-238. New York: John Wiley and Sons.

JARO, MATTHEW, et CHILDERS, DANNY (1986). Matching the 1985 census of Tampa. Article présenté à la conférence annuelle de l'American Statistical Association. Chicago.

MARKS, ELI S. (1978). The use of dual system estimation in census evaluation. Dans *Developments in Dual System Estimation*, (éd. Karol Krotki), Edmonton: University of Alberta Press.

MULRY, M.H., HOGAN, H.R., WALKER, J.R., CHAPMAN, D.R., EVAUL, J., et MOORE, R.H. (1981). Research proposal for a study of methods for 1990 decennial census coverage evaluation. Rapport technique, U.S., Bureau of the Census, Washington, D.C.

SCHENKER, NATHANIEL (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986. *Techniques d'enquête*, 14.

U.S. Bureau of the Census (1979). Evaluation. Dans *POPSTAN: Program Considerations*, Partie A, Chapitre A-13.

WOLFGANG, GLENN (1987). The pre-enumeration survey of the 1986 census of Central Los Angeles County. Article présenté à la conférence annuelle de l'American Statistical Association, San Francisco.

WOLTER, KIRK M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

ZASLAVASKY, ALAN M. (1988). Representing local area undercount by reweighting of households. Proceedings of the Bureau of the Census Fourth Annual Research Conference.

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 93–104
Statistique Canada

Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986

NATHANIEL SCHENKER[1]

RÉSUMÉ

Cet article porte sur les méthodes de traitement des données manquantes dans les enquêtes postcensitaires en vue de l'estimation de l'erreur de couverture dans le recensement; à titre d'illustration, nous analysons le test des opérations de redressement de 1986 (Diffendal 1988). Les méthodes précitées comprennent des méthodes d'imputation fondées sur le hot-deck et des modèles de régression logistique de même que des méthodes de redressement par la pondération. Nous analysons également la sensibilité des estimations de sous-dénombrement tirées du test de 1986 à la variation des modèles d'imputation.

MOTS CLÉS: Imputation; non-réponse; enquête postcensitaire; redressement par pondération; sous-dénombrement.

1. INTRODUCTION

Les données manquantes peuvent être une importante source d'incertitude dans l'estimation de l'erreur de couverture pour les recensements décennaux aux États-Unis (Freedman et Navidi 1986; Fay, Passel et Robinson 1988, Chapitre 6). Pour les recensements décennaux de 1960 et de 1980, diverses méthodes de traitement des données manquantes ont produit plusieurs estimations de l'erreur de couverture.

Le Census Bureau a appliqué de nombreux tests à des méthodes d'estimation de l'erreur de couverture pour être en mesure de traiter le problème des données manquantes et les autres problèmes qui pourraient surgir au cours du recensement décennal de 1990. Un de ces tests a été le Test des opérations de redressement (TOR) de 1986 (Diffendal 1988), qui était fondé sur le recensement de 1986 du Central Los Angeles County. Une modification des méthodes appliquées sur le terrain et du plan de sondage du TOR a permis de réduire la quantité de données manquantes par rapport à 1980 (Hogan et Wolter 1988). Malgré cela, certains problèmes persistent.

Cet article vise à décrire les problèmes qu'ont posés les données manquantes dans le TOR et la façon dont ils ont été traités dans le processus d'estimation. Dans la section 2, nous décrivons brièvement la méthode d'estimation de l'erreur de couverture appliquée dans le TOR. Dans les sections 3 à 6, nous analysons le genre et la quantité de données manquantes de même que les méthodes utilisées pour résoudre ce problème, notamment une méthode de redressement par pondération pour les cas de non-réponse au questionnaire (non-interview), une méthode d'imputation "hot-deck" pour les caractéristiques démographiques et les caractéristiques du logement manquantes ainsi qu'une méthode d'imputation fondée sur des modèles de régression logistique pour certains éléments binaires ayant trait au dénombrement dans le recensement. La section 7 présente des estimations de l'erreur de couverture calculées selon divers modèles d'imputation et diverses méthodes de traitement appliquées à des cas précis. Les estimations minimum et maximum du taux de sous-dénombrement calculées à l'aide de ces modèles ou de ces méthodes sont 8.50 et 10.16% pour les personnes d'origine hispanique,

[1] Nathaniel Schenker, Undercount Research Staff, Statistical Research Division, Bureau of the Census, Washington (DC) 20233, É.-U. Cet article est un compte rendu des recherches qui ont été faites par un membre du personnel du Census Bureau. Les opinions qui y sont exprimées sont celles de l'auteur et ne reflètent pas nécessairement la position du Census Bureau.

5.86 et 7.81% pour les personnes d'origine asiatique et 5.81 et 6.59% pour les autres personnes. Les estimations correspondantes calculées à l'aide du TOR étaient 9.85, 7.32 et 6.21% respectivement. Enfin, la section 8 présente une analyse et les conclusions.

2. ESTIMATION DE L'ERREUR DE COUVERTURE DANS LE RECENSEMENT

Diffendal (1988) expose en détail la méthode d'estimation de l'erreur de couverture du recensement utilisée dans le TOR. La présente section expose brièvement les points qu'il est nécessaire de connaître pour comprendre le reste de cet article. L'erreur de couverture a été estimée à l'aide de données d'une enquête postcensitaire (EP) menée auprès de personnes demeurant dans la région de recensement. On a commencé par prélever un échantillon d'îlots dans la région. Ensuite, chaque unité de logement contenue dans ces îlots a fait l'objet d'une enquête afin d'en connaître les occupants le jour du recensement, les occupants le jour de l'EP (et l'endroit où ces derniers demeuraient le jour du recensement) et de déterminer les caractéristiques de ces occupants.

Deux échantillons ont servi à estimer l'erreur de couverture du recensement. L'échantillon P (pour population) était composé des personnes qui demeuraient dans les îlots échantillonnés de l'EP le jour de cette enquête. On tentait d'apparier chaque membre de l'échantillon P à une personne qui avait été recensée afin de déterminer si ce membre avait lui-même été recensé; le taux d'appariement à l'intérieur de chaque domaine d'analyse servait essentiellement à estimer le taux de saisie du recensement pour ce domaine. L'échantillon D (pour dénombrement) était composé des personnes qui demeuraient dans les îlots échantillonnés de l'EP le jour du recensement; cet échantillon a servi à estimer le nombre d'enregistrements erronés (par ex., enregistrements fictifs ou redoublés) et de personnes non appariables (par ex., personnes pour lesquelles aucun nom n'était indiqué) dans le recensement à l'intérieur de chaque domaine. On tentait d'apparier chaque membre de l'échantillon D à une personne qui avait participé à l'enquête postcensitaire. Chaque cas d'appariement était considéré comme un enregistrement exact puisque l'EP indiquait que la personne devait avoir été recensée. Chaque cas de non-appariement était approfondi pour déterminer s'il s'agissait d'un enregistrement erroné ou d'un enregistrement exact qui avait été oublié dans l'EP (laquelle n'est pas censée être à l'abri des erreurs de couverture).

Si une EP était réalisée à la grandeur des États-Unis, les membres de l'échantillon P qui ont quitté le Central Los Angeles County entre le jour du recensement et le jour de l'EP participeraient quand même à cette enquête. On tenterait ensuite d'apparier ces personnes aux enregistrements du Central Los Angeles County faits le jour du recensement et on se servirait des résultats pour estimer l'erreur de couverture dans cette région. De même, les membres de l'échantillon P qui se sont établis dans le Central Los Angeles County entre le jour du recensement et le jour de l'EP contribueraient à l'estimation de l'erreur de couverture à l'extérieur de cette région. Toutefois, comme le recensement et l'EP pour le TOR ne s'appliquaient qu'au Central Los Angeles County (et non à tous les États-Unis), les personnes qui avaient quitté la région d'essai ne participaient pas à l'EP et celles qui s'y étaient établies n'étaient pas concernées. Par conséquent, les données relatives à ces deux catégories de personnes n'ont pas servi à l'estimation. (Notons toutefois que les données relatives aux personnes qui ont déménagé *sans quitter* la région d'essai ont servi à l'estimation). Cette question est analysée plus à fond dans la section 7.2.

L'estimateur "bivalent" de la taille de la population (voir Marks, Seltzer et Krotki 1974; Krotki 1978 et Wolter 1986 pour analyse et bibliographie) est défini par

$$\text{DSE} = N_p\,(\text{CEN-SUB-EE})/\text{M}, \tag{1}$$

où N_p est l'estimateur de la population de l'EP, CEN est le chiffre du recensement non rajusté, SUB est le nombre de substitutions de personnes (pour les cas de non-réponse au questionnaire) dans le recensement, EE est un estimateur du nombre d'enregistrements erronés et de personnes non appariables dans le recensement et M est l'estimateur du nombre de personnes qui font partie à la fois de la population du recensement et de la population de l'EP; les valeurs CEN et SUB sont déduites des données du recensement tandis que N_p, EE et M sont établis à partir des données des échantillons P et D. De fait, le calcul de l'estimateur bivalent équivaut à multiplier le nombre estimé d'enregistrements exacts et appariables du recensement (CEN-SUB-EE) par l'inverse du taux de saisie estimé du recensement (M/N_p).

La théorie de l'estimation "bivalente" suppose que la probabilité de saisie est la même pour tous les membres du domaine auquel s'applique l'estimateur et ce, aussi bien pour le recensement que pour l'enquête postcensitaire (Wolter 1986). Par conséquent, un groupe quelconque de personnes incluses dans le domaine n'aura pas plus de chances ni moins de chances que n'importe quel autre groupe d'être dénombré dans le recensement ou l'EP. Pour rendre cette hypothèse plus réaliste dans le TOR, on a calculé des estimateurs bivalents à l'intérieur de strates formées a posteriori en fonction des caractéristiques des personnes et du logement. Les strates formées a posteriori sont décrites dans Diffendal (1988). Il peut s'agir par exemple des locataires de sexe masculin et d'origine hispanique qui sont âgés de 30 à 44 ans et vivent dans des îlots à majorité hispanique.

En résumé, les données des échantillons P et D qu'il est nécessaire de connaître pour estimer l'erreur de couverture sont le code d'appariement (concordance vs. non-concordance) de chaque membre de l'échantillon P, le code de dénombrement (exact vs. erroné) de chaque membre de l'échantillon D et les caractéristiques des personnes et du logement pour chacun des membres des deux échantillons.

3. MÉNAGES DE L'ÉCHANTILLON P NON INTERVIEWÉS

Il se peut qu'un interviewer n'ait pas été en mesure de réaliser une interview pour une unité de logement occupée à cause, par exemple, d'un refus de répondre de la part des occupants. Sur les 5,935 unités de logement qui étaient considérées comme occupées, 32 (0.5%) ont été classées parmi les cas de non-interview. À cause de ces cas de non-interview, il n'a pas été possible d'obtenir toutes les données voulues concernant le nombre de personnes dans chaque ménage, les caractéristiques des personnes et du logement et les codes d'appariement.

Grâce au plan de sondage de l'EP (fondée sur un échantillon d'îlots), il a été facile de résoudre le problème de la non-interview de ménages de l'échantillon P. Dans chaque îlot échantillonné, les poids de sondage des ménages de non-interview ont été redistribués parmi les ménages interviewés. La repondération pour non-interview suppose essentiellement que la distribution des personnes, des caractéristiques et des codes d'appariement pour les ménages non interviewés à l'intérieur d'un îlot est la même que pour les ménages interviewés. On a utilisé cette hypothèse car les ménages d'un même îlot tendent à avoir plus de caractéristiques en commun que les ménages d'îlots différents, bien qu'il existe probablement des différences notables entre les ménages de non-interview et les ménages interviewés, surtout au point de vue de la taille (voir, par ex., Palmer 1967).

Il se peut que les données recueillies sur un ménage au moyen d'une interview par personne interposée (dans le TOR, cela signifie une interview complète réalisée avec une personne qui n'est pas membre du ménage) soient d'une qualité tellement faible que l'on est porté à classer ce ménage dans les ménages de non-interview. La qualité des données provenant des 189 interviews par personne interposée du TOR est analysée dans la section 4 tandis que la section 7 renferme quelques estimations de l'erreur de couverture établies lorsque les interviews par personne interposée sont considérées comme des non-interviews.

4. CARACTÉRISTIQUES MANQUANTES DANS LES ÉCHANTILLONS P ET D

Même lorsqu'on avait pu réalisé une interview avec un ménage de l'échantillon *P*, les données concernant les caractéristiques des personnes et du logement étaient parfois incomplètes. Il y avait aussi des données incomplètes dans le recensement et, partant, dans l'échantillon *D*.

Les variables utilisées pour la stratification a posteriori dans le TOR (Diffendal 1988) comprenaient la variable de logement Mode d'Occupation (1 = logement occupé par le propriétaire, 2 = logement loué ou occupé sans contrepartie) et les variables personnelles Sexe (1 = homme, 2 = femme), Âge (1 = 0-14, 2 = 15-29, 3 = 30-44, 4 = 45-64, 5 = 65+) et Origine Ethnique (1 = origine hispanique, 2 = origine asiatique, 3 = autre). En outre, la variable de logement Type de Construction (1 = maison unifamiliale, 2 = immeuble à logements multiples) a servi à traiter les codes d'appariement manquants dans l'échantillon *P* et les codes de dénombrement manquants dans l'échantillon *D* (voir sections 5 et 6).

Le tableau 1 renferme les chiffres concernant les caractéristiques manquantes pour les échantillons *P* et *D* au complet et pour les cas d'interview par personne interposée de l'échantillon *P*. En ce qui concerne les échantillons *P* et *D*, le taux de données manquantes le plus élevé est de 7.0% (pour la variable Origine Ethnique dans l'échantillon *D*), tous les autres taux n'excédant pas 3.5%. Les taux de données manquantes pour les interviews par personne interposée de l'échantillon *P* sont tous plusieurs fois supérieurs aux taux observés pour l'échantillon *P* en général, quoique seule la variable Mode d'Occupation (20.2%) présente un taux de données manquantes supérieur à 10%.

Les caractéristiques manquantes pour chacun des échantillons (*P* et *D*) ont été imputées par une méthode "hot-deck" comportant deux lectures de données après que les données ont été classées selon des critères géographiques. À la première lecture, les valeurs manquantes des variables Mode d'Occupation, Type de Construction et Origine Ethnique ont été imputées à l'aide des observations les plus récentes à cause de la relation étroite qui est censée exister entre ces variables et les caractéristiques géographiques. En outre, on a établi la distribution des variables Sexe et Âge en fonction du genre de ménage (ménage à une personne ou à plusieurs personnes), de l'état matrimonial, du lien avec le chef de ménage et du sexe et de l'âge du chef de ménage en se servant de toutes les observations. À la seconde lecture, les valeurs manquantes des variables Sexe et Âge ont été imputées aléatoirement à partir des distributions établies au cours de la première lecture. Pour plus de détails sur l'imputation des caractéristiques dans le TOR, voir Schenker (1987).

En résumé, le plan de sondage de l'EP (fondée sur un échantillon d'îlots) a servi non seulement à l'élaboration d'un schéma de pondération pour non-interviews (section 3) mais aussi à l'imputation de caractéristiques qui tendent à être réunies dans un même îlot, c'est-à-dire Mode d'Occupation, Type de Construction et Origine Ethnique.

Tableau 1

Chiffres sur les caractéristiques manquantes
(pourcentages entre parenthèses) pour les échantillons *P* et *D*
et les cas d'interview par personne interposée de l'échantillon *P*

Variable	Échantillon *P* (19,552 personnes)	Échantillon *D* (20,976 personnes)	Interviews par personne interposée de l'échantillon *P* (430 personnes)
Mode d'Occupation	690 (3.5)	154 (0.7)	87 (20.2)
Type de Construction	459 (2.3)	343 (1.6)	38 (8.8)
Sexe	418 (2.1)	82 (0.4)	18 (4.2)
Âge	137 (0.7)	432 (2.1)	18 (4.2)
Origine Ethnique	155 (0.8)	1463 (7.0)	17 (4.0)

NOTA: Les 19,552 personnes de l'échantillon *P* comprennent les 430 cas d'interview par personne interposée.

5. CODES D'APPARIEMENT MANQUANTS DANS L'ÉCHANTILLON P

Sur les 19,552 enregistrements de l'échantillon *P* créés à la suite d'une interview complète, 161 (0.8%) n'avaient pas de code d'appariement pour l'estimation "bivalente". Tous ces cas non résolus sauf trois pouvaient être classés dans deux grandes catégories: d'abord, 105 ménages pour lesquels on n'a pas tenté d'appariement à cause de noms incomplets ou de caractéristiques insuffisantes, puis 53 autres ménages qui ont déménagé entre le jour du recensement et le jour de l'EP et pour lesquels il a été difficile de déterminer l'adresse le jour du recensement ou de retracer le questionnaire de recensement.

Une façon classique de pallier l'absence d'un élément binaire comme le code d'appariement est d'imputer l'une des deux valeurs possibles. Par exemple, lorsqu'on a estimé le taux de sous-dénombrement pour le recensement décennal de 1980, on a imputé un code d'appariement pour chaque cas de l'échantillon *P* non résolu en se fondant sur un cas résolu qui présentait les mêmes caractéristiques (Fay, Passel et Robinson 1988, Chapitre 6). Toutefois, une méthode différente a été appliquée dans le TOR. Après avoir imputé toutes les caractéristiques manquantes à l'aide des méthodes décrites dans la section 4, on a imputé une probabilité d'appariement pour chaque code manquant, la probabilité étant estimée au moyen d'un modèle explicite qui sera décrit un peu plus loin. Dans le dénominateur de l'expression qui définit l'estimateur bivalent (équation 1), les cas non résolus sont représentés par la somme pondérée des probabilités imputées.

On a imputé des probabilités plutôt que des valeurs binaires pour deux raisons. La première est que l'imputation de valeurs binaires aléatoires est moins efficiente que l'imputation de probabilités estimées puisqu'elle produit des estimations qui ont une variance plus élevée (voir Rubin 1987, p. 15). En deuxième lieu, comme les probabilités imputées représentent une incertitude à propos des codes d'appariement manquants, ces probabilités devraient pouvoir servir à calculer une variance due à l'imputation. Cependant, comme l'estimateur bivalent (1) est non linéaire en *M*, l'imputation d'une probabilité (ou d'une moyenne) pour chaque élément binaire manquant introduit un biais dans l'estimation (voir Rubin 1987, p. 14). Les statisticiens approfondissent actuellement l'utilisation de probabilités imputées pour des données binaires manquantes.

La méthode de régression logistique suivante a servi à imputer les probabilités d'appariement. Soit X un vecteur de variables explicatives, Y = concordance ou non-concordance, et $p = \Pr(Y = \text{concordance} \mid X)$. Le vecteur de paramètres β du modèle de régression logistique

$$\text{logit}(p) = \log[p/(1-p)] = X'\beta$$

a été estimé sur la base des données relatives aux cas résolus à l'aide des méthodes bayesiennes pour régression logistique catégorique décrites dans Rubin et Schenker (1987); ces méthodes consistent à additionner des observations fractionnaires à chaque case ou cellule dans la régression logistique puis à ajuster le modèle par les méthodes classiques du maximum de vraisemblance. Ainsi, pour le cas non résolu j, étant donné $X = xj$, la probabilité d'appariement imputée a été définie par

$$\hat{p}_j = \text{logit}^{-1}(x_j'\hat{\beta}) = \exp(x_j'\hat{\beta})/[1 + \exp(x_j'\hat{\beta})],$$

où $\hat{\beta}$ désigne l'estimateur de β. Les variables de base qui ont servi à définir X sont le Mode d'Occupation, le Type de Construction, le Sexe, l'Âge et l'Origine Ethnique, de même que les variables qui indiquent s'il s'agit d'une interview ordinaire ou d'une interview par personne interposée ou s'il s'agit d'une personne qui a déménagé ou d'une personne qui n'a pas déménagé entre le jour du recensement et le jour de l'EP.

Le tableau A1 (dans l'annexe) donne les estimations des coefficients de la régression logistique. Les coefficients élevés qui se rattachent au code d'interview et au code "déménagement"

indiquent que les probabilités d'appariement imputées sont beaucoup moins élevées dans le cas des interviews par personne interposée et des personnes ayant déménagé que dans le cas des interviews ordinaires et des personnes n'ayant pas déménagé. Cette situation s'explique peut-être en partie par la difficulté d'apparier les cas d'interview par personne interposée et les cas de déménagement plutôt que par le seul fait que les taux de saisie du recensement sont moins élevés dans ces cas. Si cela est exact, le temps est peut-être venu de considérer d'autres méthodes de traitement des données manquantes; c'est ce à quoi nous nous attachons dans la section 7.

Sur les 19,391 cas résolus de l'échantillon P, 17,018 (87.8%) se sont avérés des cas de concordance. La somme (non pondérée) des 161 probabilités d'appariement imputées était de 124.66 (pour un taux d'appariement imputé de 77.4%). Même si l'on a fait usage d'un échantillon stratifié d'îlots dans le TOR, l'estimation des paramètres de régression logistique a supposé un échantillon aléatoire simple de personnes. Afin de vérifier si l'omission de la stratification n'aurait pas créé de biais, nous avons ajusté une fois de plus le modèle de régression logistique (une fois le TOR achevé) en incluant dans X les variables indicatrices des six strates d'échantillonnage (Diffendal 1988). Nous avons refait la somme des probabilités d'appariement imputées et nous avons obtenu 124.50 (77.3%). La faible incidence de ce changement sur les estimations de l'erreur de couverture dans le recensement est démontrée dans la section 7. Dans la section 8, nous examinons les conséquences des effets de plan qui pourraient découler de la formation de grappes.

6. CODES DE DÉNOMBREMENT MANQUANTS DANS L'ÉCHANTILLON D

Sur les 20,976 cas de l'échantillon D, 3,714 ont fait ou auraient dû faire l'objet d'un suivi. Après le suivi, 979 cas (4.7% de l'effectif, 26.4% des cas devant faire l'objet d'un suivi) n'avaient toujours pas de code de dénombrement. Tous ces cas non résolus sauf neuf pouvaient être classés dans quatre grandes catégories: d'abord, 498 cas qui n'ont pas fait l'objet d'un suivi alors qu'ils auraient dû; ensuite, 257 cas où la personne qui a participé à l'interview de rappel ne connaissait pas la personne en question, puis 137 cas pour lesquels l'interview n'a pas produit assez de renseignements pour déterminer un code de dénombrement, et enfin 78 cas pour lesquels il n'a pas été possible d'obtenir une interview de rappel.

On a pallié à l'absence de codes de dénombrement dans l'échantillon D en imputant une probabilité d'enregistrement erroné pour chaque cas non résolu. Dans le terme EE de l'expression définissant l'estimateur bivalent (équation 1), les cas non résolus sont représentés par la somme pondérée des probabilités imputées. La méthode d'imputation différait peu de celle utilisée pour les codes d'appariement. Un seul changement majeur: comme l'absence de codes de dénombrement ne pouvait être constatée qu'à la suite des procédures de suivi, seuls les cas résolus par suivi ont servi à l'estimation du modèle de régression logistique. Les variables de base qui ont servi à définir X pour la régression logistique sont le Mode d'Occupation, le Type de Construction, le Sexe, l'Âge et l'Origine Ethnique, de même que les variables qui indiquent si le questionnaire du recensement concernant le ménage dont fait partie la personne visée a été retourné par la poste ou si le ménage au complet ou une partie de celui-ci n'a pu être apparié avant le suivi. Le tableau 2 (dans l'annexe) donne les estimations des coefficients de la régression logistique.

Sur les 17,262 cas qui ne nécessitaient pas de suivi, 278 (1.6%) ont été classés parmi les enregistrements erronés ou les cas non appariables. Deux mille sept cent trente-cinq cas ont été résolus par suivi et 82 de ces cas (3.0%) ont été classés dans les enregistrements erronés. La somme (non pondérée) des 979 probabilités imputées était de 21.93 (2.2%). Lorsqu'on inclut les variables indicatrices des strates d'échantillonnage dans X, la somme des probabilités imputées devient 23.58 (2.4%). Comme dans le cas de l'échantillon P, ce changement a une très faible incidence sur les estimations de l'erreur de couverture; voir section 7.

7. ESTIMATION DE L'ERREUR DE COUVERTURE SUIVANT D'AUTRES MÉTHODES DE TRAITEMENT DES DONNÉES MANQUANTES ET AUTRES SOURCES DE PROBLÈME

Dans cette section, nous analysons l'incidence d'autres méthodes de traitement des données manquantes et autres sources de problème sur les estimations de l'erreur de couverture pour les trois catégories définies par la variable Origine Ethnique (hispanique, asiatique et autre). Pour une méthode de traitement et une catégorie données, soient $\hat{N}$ la somme des estimations bivalentes pour toutes les strates formées a posteriori qui correspondent à la catégorie donnée et N_c la somme des chiffres de recensement non rajustés pour ces strates. Le taux de sous-dénombrement estimé est alors $100(1 - N_c / \hat{N})\%$.

Nous envisageons tout d'abord d'utiliser les indicateurs des strates d'échantillonnage comme variables explicatives dans les régressions logistiques appliquées aux échantillons P et D pour l'imputation des probabilités d'appariement et d'enregistrement erroné traitées dans les sections 5 et 6. Le TOR a produit des estimations de taux de sous-dénombrement sans faire intervenir ces variables explicatives; ces taux sont 9.85% pour les personnes d'origine hispanique, 7.32% pour les personnes d'origine asiatique et 6.24% pour les autres. Lorsqu'on utilise les indicateurs des strates d'échantillonnage, les taux deviennent 9.82% pour les personnes d'origine hispanique, 7.31% pour les personnes d'origine asiatique et 6.21% pour les autres. L'utilisation des indicateurs de strates d'échantillonnage ne modifie les taux initiaux que de 0.03 point de pourcentage au maximum. Néanmoins, cette utilisation sera considérée dans toutes les méthodes que nous analyserons car elle produit en principe des résultats plus précis; par exemple, on devrait obtenir des erreurs types plus justes.

7.1 Méthodes qui réduisent le taux de sous-dénombrement estimé

Le taux d'appariement pour les 375 cas d'interview par personne interposée de l'échantillon P qui ont été résolus était de 78.9% comparativement à 87.8% pour l'ensemble de l'échantillon P. Même s'il est probable que les cas d'interview par personne interposée aient eu un taux de saisie moins élevé que les cas d'interview ordinaire dans le recensement, l'écart entre les taux d'appariement est peut-être attribuable en partie à l'absence ou à l'inexactitude de données provenant d'interviews par personne interposée (voir section 4). Une façon prudente de procéder consisterait à classer les 189 interviews par personne interposée dans les non-interviews et à appliquer la méthode de pondération décrite dans la section 3; ainsi, les cas d'interview par personne interposée auraient essentiellement le même taux d'appariement que les cas d'interview ordinaire. (Notons que lorsque toutes les interviews par personne interposée sont considérées comme des non-interviews, le code d'interview ne figure plus dans le modèle de régression logistique pour l'imputation des probabilités d'appariement.)

Le taux d'appariement pour les 277 cas de déménagement (entre le jour du recensement et le jour de l'EP) de l'échantillon P qui ont été résolus était de 66.1%. Bien que l'on croit généralement que les personnes ayant déménagé ont un taux de saisie moins élevé dans le recensement que les personnes n'ayant pas déménagé, le faible taux d'appariement observé dans le cas des personnes ayant déménagé est peut-être attribuable en partie aux difficultés inhérentes à l'appariement de cette catégorie de personnes, par exemple la difficulté d'obtenir l'adresse à laquelle demeurait la personne le jour du recensement. Une façon prudente de procéder consisterait à classer tous les cas de déménagement dans les cas non résolus puis à imputer des probabilités d'appariement pour les cas non résolus à l'aide d'un modèle de régression logistique qui n'utilise pas le code "déménagement" comme variable explicative. Ainsi, les personnes ayant déménagé auraient essentiellement le même taux d'appariement que les personnes n'ayant pas déménagé.

Sur les 979 cas non résolus de l'échantillon D, 257 portaient le code d'interview de suivi $W1$, qui signifiait que le répondant ne connaissait pas la personne en question. Ce code pouvait aussi indiquer que la personne en question était un individu fictif. C'est pourquoi, après le TOR,

Tableau 2

Taux de sous-dénombrement estimés (en ‰) selon l'origine ethnique et suivant diverses méthodes de traitement des cas d'interview par personne interposée de l'échantillon *P*, des cas de déménagement de l'échantillon *P* et des cas de l'échantillon *D* portant le code *W*1

Indicateur de méthode (1 = autre, 0 = TOR)			Origine hispanique	Origine asiatique	Autre
Interview par personne interposée	Personne ayant déménagé	W1			
0	0	0	9.82	7.31	6.21
0	0	1	9.30	6.76	5.83
0	1	0	9.33	7.24	6.19
0	1	1	8.80	6.69	5.81
1	0	0	9.55	6.52	6.24
1	0	1	9.03	5.96	5.86
1	1	0	9.04	6.45	6.22
1	1	1	8.51	5.90	5.84

NOTA: Les indicateurs des strates d'échantillonnage ont servi de variables explicatives dans les régressions logistiques pour l'imputation des probabilités d'appariement et d'enregistrement erroné.

tous les enregistrements portant le code *W*1 ont été vérifiés par des spécialistes de l'appariement. Ceux-ci identifiaient les enregistrements qui pouvaient laisser croire par un signe quelconque (par exemple: une note de l'interviewer) qu'il s'agissait d'un cas fictif; ils en ont relevé 118. Une méthode différente de celle utilisée dans le TOR consisterait à classer les 118 cas dans les enregistrements erronés résolus avant l'imputation. Cela aurait pour effet de hausser les taux observé et imputé d'enregistrement erroné.

Le tableau 2 donne les taux de sous-dénombrement estimés selon l'origine ethnique pour le plan factoriel 2x2x2, où les facteurs indiquent si des méthodes autres que celle utilisée dans le TOR ont été appliquées aux cas d'interview par personne interposée, aux cas de déménagement et aux cas portant le code *W*1. L'écart entre l'estimation la plus faible et l'estimation la plus élevée du taux de sous-dénombrement est de 1.31 point de pourcentage pour les personnes d'origine hispanique, de 1.41 point pour les personnes d'origine asiatique et de 0.43 point pour les autres.

Il convient de souligner qu'il y a peu d'interaction entre les méthodes de traitement des cas d'interview par personne interposée, des cas de déménagement et des cas portant le code *W*1 pour chaque origine ethnique. De fait, on peut utiliser le modèle additif simple ci-dessous pour prévoir les valeurs du tableau 2 pour chaque origine ethnique:

$$\hat{Y} = \hat{\alpha}_0 + I_p\hat{\alpha}_p + I_m\hat{\alpha}_m + I_w\hat{\alpha}_w, \qquad (2)$$

où $\hat{Y}$ est l'estimation prévue du taux de sous-dénombrement, I_p, I_m, et I_w sont les indicateurs de méthode (1 = autre, 0 = TOR) pour les cas d'interview par personne interposée, les case de déménagement et les cas portant le code *W*1 respectivement et $\hat{\alpha}_0$, $\hat{\alpha}_p$, $\hat{\alpha}_m$, et $\hat{\alpha}_w$, sont les estimations de paramètres données dans le tableau 3. Le paramètre α_0 est le taux de sous-dénombrement estimé lorsqu'aucune autre méthode que celle du TOR n'est utilisée; α_p, α_m, et α_w sont les effets de l'application d'autres méthodes aux cas d'interview par personne

Tableau 3

Estimations des paramètres du modèle additif (2) servant à prévoir les taux de sous-dénombrement estimés du tableau 2

	Origine hispanique	Origine asiatique	Autre
$\hat{\alpha}_o$	9.82	7.31	6.21
$\hat{\alpha}_p$	−0.28	−0.7925	0.03
$\hat{\alpha}_m$	−0.505	−0.0675	−0.02
$\hat{\alpha}_w$	−0.525	−0.5525	−0.38

interposée, aux cas de déménagement et aux cas portant le code *W*1 respectivement. Lorsqu'on utilise l'équation (2) pour prévoir les valeurs du tableau 2, le résiduel le plus élevé est 0.02%.

7.2 Méthode qui accroît le taux de sous-dénombrement estimé

Comme le TOR était limité à une très petite région des États-Unis, l'enquête postcensitaire n'a pas permis de recueillir des données sur les personnes qui avaient quitté la région d'essai après le recensement. Le fait d'exclure ces personnes de l'estimation revenait à supposer qu'elles avaient le même taux de saisie dans le recensement que les personnes incluses dans l'estimation. C'était là une hypothèse modérée puisque l'on sait que les personnes ayant déménagé ont habituellement un taux de saisie moindre que les personnes n'ayant pas déménagé.

Quatre cent neuf personnes sont venues s'établir dans la région d'essai entre le jour du recensement et le jour de l'EP. Ces nouveaux venus n'ont pas été inclus dans l'estimation puisque le jour du recensement, ils demeuraient à l'extérieur de la région d'essai et que, par conséquent, les données recueillies ce jour-là s'appliquent à d'autres régions. De plus, il n'était pas possible d'apparier ces nouveaux venus à des enregistrements du recensement puisqu'ils vivaient à l'extérieur de la région d'essai le jour du recensement.

Pour avoir une idée des conséquences qu'aurait l'inclusion des sortants dans l'estimation, on pourrait considérer à la place les 409 entrants et imputer des probabilités d'appariement pour ces personnes (puisque leurs codes d'appariement sont inconnus). Les combinaisons qui ont produit les estimations les plus élevées et les plus faibles dans le tableau 2 ont été appliquées aux données du TOR en tenant compte des entrants; les résultats pertinents figurent dans le tableau 4. On remarquera que les estimations obtenues par l'application de méthodes autres

Tableau 4

Taux de sous-dénombrement estimés (en %) selon l'origine ethnique lorsque les entrants sont considérés dans l'estimation avec probabilités d'appariement imputées

Indicateur de méthode (1 = autre, 0 = TOR)			Origine hispanique	Origine asiatique	Autre
Interview par personne interposée	Personne ayant déménagé	W1			
0	0	0	10.16	7.81	6.59
1	1	1	8.50	5.86	5.81

NOTA: Les indicateurs des strates d'échantillonnage ont servi de variables explicatives dans les régressions logistiques pour l'imputation des probabilités d'appariement et d'enregistrement erroné.

que celle utilisée dans le TOR pour les cas d'interview par personne interposée, les cas de déménagement et les cas portant le code *W*1 sont toutes inférieures aux estimations correspondantes du tableau 2 et l'écart maximum est de 0.04 point de pourcentage. Ces résultats étaient prévisibles puisque l'addition de cas ayant un taux d'appariement imputé comparable au taux d'appariement global n'est pas censée avoir beaucoup d'effet sur les estimations. Les estimations obtenues par l'application de la méthode utilisée dans le TOR sont toutes supérieures aux estimations correspondantes du tableau 2, les écarts étant de 0.34 point de pourcentage pour les personnes d'origine hispanique, de 0.50 point de pourcentage pour les personnes d'origine asiatique et de 0.38 point de poucentage pour les autres.

8. RÉSUMÉ ET ANALYSE

Une combinaison de méthodes de pondération et d'imputation (aléatoire et non aléatoire) a servi à traiter les données manquantes dans le TOR. Les cas de non-interview dans l'échantillon *P* ont été résolus grâce à une méthode de redressement par pondération au niveau de l'îlot. Une méthode hot-deck a servi à imputer les caractéristiques manquantes dans les deux échantillons. Pour pallier l'absence de codes d'appariement dans l'échantillon *P* et de codes de dénombrement dans l'échantillon *D*, on a imputé des probabilités estimées par des méthodes de régression logistique.

Comme nous l'avons indiqué dans les sections 5 et 6, l'utilisation de probabilités imputées pour les codes d'appariement et de dénombrement manquants devrait faciliter le calcul de la variance due à l'imputation de ces codes. Pour évaluer entièrement cette variance, il est nécessaire de mesurer la variance due à l'estimation des paramètres de la régression logistique de même que la variance due à l'imputation étant donné β (Rubin et Schenker 1986). Il nous faut donc une matrice estimée des variances-covariances pour $\hat{\beta}$. Comme on a procédé à un échantillonnage par grappes dans le TOR, l'estimation par régression logistique (section 5), qui suppose un échantillonnage aléatoire simple, ne produit pas une estimation juste de la matrice des variances-covariances. Cela ne posait pas vraiment de problèmes dans le TOR puisque celui-ci ne visait pas principalement à mesurer la variance due à l'imputation. En outre, pour ce qui a trait aux taux de non-réponse enregistrés dans le test, la variance due à l'incertitude entourant l'estimation de β sera vraisemblablement peu élevée par rapport à la variance due à l'imputation étant donné β (Rubin et Schenker 1986).

Bien qu'il soit possible en principe d'évaluer la variance due à l'imputation de codes d'appariement et de dénombrement au moyen de la méthode du TOR, on ne peut encore évaluer la variance due à l'imputation de caractéristiques manquantes (section 4). En fait, il serait possible d'évaluer cette variabilité en multipliant les caractéristiques d'imputation dans les échantillons *P* et *D* (Rubin 1987). Il faudrait alors calculer plusieurs estimateurs bivalents – un pour chaque série d'imputations.

Les modèles sur lesquels reposent les méthodes de pondération et d'imputation utilisées dans le TOR supposent que la probabilité qu'une variable soit manquante ne dépend pas de la valeur de celle-ci, étant donné les données observées. Une autre question ayant trait à l'imputation est de savoir quelle est la meilleure façon d'imputer simultanément des caractéristiques et des codes d'appariement (ou de dénombrement). La méthode utilisée dans le TOR (qui consiste à imputer tout d'abord les caractéristiques et ensuite les codes en fonction des caractéristiques imputées) suppose que les codes ne sont pas des variables explicatives utiles pour l'imputation de caractéristiques. Il serait peut-être plus indiqué d'adopter des modèles qui assouplissent les hypothèse du TOR. Rubin, Schafer et Schenker (1988) approfondissent la question.

Les données manquantes constituent une seule source d'erreur dans l'estimation de l'erreur de couverture. D'autres sources, comme l'erreur d'appariement et le non-respect de l'hypothèse des probabilités de saisie constantes (section 2), sont analysées dans Hogan et Wolter (1988). Après avoir évalué toutes ces sources d'erreur pour le TOR, Hogan et Wolter concluent à la supériorité du TOR par rapport au dénombrement initial.

REMERCIEMENTS

L'auteur tient à remercier Robert O'Brien, qui a calculé les estimations bivalentes utilisées dans la section 7. Il veut aussi exprimer sa reconnaissance aux arbitres pour leurs commentaires utiles et aux membres du comité d'étude sur le sous-dénombrement du Département de statistique de l'Université Harvard pour leur participation enrichissante aux discussions sur le traitement des données manquantes dans l'estimation du sous-dénombrement.

ANNEXE

RÉSULTATS DE LA RÉGRESSION LOGISTIQUE

Tableau A1

Résultats de la régression logistique pour l'échantillon *P*

Variable explicative	Codes	Coefficient estimé
Ordonnée à l'origine		1.47
Code d'interview	1 si interview ordinaire, −1 si interview par personne interposée	.36
Code "déménagement"	1 si la personne n'a pas déménagé, −1 si la personne a déménagé	.60
Mode d'occupation	1 si logement occupé par le propriétaire, −1 dans les autres cas	.46
Type de construction	1 si maison unifamiliale, −1 si immeuble à logements multiples	−.16
Sexe	1 pour homme, −1 pour femme	−.09
Âge 1	1 si de 0 à 14 ans, −1 si 65 et plus, 0 dans les autres cas	−.06
Âge 2	1 si de 15 à 29 ans, −1 si 65 ans et plus, 0 dans les autres cas	−.46
Âge 3	1 si de 30 à 44 ans, −1 si 65 ans et plus, 0 dans les autres cas	−.02
Âge 4	1 si de 45 à 59 ans, −1 si 65 ans et plus, 0 dans les autres cas	.13
Origine ethnique 1	1 si hispanique, −1 si autre, 0 si asiatique	−.14
Origine ethnique 2	1 si asiatique, −1 si autre, 0 si hispanique	.11

Tableau A2

Résultats de la régression logistique pour l'échantillon *D*

Variable explicative	Codes	Coefficient estimé
Ordonnée à l'origine		−3.45
Code de questionnaire	1 si retourné par la poste, −1 dans les autres cas	.01
Code de pré-suivi	1 si non-concordance partielle du ménage, −1 si non-concordance totale	−.20

Tableau A2 - Suite

Résultats de la régression logistique pour l'échantillon *D*

Variable explicativ	Codes	Coefficient estimé
Mode d'occupation	1 si logement occupé par le propriétaire, −1 dans les autres cas	.36
Type de construction	1 si maison unifamiliale, −1 si immeuble à logements multiples	.17
Sexe	1 pour homme, −1 pour femme	.08
Âge 1	1 si de 0 à 14 ans, −1 si 65 et plus, 0 dans les autres cas	−.30
Âge 2	1 si de 15 à 29 ans, −1 si 65 ans et plus, 0 dans les autres cas	−.04
Âge 3	1 si de 30 à 44 ans, −1 si 65 ans et plus, 0 dans les autres cas	−.34
Âge 4	1 si de 45 à 59 ans, −1 si 65 ans et plus, 0 dans les autres cas	.10
Origine ethnique 1	1 si hispanique, −1 si autre, 0 si asiatique	−.02
Origine ethnique 2	1 si asiatique, −1 si autre, 0 si hispanique	−.38

BIBLIOGRAPHIE

DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County, *Techniques d'enquête*, 14.

FAY, R.E., PASSEL, J.S., et ROBINSON, J.G. (1988). *The Coverage of Population in the 1980 Census*, 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.

FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression Models for Adjusting the 1980 Census, *Statistical Science*, 1, 3-39.

HOGAN, H.R., et WOLTER, K.M. (1988). Mesure de l'erreur dans une enquête post-censitaire, *Techniques d'enquête*, 14.

KROTKI, K. (1978). *Developments in Dual System Estimation of Population Size and Growth*, Edmonton, The University of Alberta Press.

MARKS, E.S., SELTZER, W., et KROTKI, K.J. (1974). *Population Growth Estimation*, New York: The Population Council.

PALMER, S. (1967). On the Character and Influence of Nonresponse in the Current Population Survey, *Proceedings of the Social Statistics Section*, American Statistical Association, 73-80.

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

RUBIN, D.B., SCHAFER, J.L., et SCHENKER, N. (1988). Imputation Strategies for Estimating the Undercount, *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census.

RUBIN, D.B., et SCHENKER, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse, *Journal of the American Statistical Association*, 81, 366-374.

RUBIN, D.B., et SCHENKER, N. (1987). Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior, *Sociological Methodology*, 17, 131- 144.

SCHENKER, N. (1987). Handling Missing Data in the 1986 Test of Adjustment Related Operations, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

WOLTER, K.M. (1986). Some Coverage Error Models for Census Data, *Journal of the American Statistical Association*, 81, 338-346.

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 105–124
Statistique Canada

Mesure de l'erreur dans une enquête post-censitaire

HOWARD HOGAN et KIRK WOLTER[1]

RÉSUMÉ

Le U.S. Bureau of the Census aura recours à une enquête post-censitaire pour évaluer le taux de couverture du recensement décennal de 1990. Cet article décrit les nouvelles méthodes que le Bureau a mises au point et expérimentées pour accroître le degré de précision des estimations de cette enquête. Il étudie aussi les catégories d'erreur qui surviennent dans une enquête post-censitaire de même que les moyens qui permettent de juger de l'exactitude des résultats. À cette fin, les auteurs se servent d'une enquête post-censitaire qui a été réalisée récemment à titre expérimental.

MOTS CLÉS: Recensement; sous-dénombrement; sur-dénombrement; évaluation du taux de couverture.

1. INTRODUCTION

Cet article porte sur les recherches effectuées récemment au U.S. Bureau of the Census pour accroître et mesurer la précision des estimations d'une enquête post-censitaire. À l'origine, ces recherches visaient principalement à élaborer un ensemble cohérent de principes, de méthodes et d'opérations statistiques devant permettre de redresser les chiffres du recensement des États-Unis en fonction de l'erreur de couverture. Les résultats exposés dans cet article montrent qu'il est désormais possible de produire à l'aide d'une enquête post-censitaire (EP) des estimations de la population totale qui sont plus proches de la réalité que les estimations du recensement.

Compte tenu de la décision du Département du commerce des É.-U. de ne pas redresser les chiffres du recensement de 1990 eu égard à l'erreur de couverture, les méthodes de l'EP analysées ici serviront à évaluer soigneusement le taux de couverture de ce recensement. Voir Département du commerce des É.-U. (1987). Les résulats de cette évaluation serviront ensuite à informer les utilisateurs des limites du recensement, à orienter la planification des recensements futurs et à améliorer les estimations démographiques du Bureau of the Census pour les années qui suivent l'année du recensement.

L'EP utilise deux échantillons pour mesurer l'erreur de couverture nette. D'abord, un échantillon de personnes qui devraient normalement avoir été recensées initialement est interviewé après le recensement et sert à évaluer le nombre des omissions. Nous l'appelons l'échantillon P (pour population). Il nous faut aussi un échantillon d'enregistrements du recensement pour évaluer le nombre de répétitions et les autres erreurs contenues dans les comptes du recensement. Nous l'appelons l'échantillon D (pour dénombrement). L'estimateur dual (ED) nous permet de déduire de ces échantillons une estimation de la population totale. Voir Diffendal (1988) pour une analyse détaillée des échantillons et du modèle dual. Sauf indication contraire, nous utiliserons dans cet article la notation adoptée par Diffendal.

Le Bureau of the Census a réalisé une EP en rapport avec le recensement de 1980. L'échantillon P était constitué des membres des ménages qui avaient été enquêtés lors des sondages d'avril et d'août de la Current Population Survey (CPS). Pour une description de la CPS, voir U.S. Bureau of the Census (1978). L'échantillon D était un échantillon distinct et

[1] U.S. Bureau of the Census, Washington (D.C.) 20233. Howard Hogan est chef du Undercount Research Staff. Kirk Wolter est chef de la Statistical Research Division. Cet article est un compte rendu des recherches qui ont été faites par le personnel du U.S. Bureau of the Census. Les opinions qui y sont exprimées sont celles des auteurs et ne reflètent pas nécessairement la position de l'organisme.

indépendant constitué de personnes qui vivaient dans des unités de logement recensées. En outre, le Bureau of the Census avait produit une autre série d'estimations du taux de sous-dénombrement fondées sur une analyse globale des registres des naissances et des décès, de fichiers administratifs et de recensements antérieurs. Ce programme, appelé analyse démographique, sera évoqué à quelques reprises dans notre étude. Le Bureau of the Census n'a pas redressé les chiffres du recensement de 1980 en fonction de l'erreur de couverture car il considérait que les estimations de l'EP étaient faussées par des données manquantes ou inexactes. En outre, les résultats de l'analyse démographique laissaient à désirer du fait, notamment, que l'on ne disposait pas de données suffisantes sur le nombre d'immigrants sans statut légal et qu'il n'existait pas de méthode satisfaisante pour établir des estimations au niveau de l'État ou de la municipalité. Voir U.S. Bureau of the Census (1988).

Ces dernières années, nous avons élaboré pour l'EP un nouveau plan de sondage et de nouvelles méthodes qui atténuent les problèmes que nous avons connus en 1980 sans en créer d'autres d'importance majeure. Le nouveau plan de sondage de l'EP est fondé sur un échantillon aréolaire de blocs de recensement à partir duquel est constitué l'échantillon P. L'échantillon P comprend toutes les personnes qui demeurent dans les blocs échantillonnés au moment de l'interview de l'EP. Les intervieweurs vont dans chaque unité de logement et déterminent à quel endroit vivaient les occupants au moment du recensement.

À l'aide d'un logiciel et de méthodes d'appariement élaborés récemment (Jaro 1988), nous tentons d'apparier toutes les personnes de l'échantillon P aux enregistrements du recensement correspondants. Des commis vérifient les résultats de l'appariement informatisé et attribuent à chaque personne de l'échantillon P un code de dénombrement qui indique si la personne a été enregistrée ou manquée lors du dénombrement initial. Pour ce qui a trait aux personnes qui ont déménagé entre le jour du recensement et le jour de l'EP, nous associons l'adresse déclarée le jour du recensement au bloc approprié et tentons d'établir l'appariement à cet endroit. Il arrive que l'appariement ne soit pas décisif pour tous; pour les quelques cas en suspens, une interview de rappel est alors nécessaire pour recueillir des renseignements additionnels ou résoudre les incohérences que renferment les données déjà recueillies. Après le suivi, les commis attribuent un code de dénombrement aux personnes de l'échantillon P pour lesquelles on a complété une interview de rappel. Malgré toutes ces étapes, il est possible qu'un très petit nombre de cas demeurent en suspens; nous imputons alors à chacun d'eux un code de dénombrement en appliquant des méthodes statistiques appropriées pour les données manquantes (Schenker 1988).

En ce qui a trait aux personnes de l'échantillon D, on attribue également à chacune d'elles un code de dénombrement indiquant si la personne a été enregistrée correctement ou incorrectement lors du recensement. Dans la section 6, nous décrivons ce qui constitue un enregistrement incorrect ou erroné (EE); par ailleurs, tous les enregistrements non erronés sont considérés comme des enregistrements corrects (EC). Un grand nombre des personnes interviewées comme membres de l'échantillon P ont aussi été enregistrées lors du recensement. Les deux échantillons se chevauchent donc dans une forte proportion. Tous les membres de l'échantillon D qui font aussi partie de l'échantillon P (conformément aux résultats de l'appariement informatisé et manuel) sont reconnus automatiquement comme des enregistrements corrects (EC). Toutefois, le chevauchement n'est pas parfait. L'échantillon D pourra compter des personnes qui ne figurent pas dans l'échantillon P et vice versa. D'autres personnes dans le bloc pourront être recensées par erreur. Des intervieweurs pourront fabriquer des enregistrements. Chaque personne de l'échantillon D qui n'aura pu être appariée à une personne de l'échantillon P devra faire l'objet d'une interview de rappel. Cette interview permet de recueillir suffisamment de données pour déterminer si une personne de l'échantillon D a été enregistrée correctement lors du recensement.

Le nouveau plan de sondage de l'EP a été expérimenté en 1986 à l'occasion d'un recensement d'essai réalisé à Los Angeles. Ce recensement, qui a été désigné comme le Test des opérations de redressement (TOR), portait sur 190 blocs renfermant près de 6,000 unités de

Tableau 1

Erreurs du TOR et estimations de l'effet moyen de la correction d'erreur sur le taux de sous-dénombrement estimé

Sources d'erreur	Effet moyen sur le taux de sous-dénombrement estimé
Erreur d'appariement	−1.0%
Déclaration de l'adresse le jour du recensement	−1.0%
Fabrication dans l'interview de l'EP	−1.0%
Données manquantes	0.0%
Erreur dans l'estimation du nombre d'enregistrements erronés	−0.5%
Conciliation des taux de sur-dénombrement et de sous-dénombrement bruts	0.0%
Biais de corrélation	+2.3%
Erreur aléatoire	0.0%

logement et 20,000 personnes. Ce test a permis d'établir un taux de sous-dénombrement net estimé d'environ 9%. Pour plus de détails sur les méthodes et les résultats du TOR, voir Diffendal (1988) et Schenker (1988).

Nous avons aussi expérimenté le nouveau plan de sondage dans une région rurale du Mississippi la même année. Nous avons utilisé à cette fin un échantillon de 271 blocs comprenant environ 3,250 unités de logement et 8,000 personnes. Ce test a permis d'établir un taux de sous-dénombrement estimé de 5.5%. Pour plus de détails sur les résultats et les méthodes utilisées, voir Anolik (1988). Bien que les données de ce dernier test n'aient pas été analysées aussi minutieusement que les données du TOR, nous nous reporterons parfois aux résultats de l'analyse au cours de cet article.

Une question importante est de savoir si la nouvelle EP peut produire des estimations démographiques plus justes que ne peut le faire le recensement. En théorie, les estimations de l'EP devraient être considérées comme les plus justes mais, en pratique, la réalisation de l'EP et du recensement, de même que l'analyse des résultats, ne sont pas à l'abri des erreurs non dues à l'échantillonnage. Une analyse sérieuse est nécessaire pour évaluer la justesse relative de ces deux types d'enquête. Dans cet article, nous nous attachons à évaluer la structure d'erreur du TOR de 1986.

Huit sources d'erreur peuvent influer sur les estimations du taux de sous-dénombrement produites par l'EP: l'erreur d'échantillonnage plus sept sources d'erreur non due à l'échantillonnage. Le tableau 1 présente ces sources d'erreur ainsi qu'une évaluation sommaire de leur incidence sur les données du TOR. La seconde colonne du tableau indique l'effet des erreurs sur le taux de sous-dénombrement estimé. Par exemple, si nous corrigeons toutes les "erreurs d'appariement", le taux de sous-dénombrement estimé baissera d'environ un point de pourcentage, soit de 9 à 8%. Certaines catégories d'erreur, comme les "données manquantes" et l'"erreur aléatoire", peuvent contribuer à accroître ou à diminuer le taux de sous-dénombrement mais nous croyons qu'elles n'introduisent pas de biais majeur dans les données du TOR. Les chiffres de la seconde colonne ont été déterminés uniquement en fonction de la source d'erreur correspondante (c'est-à-dire que l'on n'a pas tenu compte de l'effet que pouvaient avoir concurremment les autres sources d'erreur).

Par construction, les huit catégories d'erreur tendent à s'exclure mutuellement et à s'additionner. Il peut y avoir des chevauchements ou des interactions entre les diverses catégories, mais nous croyons qu'ils sont négligeables et nous n'en tenons pas compte ici. Globalement, nous calculons l'effet combiné des erreurs comme suit:

$$(-1.0 - 1.0 - 1.0 + 0.0 - 0.5 + 0.0 + 2.3 + 0.0) \text{ pour cent} = -1.2 \text{ pour cent.}$$

Par conséquent, si nous devions tenir compte de l'effet combiné des erreurs dans l'estimation du taux de sous-dénombrement, celui-ci serait ramené de 9.0% à environ 7.8%. Le chiffre redressé (7.8%) peut être considéré à peu près comme la moyenne d'une distribution a posteriori de l'erreur pour le taux de sous-dénombrement déterminé à l'aide du TOR. Le Bureau of the Census s'emploie actuellement à élaborer une distribution a posteriori de l'erreur complète (voir Mulry et Spencer 1988).

Comme l'estimation initiale du TOR (9%) est beaucoup plus près du pourcentage redressé (7.8%) que celui-ci ne l'est de zéro, nous en concluons que les données originales du TOR reflètent mieux la réalité que les données du recensement.

Dans les huit sections qui vont suivre, nous allons examiner les catégories d'erreur une par une. Chaque section expose les méthodes utilisées dans l'EP de 1980 et les problèmes qui y étaient rattachés, de même que les perfectionnements permettant de réduire la marge d'erreur qui ont été éprouvés dans le TOR. En outre, nous décrivons le mode d'évaluation de chaque source d'erreur et présentons les résultats qui ont été à l'origine de nos conclusions. Finalement, la section 10 présente un résumé des résultats de notre analyse et propose certaines orientations pour la recherche future.

2. ERREUR D'APPARIEMENT

Deux raisons générales peuvent expliquer une mauvaise classification des membres de l'échantillon P:

a) soit que l'information fournie par le répondant ou l'intervieweur est inexacte;
b) soit que l'information fournie est exacte mais utilisée incorrectement.

La première catégorie concerne l'enregistrement de l'adresse déclarée le jour du recensement et la fabrication d'enregistrements dans l'EP, deux sujets traités dans les sections 3 et 4 respectivement. La présente section porte sur les erreurs d'appariement (catégorie b), qui surviennent même lorsqu'il s'agit de personnes véritables et que l'adresse le jour du recensement a été déclarée correctement. Autrement dit, nous parlons ici des erreurs d'appariement attribuables à des erreurs de traitement.

Dans le nouveau plan de sondage de l'EP, l'appariement revêt deux formes: appariement automatisé par lots et appariement manuel assisté par ordinateur. Une personne de l'échantillon P est reconnue comme "non recensée" lorsqu'on a recueilli une quantité suffisante de données pour l'appariement mais qu'il n'est pas possible d'apparier la personne en question à un enregistrement du recensement. Des erreurs surviennent lorsqu'on ne dispose pas de données suffisantes pour l'appariement et que l'on tente néanmoins d'apparier la personne en question ou lorsqu'on recherche parmi les bons questionnaires de recensement sans pouvoir établir une concordance même si la personne a effectivement été recensée.

Il arrive qu'une personne de l'échantillon P soit appariée au mauvais enregistrement. Cela se produit le plus souvent à l'intérieur d'une même famille, où des enfants peuvent avoir des noms et des âges semblables, et dans les quartiers "ethniques", où certains noms sont anormalement répandus. En règle générale, les fausses concordances sont plus rares que les fausses non-concordances car les cas de concordance peuvent être vérifiés facilement par des commis préposés à l'appariement manuel. Les fausses concordances introduisent un biais dans l'estimateur bivalent seulement si le membre de l'échantillon P n'a effectivement pas été recensé.

Un des principaux changements au plan de sondage de l'EP depuis 1980 est l'utilisation d'un même échantillon de blocs pour les échantillons P et D; cela permet un meilleur contrôle de l'erreur d'appariement. L'échantillonnage de blocs permet de classer toutes les personnes enquêtées (échantillons P et D) en trois catégories:

— incluses dans les échantillons P et D
— incluses dans l'échantillon P mais non dans l'échantillon D
— incluses dans l'échantillon D mais non dans l'échantillon P.

Cette classification, qui n'était pas possible selon le plan de 1980, rend le processus d'appariement plus fiable. Par exemple, on peut désormais distinguer les personnes qui ont le même nom dans un quartier ethnique en utilisant tous les renseignements tirés d'un échantillon de blocs. De même, l'échantillon de blocs facilite le débrouillement des erreurs d'adresses. Le choix du bloc de recensement comme unité d'échantillonnage a aussi pour effet de réduire les erreurs de géocodage par rapport à l'EP de 1980, où l'échantillon P était fondé sur des grappes de quatre unités de logement de la CPS et sur la géographie du recensement de 1970.

L'appariement est particulièrement difficile pour les personnes de l'échantillon P qui demeuraient à une autre adresse le jour du recensement (personnes ayant déménagé). Pour cette catégorie de personnes, il est indispensable d'associer l'adresse le jour du recensement déclarée lors de l'EP à la région géographique appropriée avant de procéder à l'appariement. Cette opération posait des problèmes dans l'EP de 1980 et ces problèmes ne sont pas nécessairement résolus avec le nouveau plan de sondage. Toutefois, le Bureau of the Census mettra en application un nouveau système géographique automatisé pour le recensement de 1990 (voir Marx et Saalfeld 1988) et nous espérons que cette innovation permettra d'associer avec célérité et justesse l'adresse des personnes ayant déménagé à la région géographique appropriée.

Lors du TOR de 1986, environ 74% des personnes de l'échantillon P ont été appariées par ordinateur. En outre, 12% des membres de ce même échantillon ont été reconnus comme des cas de "concordance probable" par l'ordinateur. Des commis formés spécialement à cette fin ont passé en revue tous les cas que l'ordinateur n'a pas reconnus comme des cas de "concordance", y compris tous les cas de "concordance probable".

Les résultats de l'EP de 1986 au Mississippi montrent que l'efficacité du système d'appariement informatisé ne se limite pas aux régions urbaines, caractérisées par des numéros civiques, des noms de rue et une géographie bien définie. Dans le test effectué au Mississippi, les adresses étaient composées le plus souvent d'une route rurale et d'un numéro de case postale. Les blocs, de forme irrégulière, étaient définis par des limites invisibles comme un cours d'eau intermittent ou la frontière d'un comté. Néanmoins, l'ordinateur a pu faire l'appariement pour 68% des personnes.

Nous avons réalisé deux études dans le but d'évaluer l'importance de l'erreur d'appariement dans le TOR. Dans la première étude, des spécialistes du bureau chef ont prélevé un sous-échantillon de 35 blocs et procédé à un nouvel appariement. On a effectué cette opération indépendamment de l'appariement initial, puis on a déterminé quels étaient les écarts entre les résultats des deux opérations. À cause de la rigueur avec laquelle s'est fait le nouvel appariement, nous croyons que les résultats de cet appariement reflètent la réalité tandis que les écarts entre les résultats du premier et du second appariement reflètent le biais dont sont entachés les résultats du premier appariement. Seules les personnes n'ayant pas déménagé ont été considérées. En outre, l'étude s'est limité à un nouvel appariement à l'intérieur des blocs, de sorte qu'elle n'a pas permis d'évaluer avec exactitude le nombre de "fausses non-concordances" qui pouvaient être attribuables au fait que l'enregistrement du recensement se trouvait à l'extérieur du bloc visé par l'EP.

Le tableau 2 expose les résultats de l'étude pour l'échantillon P sous la forme d'une table où se recoupent les codes d'appariement obtenus à la suite de l'appariement initial du TOR et ceux obtenus à la suite du nouvel appariement.

On estime que les résultats du TOR renferment environ 88 fausses non-concordances et 18 fausses concordances et que 111 (55 + 56) cas qui ont été classés initialement comme appariés ou non appariés auraient dû être reconnus comme non résolus. Dans le cours normal de l'estimation, les cas non résolus devraient être traités au moyen des méthodes appliquées aux données manquantes (voir Schenker 1988). Il ressort du tableau 2 que le taux d'appariement observé (c'est-à-dire le nombre de cas appariés divisé par le total des cas appariés et des cas non appariés) est de .879 pour l'appariement initial et de .885 pour le nouvel appariement et que, par conséquent, le taux d'appariement initial est entaché d'un biais par défaut d'environ 0.6%.

Tableau 2

Résultats de l'étude sur le nouvel appariement: échantillon (pondéré)[a]

Résultats de l'appariement initial	Résultats du nouvel appariement			
	Recensés	Non recensés	Non résolus	Total
Recensés	16,623	18	55	16,696
Non recensés	88	2,164	56	2,308
Non résolus	17	0	132	149
Total	16,728	2,182	243	19,153

[a] Les données sont pondérées aux totaux de l'échantillon P.

La seconde étude visait à évaluer l'importance de l'erreur d'appariement pour les personnes ayant déménagé. Parmi les personnes qui n'ont pu être appariées dans le TOR, 90 ont déclaré qu'elles avaient déménagé entre le jour du recensement et le moment où s'est déroulée l'EP. Pour les personnes ayant déménagé, la recherche se fait en fonction de l'adresse déclarée le jour du recensement. Pour évaluer la qualité du processus d'appariement, nous avons réexaminé les 90 cas de déménagement à l'aide de méthodes plus rigoureuses. Cet exercice nous a permis de découvrir 11 nouveaux cas de concordance, ce qui a fait passer le taux d'appariement observé pour les personnes ayant déménagé et faisant partie du champ de l'enquête de .661 à .719, pour une augmentation de .058. Bien que le taux de "fausses non-concordances" (11/90 = .122) pour les personnes ayant déménagé soit supérieur à celui observé pour les personnes n'ayant pas déménagé, le premier groupe de personnes représente une proportion relativement faible de l'échantillon P. Si nous corrigeons le biais par défaut dont est entaché le taux d'appariement pour les personnes ayant déménagé et celles n'ayant pas déménagé (5.8 et 0.6% respectivement), nous nous trouvons à réduire de 0.7% le taux de sous-dénombrement du TOR.

Ces calculs ne tiennent pas compte de la possibilité de nouveaux cas de concordance dans l'hypothèse où l'objet de la première étude irait au-delà des limites des blocs visés par l'EP (Thompson, Whitford et Stoudt 1987). Toutefois, si nous nous fondons sur les résultats de l'appariement informatisé pour la région d'essai de Los Angeles, nous en concluons que le codage géographique a été fait correctement et que l'observation de nouveaux cas de concordance pourrait entraîner une baisse additionnelle du taux de sous-dénombrement estimé du TOR d'au plus 0.3%.

3. DÉCLARATION DE L'ADRESSE LE JOUR DU RECENSEMENT

Dans le nouveau plan de sondage de l'EP, nous tentons, comme dans le plan de 1980, d'apparier les personnes de l'échantillon P à l'enregistrement qui correspond à l'adresse le jour du recensement. Pour faciliter l'appariement, l'intervieweur doit demander à quel endroit vivait chaque membre du ménage le jour du recensement. Il cherche ensuite à savoir si ces personnes auraient pu demeuré à une autre adresse ce même jour, par exemple sur un campus de collège ou d'université, sur une base militaire ou à bord d'un bâtiment de la marine ou encore dans une résidence secondaire. Si l'adresse le jour du recensement déclarée incorrectement lors de l'interview de l'échantillon P, il y a alors risque de classer par erreur les membres de ce ménage parmi les personnes non recensées, ce qui aurait pour effet d'introduire un biais par excès dans l'estimation du taux de sous-dénombrement.

Pour évaluer l'importance de l'erreur de déclaration d'adresses, nous avons réinterviewé un sous-échantillon de personnes appariées et des personnes non appariées après que les estimations initiales du taux de sous-dénombrement aient été produites. Cette interview de rappel

Tableau 3

Tailles d'échantillons pour le suivi effectué après la production des estimations initiales

Résultat de l'appariement initial	Concordance totale du ménage	
	dans l'échantillon P	réinterviewés
Concordance totale du ménage	4,662	50
Concordance partielle du ménage	609	50
Non-concordance totale avec élément de contradiction	160	64
Non-concordance totale sans élément de contradiction	357	109

suivait de six mois l'interview initiale de l'EP et de dix mois le recensement. Avant d'exposer les résultats de cette étude, nous signalons ici deux contraintes qu'il faut considérer. La première est le risque d'une plus grande erreur de rappel que dans l'interview originale. La seconde est la possibilité d'effritement de la confiance qu'avait inspirée la campagne de publicité pour le recensement, cette contrainte pouvant constituer un problème sérieux dans les régions qui comptent un grand nombre d'immigrants sans statut légal, qui craignent tout contact avec des représentants de l'État.

Le tableau 3 décrit la composition du sous-échantillon. Dans la plupart des cas, le ménage de l'EP concorde parfaitement avec celui du recensement (concordance totale). Pour les cas de concordance partielle, quelque-uns des membres du ménage de l'EP concordent avec l'enregistrement du recensement, d'autres non.

La catégorie "non-concordance totale du ménage avec élément de contradiction" constitue ce que nous appelons le problème "Dufour-Paradis". La famille Dufour a été recensée à une certaine adresse et le suivi exécuté pour l'échantillon D a confirmé cette adresse. Or, l'interview réalisée pour l'échantillon P a révélé qu'une famille Paradis demeurait à cette adresse le jour du recensement. Il y a là une contradiction et elle s'explique peut-être par le fait que les Paradis ont mal indiqué l'adresse où ils demeuraient le jour du recensement. La catégorie "non-concordance totale du ménage sans élément de contradiction" ne renferme pas de contradictions apparentes; par exemple, l'unité de logement peut avoir été oubliée lors du recensement ou classée dans les logements inoccupés.

Le tableau 4 donne les résultats pour les personnes de l'échantillon en distinguant les personnes appariées initialement des personnes non appariées initialement dans les cas de concordance partielle. Comme prévu, le pourcentage de personnes pour lesquelles l'adresse a été confirmée varie considérablement d'une strate à l'autre. Les adresses ont pratiquement toutes été confirmées en ce qui a trait à la catégorie "concordance totale" tandis que le taux de confirmation le plus faible a été enregistré pour la catégorie "non-concordance totale du ménage avec élément de contradiction". De 13 à 17% des personnes non appariées, selon l'une ou l'autre des trois catégories concernées, ont donné de nouvelles adresses. Chose intéressante, on a indiqué de nouvelles adresses pour 10% des personnes appariées qui sont membres de ménages partiellement appariés, ce qui n'est pas beaucoup moins que le pourcentage correspondant pour les personnes non appariées des mêmes ménages. Il y a peu de chances que la nouvelle adresse soit exacte à moins que l'on ait commis la même erreur lors du recensement et dans l'interview initiale de l'échantillon P. La variabilité des résultats observés nous confirme dans notre opinion qu'un suivi réalisée plusieurs mois après l'interview initiale de l'échantillon P se solde parfois par des réponses différentes (à cause de la crainte et des erreurs de rappel) mais ne permet pas nécessairement d'obtenir une adresse plus juste.

Tableau 4

Résultats du suivi effectué après production des estimations initiales, personnes (non-pondérées)

Résultat	Non-concordance totale du ménage				Concordance partielle du ménage				Concordance totale du ménage	
	avec élément de contradiction		sans élément de contradiction		Personnes non appariées		Personnes appariées			
	#	%	#	%	#	%	#	%	#	%
Adresse confirmée	64	33	252	73	61	75	138	90	164	99
Nouvelle adresse donnée	32	17	46	13	13	16	15	10	1	1
Fabrication possible	70	36	23	7	2	2	0	0	0	0
Non-interview	27	14	24	7	5	6	0	0	0	0
Total	193	100	345	100	81	100	153	100	165	100

Nota: Le symbole # signifie le nombre de personnes faisant partie du sous-échantillon utilisé pour le suivi. Le symbole % signifie le pourcentage du total de colonne.

Certains renseignements nous permettent de croire que 95 cas auraient possiblement inventés lors de l'interview initiale de l'échantillon P. La plupart de ces cas (70) se trouvent dans la catégorie "non-concordance totale du ménage avec élément de contradiction".Ce problème est traité en détail dans la section 4. Par ailleurs, il y a des cas où la réinterview n'a pas été réalisée en entier ou n'a pas produit suffisamment de données pour que l'on puisse classer les personnes dans l'une ou l'autre catégorie. Si l'interview avait été menée correctement, on aurait peut-être relevé une nouvelle adresse pour certains de ces cas.

Si nous pondérons les données du tableau 4 aux totaux de l'échantillon P, nous estimons qu'environ 3.1% des personnes de cet échantillon ont été considérées par erreur comme des personnes n'ayant pas déménagé selon l'interview initiale. En ce qui concerne les personnes qui ont déménagé sans quitter la région d'essai, nous avons pu tenter un appariement à la nouvelle adresse et constaté qu'un tiers des cas avaient été dénombrés lors du recensement d'essai de Los Angeles. Pour évaluer l'effet probable des erreurs de déclaration, surtout si nous considérons le TOR comme le test d'une EP menée à l'échelle nationale, nous supposons que les personnes qui ont indiqué des adresses à l'extérieur de la région d'essai auraient été recensées dans la même proportion que les personnes qui ont indiqué des adresses dans la région d'essai. Ainsi, le tiers des personnes qui ont été considérées par erreur comme des personnes n'ayant pas déménagé auraient été appariées et classées parmi les personnes recensées. Le redressement des estimations pour l'erreur de déclaration a pour effet de réduire de un pourcent l'estimation du taux de sous-dénombrement.

4. FABRICATION DANS L'INTERVIEW DE L'EP

Malgré tout l'effort que l'on met à former et à suivre les intervieweurs, un intervieweur pour l'EP peut parfois inventer un ménage au lieu de réaliser une interview en bonne et due forme. Les enregistrements fabriqués ne peuvent être appariés aux enregistrements du recensement. Ce phénomène contribuera à gonfler l'estimation du taux de sous-dénombrement dans la mesure où les enregistrements fabriqués pour une adresse donnée tiendront la place de personnes qui ont effectivement été recensées.

Le nouveau plan de sondage de l'EP a été conçu de manière à réduire au minimum le taux de fabrication. Il prévoit des contrôles qualitatifs fréquents fondés sur la réinterview. Plusieurs fois par semaine, on prélève des échantillons dans chaque bloc pour vérifier le travail de chaque intervieweur. Une vérification aussi serrée n'était pas possible dans l'EP de 1980 puisque la tâche des intervieweurs était moins concentrée géographiquement. En outre, la formation et la supervision des intervieweurs ont aussi été améliorées depuis ce temps. On a mis en place des programmes de recyclage et des mécanismes de rétroaction pour faire en sorte que les intervieweurs ne répètent pas leurs erreurs.

Deux études ont permis d'évaluer l'importance de la fabrication dans le TOR de 1986. En premier lieu, des contrôles qualitatifs rigoureux ont été effectués pendant la collecte des données pour l'échantillon P; ces contrôles portaient aussi bien sur le listage d'adresses que sur l'interview. Essentiellement, ils n'ont révélé l'existence que d'un petit nombre d'enregistrements fabriqués. Quelques jours après l'interview initiale de l'échantillon P, des commis ont soumis les résultats de 2,070 interviews à un contrôle qualitatif consistant à vérifier la composition des ménages. Sur ces 2,070 interviews, 59 n'ont pas satisfait aux critères du contrôle qualitatif. On a analysé minutieusement ces cas pour déterminer combien d'entre eux étaient le produit d'une fabrication. À cette fin, on a tenté d'apparier chaque membre du ménage à l'enregistrement du recensement en se fondant sur les données fournies par le premier intervieweur (et non sur celles fournies par le commis affecté au contrôle qualitatif); une concordance impliquerait que le premier intervieweur avait recueilli des données justes au sujet de cette personne. Pour invalider cette hypothèse, il faudrait qu'il y ait eu une fabrication identique au moment du recensement. Seulement 13 des 59 cas ont été jugés comme des cas possibles de fabrication; par exemple, aucun membre du ménage figurant sur la liste originale de l'EP n'avait pu être apparié à un enregistrement du recensement. Par conséquent, le taux de fabrication estimé pour le contrôle qualitatif était de 0.6%.

La seconde source de données nous permettant d'évaluer l'importance de la fabrication est le suivi effectué après production des estimations initiales, qui a été décrit dans la section 3. D'après les données du tableau 4, nous estimons qu'environ 1.2% des personnes de l'échantillon P peuvent avoir été le fruit d'interviews fictives. Ce taux de fabrication est environ le double de celui estimé par suite du contrôle qualitatif. Nous croyons qu'une bonne partie de la différence est imputable à un mauvais intervieweur dont les actions ont été dévoilées par l'interview de rappel, après avoir manifestement échappé au contrôle qualitatif. La différence de taux de fabrication s'explique peut-être aussi par le fait que le suivi exagère le degré de fabrication; en effet, des propriétaires et d'autres répondants peuvent nier l'existence de personnes qui occupent des unités de logement transformées sans autorisation ou qui sont au pays sans statut légal.

Afin de déterminer une limite supérieure pour l'effet de la fabrication dans le TOR, nous utilisons le taux de fabrication le plus élevé (.012) et nous supposons que le taux d'appariement pour les personnes de l'échantillon P qui auraient été identifiées à l'aide d'interviews convenables est le même que celui observé pour les cas non fabriqués, soit environ .88. Ces hypothèses permettent de ramener le taux de sous-dénombrement à environ 7.9%, soit à environ 1.1% de moins que le taux initial de 9%. Si on utilisait le taux de fabrication le moins élevé (.006), on obtiendrait par les mêmes calculs un taux de sous-dénombremnent de 8.4%, soit environ .6% de moins que le taux enregistré dans le TOR. Le tableau 1 indiquait un effet de 1% pour la fabrication, ce qui correspond à peu près à la limite supérieure calculée ci-dessus.

5. DONNÉES MANQUANTES

Si nous voulons mesurer correctement les faibles erreurs de couverture, l'EP doit produire un ensemble de données aussi complet que possible, qui ne présente pas une forte proportion de données manquantes. Malheureusement, l'enquête de 1980 était caractérisée par une très forte proportion de données manquantes (U.S. Bureau of the Census 1988). Les modifications apportées au plan de sondage pour l'EP devraient désormais permettre de réduire cette proportion.

Tableau 5

Taux de données manquantes des EP (%)

Source	EP de 1980		TOR de 1986
	Avril	Août	
Échantillon P			
Non-interview (ménage)	4.4	5.3	0.5
Code de dénombrement indéterminé (personne)	4.0	4.4	0.8
Total	8.4	9.7	1.3
Interview par procuration (ménage)	a	a	3.2
Échantillon D			
Non-interview (ménage)	1.1	1.1	SO
Code géographique indéterminé (ménage)	1.6	1.6	SO
Code de dénombrement indéterminé (personne)	2.0	2.0	4.7
Total	4.7	4.7	4.7

[a] Pourcentage inconnu.
NOTA: S.O. signifie "sans objet".

Premièrement, en raison du délai serré prévu pour les interviews de la CPS, les interviews initiales pour l'échantillon P de 1980 ont été réalisées sur une période d'une semaine. Le nouveau plan de sondage pour l'EP prévoit une période d'interview de trois semaines plus une semaine additionnelle s'il survient des difficultés particulières. La prolongation de la période d'interview a pour effet de réduire le taux de non-interview des ménages. On a également réussi à réduire le taux de non-interview des ménages en utilisant un échantillon de blocs de recensement au lieu des grappes de quatre unités de logement tirées d'une liste comme dans le cas de la CPS. L'échantillon de blocs permet à l'intervieweur de visiter une unité de logement à plusieurs reprises (peut-être entre deux visites d'unités de logement situées dans le même bloc) sans que cela n'entraîne des frais de déplacement trop élevés.

Les interviews de rappel incomplètes expliquent une bonne partie des codes de dénombrement manquants pour l'échantillon P dans l'EP de 1980 (2.6% pour avril et 2.8% pour août). Nous tentons actuellement d'amenuiser ce problème en recueillant durant l'interview initiale les données qui permettront de déterminer si une personne a été recensée ou manquée, éliminant ainsi, dans la plupart des cas, la nécessité d'un suivi. En outre, l'amélioration de la rapidité et de la qualité de l'appariement, grâce au nouveau système d'appariement informatisé, aura pour effet de réduire le nombre de cas exigeant un suivi.

Selon le nouveau plan de sondage de l'EP, les échantillons P et D se chevauchent, ce qui fait que la plupart des renseignements nécessaires à la détermination des codes de dénombrement pour l'échantillon D sont recueillis tôt dans le processus, soit durant l'interview initiale de l'échantillon P. L'utilisation d'un échantillon de blocs et l'application d'une géographie du recensement améliorée permettent aussi de réduire la proportion de cas de l'échantillon D pour lesquels on ne peut évaluer le degré d'exactitude du géocodage de recensement. Enfin, des améliorations ont été apportées au traitement des données manquantes (Schenker 1988).

Comme l'indique le tableau 5, les taux de données manquantes pour l'échantillon P dans le TOR sont très inférieurs aux taux observés dans l'EP de 1980. Le taux de données manquantes

pour l'ensemble de l'échantillon D dans le TOR est égal au taux observé dans l'EP de 1980; ceci est toutefois attribuable à une erreur opérationnelle dans le TOR et des réductions des taux de données manquantes devraient, dans l'avenir, être aussi prononcées que celles observées dans le cas de l'échantillon P.

Même si nous avons obtenu des taux de données manquantes relativement faibles dans le TOR, il importe d'analyser l'effet des données manquantes sur les taux de sous-dénombrement estimés. À cette fin, nous avons calculé plusieurs séries d'estimations du taux de sous-dénombrement pour le TOR au moyen de diverses méthodes de traitement des données manquantes, des cas d'interview par procuration dans l'échantillon P, des cas de déménagement dans l'échantillon P et de certains cas non résolus de l'échantillon D. Voir Schenker (1988) pour une description détaillée des diverses estimations, qui varient de 7.8% à 9.4%. Deux des méthodes de traitement considérées dans Schenker (1988) visent à résoudre des problèmes que nous analysons ailleurs dans cet article; il s'agit du traitement des personnes qui ont déménagé sans quitter la région d'essai (sections 2 et 3) et du traitement des cas résolus de l'échantillon D qui peuvent constituer des enregistrements fictifs (section 6). Dans le tableau 1, les effets de ces méthodes de traitement sont attribués à d'autres sources d'erreur que les données manquantes et ils expliquent en majeure partie l'écart entre le taux de sous-dénombrement estimé du TOR (9%) et la plus faible estimation obtenue par Schenker (7.8%). Lorsque nous considérons les autres méthodes de traitement analysées dans Schenker (1988), la différence dans le taux de sous-dénombrement estimé varie de -0.3% à 0.3%. Il s'agit de différences relativement faibles pour lesquelles il est difficile de dire dans quel sens est orienté l'effet. C'est pourquoi nous avons indiqué dans le tableau 1 un effet moyen de 0.0% pour les données manquantes.

6. ERREUR DANS L'ESTIMATION DU NOMBRE D'ENREGISTREMENTS ERRONÉS

Pour estimer l'erreur de couverture nette, il faut estimer le nombre d'enregistrements erronés (EE) créés lors du recensement. Les EE comprennent les catégories suivantes:

i) fabrication dans le recensement, par laquelle le recenseur invente des personnes au lieu de réaliser une interview en bonne et due forme ou par laquelle le répondant fournit le nom de personnes fictives;
ii) répétition d'enregistrements;
iii) personnes nées après le jour du recensement et personnes décédées avant ce jour;
iv) personnes recensées avec si peu de renseignements qu'il n'est pas possible de les apparier aux enregistrements de l'EP.

Toutes ces catégories sont estimées au moyen de l'échantillon D. En outre, certaines erreurs de géocodage dans le recensement sont considérées comme des enregistrements erronés; cette question s'inscrit dans le problème de l'équilibrage analysé dans la Section 7.

Dans l'EP de 1980, l'échantillon D était un échantillon distinct et indépendant de 110,000 ménages recensés. Les intervieweurs ont visité de nouveau les unités de logement huit mois après le jour du recensement pour vérifier si les enregistrements d'alors étaient corrects ou erronés. De plus, on a situé chaque unité de logement sur une carte pour vérifier si on lui avait attribué le bon code géographique et des commis ont examiné les dossiers du recensement pour déceler les enregistrements répétés.

Depuis 1980, l'échantillon D a fait l'objet de deux modifications majeures. Premièrement, comme nous l'avons déjà mentionné, les échantillons P et D reposent désormais sur le même échantillon de blocs. Nous avons constaté que le chevauchement des échantillons P et D a pour effet de réduire les erreurs de codage géographique. En deuxième lieu, la plupart des données de l'échantillon D seront recueillies en juillet, soit à peine trois mois suivant le jour du recensement. Désormais, on peut dire qu'une personne de l'échantillon D a été enregistrée correctement si elle est comptée dans l'échantillon P en juillet et qu'elle peut ensuite être appariée

à l'enregistrement correspondant de l'échantillon D. Les cas non appariés de l'échantillon D doivent faire l'objet d'une interview de rappel qui n'a lieu que six mois après le jour du recensement. Le resserrement des délais d'exécution a pour effet de réduire les taux de données manquantes, d'atténuer le rôle des interviews par procuration et d'améliorer la qualité des données recueillies.

On relève quatre grandes sources d'erreur dans l'estimation du nombre d'enregistrements erronés:

i) erreurs de réponse dans l'interview de l'échantillon D (il s'agit de l'interview de l'échantillon P pour la plupart des cas et de l'interview de rappel pour tous les autres cas), ou erreurs de codage commises par les préposés au traitement des questionnaires;
ii) erreur commise par un intervieweur ou un membre du personnel lors de l'attribution du code géographique du bureau à une personne de l'échantillon D;
iii) erreur dans la recherche d'enregistrements répétés;
iv) erreurs commises en classant une personne de l'échantillon D parmi les cas qui n'offrent pas suffisamment de données pour l'appariement.

Il y a aussi les erreurs dues à la non-réponse dans l'interview de l'échantillon D, comme nous l'avons vu dans la Section 5, et l'erreur d'échantillonnage, que nous verrons dans la Section 9.

Les erreurs de réponse sont souvent liées à l'identification d'un membre de l'échantillon D comme une personne "fictive". Parfois, l'intervieweur constate que la personne qui demeure dans une unité de logement (ou un autre répondant admissible) ne connaît pas les personnes dont le nom figure sur la liste du recensement pour cette adresse. En règle générale, le répondant en question est une personne qui est déménagée à cette adresse après le recensement et qui ne connaît tout simplement pas les personnes qui occupaient le logement le jour du recensement. Ces cas de l'échantillon D devraient être identifiés comme des cas de non-réponse. Toutefois, s'il s'agissait d'enregistrements fabriqués, aucun répondant ne pourrait prétendre connaître les personnes en question.

Au moment de l'expérimentation du nouveau plan de sondage dans le TOR, les intervieweurs affectés à l'échantillon D devaient déterminer si des enregistrements de cet échantillon étaient fictifs et devaient indiquer les motifs de leurs décisions. À l'origine, les commis exigeaient des preuves solides avant de reconnaître une personne de l'échantillon D comme fictive. Ces données ont servi à établir les premières estimations de la population totale et du taux de sous-dénombrement selon le TOR. Nous nous sommes rendu compte plus tard que les règles de codage étaient interprétées trop rigoureusement et nous avons demandé à des spécialistes de réexaminer tous les cas de l'échantillon D qui avaient été classés parmi les cas de "non-interview" (inconnu du répondant) pour déterminer s'il ne s'agissait pas plutôt de cas "fictifs". Sur 257 cas réexaminés, 118 ont été reconnus comme fictifs par les spécialistes. Les données corrigées ont servi par la suite à établir de nouvelles estimations pour le TOR (Schenker 1988).

En règle générale, le codage géographique des questionnaires du recensement a été reconnu comme très bon dans la région d'essai de Los Angeles, qui était un ancien quartier formé de grands blocs bien circonscrits. Nous n'avons pas évalué formellement l'effet des erreurs de codage géographique sur le nombre estimé d'enregistrements erronés mais nous croyons que ces erreurs sont négligeables. Dans d'autres régions des É.-U. toutefois, ce genre d'erreur pourrait ne pas être négligeable en raison de cartes de mauvaise qualité, d'adresses incomplètes ou d'une confusion à propos de la position géographique crée par de nouvelles constructions.

Par exemple, contrairement au test de Los Angeles, le test du Mississippi pour 1986 a donné lieu à des problèmes de codage géographique. Nous avons constaté dans ce test que 2.22% des enregistrements de l'échantillon D étaient répétés. De ce groupe, 35% se trouvaient à l'extérieur du bloc échantillonné.

Bien que nous ayons pu trouver de nombreux enregistrements répétés à l'extérieur de l'îlot échantillonné, nous ne sommes pas sûrs de les avoir tous trouvés car la recherche de tels enregistrements ne constituait pas une activité distincte. Nous nous sommes contentés de relever

Tableau 6

Résultats de l'étude sur le nouvel appariement: échantillon D (pondéré)[a]

Résultats initiaux	Résultats du nouvel appariement			
	Enregistrement correct	Enregistrement erroné	Cas non résolu	Total
Enregistrement correct	19,153	28	88	19,269
Enregistrement erroné	41	283	1	325
Cas non résolu	140	100	223	463
Total	19,334	411	312	20,057

[a] Les données sont pondérées aux totaux de l'échantillon D.

les enregistrements répétés durant l'exécution d'autres opérations de l'EP et par conséquent, plusieurs ont dû échapper à notre attention. Dans la prochaine EP, une activité spéciale sera prévue pour déceler les enregistrements répétés.

Il arrive que le recensement fournisse si peu de renseignements sur une personne que même si celle-ci subissait une interview en bonne et due forme dans l'échantillon P, il ne serait pas possible de l'apparier à un enregistrement de l'échantillon D. Pour pallier à cette difficulté, il conviendrait de classer ces cas dans les EE de manière à obtenir une estimation juste de la population totale. Ce problème se compare au problème de l'équilibrage géographique analysé dans la section 7. Dans l'EP de 1980, il était très difficile d'être conséquent à cause du non-chevauchement des échantillons D et P; des cas similaires ont été classés comme "non appariable" dans l'échantillon D et comme "appariable" dans l'échantillon P, ce qui a introduit un biais dans l'estimateur dual. Comme les échantillons P et D se chevauchent dans le nouveau plan de sondage de l'EP, nous voyons à ce que des règles identiques s'appliquent dans les deux cas, ce qui élimine le biais.

Lors d'une autre évaluation du TOR, ainsi que dans le cadre de l'étude sur le nouvel appariement dont il a été question plus haut (voir section 2), des spécialistes du bureau chef ont réexaminé les cas de l'échantillon D contenus dans un sous-échantillon de 35 blocs. Comme dans la section 2, nous avons effectué le nouvel appariement indépendamment de l'appariement initial et nous nous sommes ensuite prononcé sur les cas où il y avait des écarts entre les résultats des deux opérations. En définitive, nous croyons que les résultats du nouvel appariement reflètent le plus fidèlement possible les codes de dénombrement réels des personnes de l'échantillon D tandis que les écarts entre les résultats de l'appariement initial et ceux du nouvel appariement peuvent être considérés comme une mesure du biais attribuable à l'erreur dans l'opération initiale.

Les résultats pertinents figurent dans le tableau 6. Il convient de souligner que la plupart des changements ont porté sur des cas qui avaient été classés initalement parmi les cas "non résolus". Bon nombre de ces cas sont ceux que nous avons analysés plus haut, c'est-à-dire ceux qu'il fallait classer soit parmi les cas "fictifs" ou soit parmi les cas de "non-réponse". Les données du tableau 6 nous portent à conclure que de meilleures procédures s'imposent pour identifier les cas de l'échantillon D comme fictifs. Nous travaillons actuellement à l'élaboration de nouvelles procédures qui seront mises en application dans la prochaine EP du Bureau of the Census, qui sera combinée avec une répétition générale du recensement de 1990 prévue pour 1988.

D'après l'étude sur le nouvel appariement, nous croyons que le taux initial d'EE,

$$\frac{325}{325 + 19{,}269} = .016$$

devrait être porté à environ

$$\frac{411}{411 + 19{,}334} = .021.$$

Par conséquent, le taux de sous-dénombrement calculé initialement dans le TOR devrait être abaissé d'environ 0.5 point de pourcentage, ce qui donnerait un taux de sous-dénombrement estimé d'environ 8.5%.

7. CONCILIATION DES TAUX DE SUR-DÉNOMBREMENT ET DE SOUS-DÉNOMBREMENT BRUTS

Pour estimer l'erreur de couverture nette, les méthodes et principes servant à mesurer le taux de sur-dénombrement brut doivent être compatibles avec les méthodes et principes utilisés pour mesurer le taux de sous-dénombrement brut. C'est ce que nous appelons la condition d'"équilibrage". Nous décrivons ci-dessous en termes élémentaires comment réaliser l'équilibre dans l'EP.

Une façon d'aborder le problème est d'envisager l'estimateur dual sous la forme

$$\hat{N}_{++} = (\hat{N}_{1+}\hat{N}_{+1})/\hat{N}_{11},$$

où

$$\hat{N}_{11} = M,$$

le nombre pondéré de personnes de l'échantillon P qui ont pu être appariées, et

$$\hat{N}_{+1} = N_p,$$

le nombre pondéré de personnes comprises dans l'échantillon P. On trouvera la définition de tous les symboles dans Diffendal (1988).

Comme nous ne pouvons scruter tous les questionnaires du recensement, le nombre observé de personnes appariées (M) sera inférieur au nombre réel de concordances dans les deux systèmes. Pour maintenir les coûts à un niveau acceptable, nous limitons l'appariement pour un cas donné à un "secteur de recherche", constitué habituellement du bloc échantillonné et d'un ou deux anneaux de blocs situés autour de celui-ci.

En conséquence, le terme $\hat{N}_{11}$ est l'estimateur de kN^*_{11}, où $0 \leq k \leq 1$ est la probabilité conditionnelle qu'une personne recensée soit comptée dans le bon secteur de recherche, et N^*_{11} est l'estimateur que l'on obtiendrait pour N_{11} dans l'EP si on pouvait exécuter la recherche dans l'ensemble de la population.

Pour construire un estimateur convergent de la taille de la population, nous devons réduire le nombre de personnes recensées par le facteur k. Comme la recherche d'enregistrements erronés pour l'échantillon D (par ex., enregistrements répétés) se fait dans le secteur de recherche et non dans l'ensemble de la population, le terme $\hat{N}_{1+}$ est l'estimateur de $k\,N^*_{1+}$, où N^*_{1+} est l'estimateur que l'on obtiendrait pour N_{1+} si la recherche d'enregistrements erronés pouvait se faire dans l'ensemble de la population.

Si nous supposons que les secteurs de recherche sont uniformes, l'ED devient un estimateur convergent de N_{++}. Signalons que dans ce modèle d'équilibrage, nous n'estimons pas la probabilité k; nous comptons plutôt sur des secteurs de recherche uniformes pour retrancher cette probabilité de l'ED.

Dans l'EP de 1980, il était impossible d'équilibrer les échantillons P et D parce qu'ils ne se chevauchaient pas. Les adresses de la CPS (ou de l'échantillon P) étaient codées suivant la

géographie du recensement. Le secteur de recherche devait être limité aux environs immédiats de l'adresse de la CPS mais comme les adresses de cette enquête étaient fondées sur la géographie du recensement de 1970, il était difficile de les coder suivant la géographie du recensement de 1980 et la recherche se trouvait être effectuée dans un plus grand secteur. Comme le secteur de recherche pour l'échantillon P s'était élargi, celui pour l'échantillon D devait aussi s'élargir. Nous croyons que cela a engendré des incohérences entre les secteurs de recherche respectifs et, par conséquent, introduit un biais dans l'ED.

Dans le TOR, nous avons exécuté le double appariement entre les personnes des échantillons P et D à l'intérieur des blocs choisis. Tout le long de l'appariement informatisé et de l'appariement manuel, la géographie et les secteurs de recherche étaient uniformes, bien circonscrits et parfaitement sous contrôle. C'est pourquoi le problème d'équilibrage n'a pas introduit de biais notable dans les résultats du test de Los Angeles.

8. BIAIS DE CORRÉLATION

Pour que l'estimateur dual soit un estimateur convergent de la taille réelle de la population N_{++}, nous devons poser deux hypothèses d'indépendance:

i) indépendance des enquêtes;
ii) indépendance des probabilités de sélection vis-à-vis des caractéristiques des personnes.

De plus, on suppose souvent l'indépendance des probabilités de sélection des personnes mais si cette hypothèse n'est pas respectée, cela n'a pas pour effet de fausser sensiblement l'estimation de la population totale. (Wolter 1986b et Cowan et Malec 1986).

L'hypothèse de l'indépendance des enquêtes ne tient plus lorsque la fréquence de saisie d'une personne pour le recensement influe sur la probabilité de saisie de cette personne dans l'EP. L'estimateur N_{++} souffre d'un biais par défaut lorsque la probabilité de saisie dans l'EP augmente du fait que la personne a été saisie dans le recensement, et d'un biais par excès lorsque la probabilité de saisie dans l'EP diminue pour la même raison.

Les données de l'EP d'avril 1980 renferment peut-être un biais apréciable du fait que l'hypothèse de l'indépendance des enquêtes n'a pas été respectée. Cela est dû au fait que les répondants pourraient avoir confondu la CPS de mars ou d'avril avec le recensement.

L'hypothèse de l'indépendance des probabilités de sélection vis-à-vis des caractéristiques des personnes ne tient plus lorsque les probabilités de saisie dans le recensement varient d'une personne à l'autre. Le biais qui en découle (appelé biais d'hétérogénéité ou biais de corrélation) est habituellement considéré comme un biais par défaut puisque les personnes qui ont une forte probabilité de saisie dans le recensement tendent aussi à avoir une forte probabilité de saisie dans l'EP et inversement, les personnes qui ont une faible probabilité de saisie dans le recensement ont aussi tendance à avoir une faible probabilité de saisie dans l'EP.

Sekar et Deming (1949) ont proposé la post-stratification comme moyen de réduire le biais d'hétérogénéité. Dans la pratique, toutefois, il est peu probable que cette méthode s'avère parfaitement efficace; il subsiste inévitablement des différences de probabilités de saisie à l'intérieur des post-strates.

Dans le modèle dual, on estime le nombre de personnes qui ont été manquées par les deux systèmes (N_{22}) au moyen de l'expression suivante:

$$\hat{N}_{22} = \hat{N}_{12}\hat{N}_{21}/\hat{N}_{11},$$

qui correspond à l'équation (2) dans Diffendal (1988). Comme l'estimateur dual peut être exprimé sous la forme

$$\hat{N}_{++} = \hat{N}_{11} + \hat{N}_{12} + \hat{N}_{21} + \hat{N}_{22},$$

Tableau 7

Taux de sous-dénombrement (%) pour les personnes de race noire et pour l'ensemble de la population dans les recensements de 1950, 1960 et 1980 aux É.-U., et écarts entre ces taux

Source	Personnes de race noire	Population totale	Écart
1950			
EP	3.2	1.4	1.8
AD	9.6	4.4	5.2
1960			
EP	3.8	1.9	1.9
AD	8.3	3.3	5.0
1980			
EP[a]			
Bas	1.1	−1.0	2.1
Moyen	6.9	1.4	5.5
Élevé	5.7	2.1	3.6
AD	5.9	1.4	4.5

[a] L'EP de 1980 a produit 12 séries d'estimations. Les trois estimations présentées ici pour chaque catégorie sont tirées respectivement de la série la plus faible, de la série moyenne et de la série la plus élevée par rapport au taux de sous-dénombrement total estimé.

et que $\hat{N}_{11}$, $\hat{N}_{12}$ et $\hat{N}_{21}$ sont des estimateurs directement basés sur le plan de sondage, le biais qui découle du non-respect des hypothèses d'indépendance se retrouve uniquement dans $\hat{N}_{22}$, l'estimateur de N_{22}.

Nous pouvons analyser le biais de corrélation présent dans le recensement de 1980 et des recensements antérieurs en comparant N_{++} à des estimations indépendantes de la population totale tirées d'une analyse démographique (AD). Le tableau 7 contient les données pertinentes tirées de recensements récents. Si l'on considère les estimations de l'analyse démographique comme valeurs de référence, la comparaison expose le biais total dont est entaché l'estimateur dual, notamment le biais de corrélation et les autres sources d'erreur. Nous croyons que le biais par défaut exposé dans ces estimations est largement imputable au biais de corrélation. L'EP de 1950 a sous-estimé sérieusement la taille de la population, le taux de sous-dénombrement et l'écart entre le taux de sous-dénombrement pour la population totale et celui pour les personnes de race noire à cause probablement du biais d'hétérogénéité et du biais découlant de la dépendance des enquêtes. Il convient toutefois de souligner que si les données de l'EP de 1950 avaient servi à redresser les chiffres du recensement de la même année, l'écart entre les taux de sous-dénombrement des deux groupes serait passé de 6.4% à environ 4.6%.

L'EP de 1960 a sous-estimé tout aussi sérieusement la taille de la population, le taux de sous-dénombrement et l'écart entre les taux de sous-dénombrement des deux groupes à cause probablement du biais de corrélation. Si les données de l'EP de 1960 avaient servi à redresser les chiffres du recensement de cette même année, l'écart entre les taux de sous-dénombrement serait passé de 5.3% à environ 3.4%.

Il n'y a pas eu d'enquête post-censitaire en 1970. Celle de 1980 a produit 12 séries d'estimations du taux de sous-dénombrement fondées sur les résultats des sondages d'avril et d'août et sur diverses séries d'hypothèses. Les taux de sous-dénombrement calculés au moyen de l'AD se situent à peu près au centre des séries d'estimations produites par l'EP. Le biais de corrélation n'est pas aussi évident ici que dans les EP de 1950 ou de 1960 à cause principalement des améliorations qui ont été apportées en 1980 dans le but de réduire la dépendance positive des enquêtes. Nous croyons qu'il existe encore un biais d'hétérogénéité mais que celui-ci est masqué par d'autres erreurs de l'EP et un biais imputable à une dépendance négative des enquêtes.

Dans le nouveau plan de sondage de l'EP, nous tentons de réduire au minimum le biais dû à la dépendance des enquêtes en fixant la date du début des opérations de l'EP après la date de la fin des principales opérations sur le terrain pour le recensement. Contrairement à ce qui s'est fait lors de l'EP d'avril 1980, cette façon de procéder favorisera le plus possible l'indépendance entre le recensement et l'EP. En outre, nous appliquons désormais dans nos bureaux régionaux des procédures qui favorisent une telle indépendance; par exemple, pour l'EP les intervieweurs sont affectés à une région différente de celle où ils ont travaillé (si c'est le cas) durant le recensement.

Il sera difficile de supprimer le biais de corrélation dû à l'hétérogénéité dans les futures EP. Les seules solutions qui s'offrent à nous sont une post-stratification plus efficace et une combinaison quelconque des données de l'EP et de l'AD, peut-être sur la base des quotients hommes/femmes établis par l'AD. Voir Wolter (1986c) et Choi, Steel et Skinner (1988). Au cours de la dernière décennie, nous avons expérimenté diverses méthodes de post-stratification, dont celles faisant intervenir des variables comme le mode d'occupation, le pourcentage de questionnaires du recensement retournés par la poste et l'état matrimonial. Ces méthodes présentent un certain intérêt pour l'avenir. Voir Diffendal (1988).

Les écarts de taux de sous-dénombrement observés dans le TOR sont conformes aux prévisions. Dans le recensement de la population des É.-U., les hommes ont habituellement un taux de couverture plus faible que celui des femmes. C'est ce que les résultats de l'analyse démographique indiquent constamment. Les quotients de masculinité (nombre d'hommes pour 100 femmes) observés dans le TOR sont supérieurs à ceux observés dans le recensement pour les personnes d'origine hispanique et celles qui ne sont ni d'origine hispanique ni d'origine asiatique. De même, pour le groupe de personnes de 30 à 44 ans, les quotients de masculinité sont beaucoup plus élevés dans le TOR que dans le recensement (de 1.1 à 3.4 hommes de plus pour 100 femmes). Ces résultats sont conformes à ceux de l'analyse démographique de 1980 pour l'ensemble du pays. Par conséquent, nous sommes d'avis que les quotients de masculinité observés dans le TOR sont ceux qui se rapprochent le plus des quotients réels et malgré que le gain soit limité par le biais de corrélation, l'EP permet encore d'évaluer l'écart des taux de sous-dénombrement.

Le tableau 8 donne sous forme de table de contingence les résultats du TOR de 1986 sans post-stratification. Le nombre estimé de personnes qui n'ont été recensées par aucun des deux systèmes,

$$\hat{N}_{22} = 5{,}870,$$

est du même ordre de grandeur que le nombre de substitutions dans le recensement (5,259) et le nombre d'enregistrements erronés (6,426). Environ le huitième du nombre estimé de personnes manquées dans le recensement ($\hat{N}_{12} + \hat{N}_{22} = 44{,}373$) est rattaché à la case (2,2). Par conséquent, le chiffre du sous-dénombrement peut être attribué en majeure partie à l'estimation pure et simple et non au modèle dual.

Afin d'illustrer l'effet du biais de corrélation, doublons l'effectif de la case (2,2). Cela a pour effet de hausser le taux de sous-dénombrement estimé d'environ 1.4%. À la suite d'une analyse de l'EP de 1980, Ericksen et Kadane (1985) proposent de multiplier l'effectif de la case (2,2) par un facteur de 2.7, ce qui a pour effet de hausser le taux de sous-dénombrement estimé de 2.3%.

D'autres renseignements nous éclairent sur la question du biais de corrélation. En effet, trois anthropologistes ont prêté leur concours au U.S. Bureau of the Census pour agir à titre de participants ou d'observateurs dans le test de Los Angeles. Leurs observations ne permettent pas de mesurer directement le biais de corrélation mais elles nous révèlent jusqu'à quel point les catégories de personnes manquées dans les recensements et les EP se ressemblent. Selon le rapport des anthropologistes, des personnes pour lesquelles la probabilité de saisie est très faible tendraient à être manquées dans le recensement comme dans l'EP et par conséquent, les données du TOR pourraient être entachées d'un biais par défaut appréciable. Voir Hainer et coll. (1988) et Hines (1988).

Tableau 8

Estimations duales pour le recensement d'essai de 1986 à Los Angeles

		EP		
		Dénombrés	Manqués	Total
Enregistrements corrects du recensement[a]	Dénombrés	298,204	45,463	343,667
	Manqués	38,503	5,870	44,373
	Total	336,707	51,333	388,040

[a] Enregistrements corrects du recensement = total des enregistrements du recensement − substitutions − enregistrements erronés.

Compte tenu des données disponibles, nous n'avons pas de moyen précis pour évaluer l'importance du biais de corrélation contenu dans les données du TOR. Néanmoins, si nous nous fondons sur les travaux évoqués ci-dessus, nous avançons que le taux de sous-dénombrement déterminé par le TOR devrait être d'au moins 2.3% de plus.

9. ERREUR ALÉATOIRE

L'erreur d'échantillonnage influe sur les estimations du nombre de cas appariés, du nombre d'enregistrements erronés et des totaux de l'échantillon P. Les chiffres du recensement et le nombre de substitutions de personnes dans le recensement reposent sur un dénombrement complet de la population (100%) et ne sont donc pas touchés par l'erreur d'échantillonnage. L'écart type estimé pour le taux de sous-dénombrement est 0.007. Ainsi, un intervalle de confiance de 95% pour le taux de sous-dénombrement est $.09 \pm 2(.007) = (.076, .104)$ si l'on suppose une distribution normale.

Diffendal (1988) présente les erreurs types estimées pour les facteurs de redressement du TOR définis par $Y = \hat{N}_{++}/CEN$ et utilise un modèle axé sur les composantes de la variance pour lisser les valeurs de Y dans le but d'atténuer les effets de l'erreur d'échantillonnage. Dans la plupart des cas, le lissage a réduit de façon substantielle les erreurs types estimées, particulièrement en ce qui a trait aux domaines (ou petites régions). Nous croyons que cette forme de lissage pourra être utilisée de façon profitable dans les futures EP.

10. CONCLUSION

Après le recensement de 1980, le U.S. Bureau of the Census a réexaminé son programme d'évaluation du taux de couverture et en a relevé les lacunes. Nous avons alors mis sur pied un programme de recherches puis avons élaboré une nouvelle méthode d'évaluation du taux de couverture destinée à réduire ces lacunes. Au cours des huit dernières années, le nouveau plan de sondage de l'EP a fait l'objet de tests importants qui ont démontré sa grande supériorité par rapport au plan de l'EP de 1980.

Cet article nous a donné l'occasion d'analyser les résultats de notre programme de recherches dans la perspective du TOR de 1986. On ne pourra peut-être jamais avoir une EP parfaite. Cependant, le nouveau plan de sondage ne renferme pas de lacune assez grande pour réfuter les résultats de l'EP. Pour les raisons énoncées dans la section 1, nous croyons que l'effet combiné des erreurs dans l'estimation du taux de couverture dans le TOR est moindre que l'erreur contenue dans le recensement à Los Angeles.

Un des principaux avantages du TOR est d'attirer notre attention sur de nouveaux sujets et des problèmes mineurs qui justifient une recherche plus poussée. Par exemple, l'interview

initiale de l'EP visait entre autres à recueillir les données qui devaient permettre de dire si une personne de l'échantillon P avait été manquée lors du recensement. Nous travaillons actuellement à rendre le questionnaire plus révélateur en y incorporant d'autres questions de sélection qui visent à mieux cerner les personnes ayant déménagé. Les futures enquêtes post-censitaires comporteront également des interviews de suivi pour la plupart des personnes ayant déménagé et pour les ménages de l'échantillon P qui ont déclaré ne pas avoir déménagé mais pour qui, selon certains indices, ce ne serait pas le cas. Nous croyons que ces interviews de suivi seront un moyen de réduire au minimum le nombre de déclarations fautives.

Il faut continuer d'améliorer et de tester les procédures de contrôle qualitatif qui visent à déceler les cas de fabrication dans l'EP et à en corriger les effets. Le contrôle qualitatif devrait permettre de vérifier non seulement les noms figurant sur la liste de l'EP, mais aussi d'autres éléments d'information. Par ce contrôle, nous devrions pouvoir déceler les enregistrements que l'on a fabriqués partiellement en lisant le nom sur la boîte aux lettres ou en l'obtenant du propriétaire pour ensuite fabriquer les caractéristiques. Nous sommes aussi à modifier les formulaires de suivi de l'EP pour faciliter la détection de personnes fictives.

Dans les futures enquêtes post-censitaires, nous chercherons spécialement à réduire au minimum le nombre de données manquantes en tâchant particulièrement d'éliminer la nécessité d'un suivi. Par ailleurs, plus il y aura de cas qui devront faire l'objet d'un suivi, plus grande sera la proportion de cas qui n' auront pu être résolus. Il est donc nécessaire de trouver une façon de traiter convenablement ces cas.

Malgré les résultats acceptables qu'a produit le TOR, il convient d'examiner tous les aspects de la question avant de conclure que les résultats de l'EP de 1990 seront plus près de la réalité que les résultats du recensement de cette même année. Le niveau de sous-dénombrement net observé dans le test de Los Angeles était élevé par rapport à ce que l'on devrait observer dans un recensement à l'échelle du pays. Une EP nationale aura-t-elle un degré d'erreur suffisamment faible pour produire des estimations démographiques plus précises?

Nous pensons que le recensement de 1990 révèlera des secteurs qui auront un fort taux de sous-dénombrement ou, peut-être, de sur-dénombrement malgré un faible taux de sous-dénombrement net à l'échelle nationale. L'EP devrait donc produire des estimations démographiques plus justes pour les secteurs où la population est la plus difficile à dénombrer. Si, d'ici deux ans, nous perfectionnons la nouvelle version de l'EP, nous pourrions peut-être aussi obtenir des estimations démographiques plus justes pour les secteurs où la population est moins difficile à recenser.

Nous pensons aussi que l'EP comportera moins d'erreurs si le taux de sous-dénombrement est moins élevé. Les secteurs stables, caractérisés par des cartes de bonne qualité, des adresses clairement définies, un faible nombre de personnes ayant déménagé et des répondants coopératifs, causeront relativement peu de problèmes tant pour le recensement que l'EP. Les erreurs de traitement résiduelles pourraient créer un seuil de précision au-delà duquel l'EP ne pourrait aller, peu importe le taux de sous-dénombrement net réel. Nous ne pourrons vérifier cela qu'à l'exécution de l'EP de 1990. Il pourrait alors y avoir des secteurs pour lesquels les estimations de l'EP seraient plus justes que les estimations du recensement et d'autres (la plupart de ceux qui restent) pour lesquels elles seraient aussi justes sinon presque aussi justes que les estimations du recensement. La théorie statistique devrait nous permettre de trouver un moyen de produire une meilleure estimation en combinant les résultats du recensement avec ceux de l'EP.

REMERCIEMENTS

Cet article est le fruit du travail de nombreuses personnes à part les auteurs; ceux-ci tiennent à remercier Dan Childers, Carol Corby, Gregg Diffendal, Charisse Jeffries, Arona Pistiner, Nathaniel Schenker, Maria Urrutia et Kirsten West. Les auteurs veulent aussi exprimer toute leur gratitude aux trois arbitres pour leurs nombreux commentaires utiles.

BIBLIOGRAPHIE

ANOLIK, I. (1988). The rural post-enumeration survey in east central Mississippi. Statistical Research Division Report, Series RR 88/10. U.S. Bureau of the Census, Washington, D.C.

CHOI, C.Y., STEEL, D.G., et SKINNER, T.J. (1988). Adjusting the 1986 Australian Census for under-enumeration. *Proceedings of the Census Bureau Fourth Annual Research Conference*. Bureau of the Census, Washington, D.C.

CITRO, C.F., et COHEN, M.L. (1985). *The Decennial Census: New Directions for Methodology in 1990*. Washington: National Academy Press.

COWAN, C.D., et MALEC, D. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.

DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County. *Techniques d'enquête*, 14.

ERICKSEN, E.P., et KADANE, J. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-114.

FAY, R.E., PASSEL, J.S., ROBINSON, G., et COWAN, C.D. (1988). The coverage of population in the 1980 Census. Technical report PHC 80-E4. Bureau of the Census, Washington, D.C.

HAINER, P., HINES, C., MARTIN, E., et SHAPIRO, G. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.

HINES, C. (1988). The role of participant observation research in understanding the census undercount. Paper presented at the Population Association of America Annual Meetings, New Orleans, La.

JARO, M. (1988). Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* (à paraître).

MARX, R.W., et SAALFELD, A.J. (1988). Programs for assuring map quality at the Bureau of the Census. *Proceedings of the Bureau of the Census Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.

MULRY, M., et SPENCER, B. (1988). Total error in dual system estimates of population size. *Proceedings of the Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.

SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

SCHENKER, N. (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986. *Techniques d'enquête*, 14.

THOMPSON, J.H., WHITFORD, D., et STOUDT, D. (1987). Note de service à Howard Hogan, Suject: Review of 1986 PES Matching, avril 21, 1987.

U.S. BUREAU OF THE CENSUS (1978). *The Current Population Survey: Design and Methodology*, Technical Paper No. 40, Washington, D.C.

U.S. DEPARTMENT OF COMMERCE (1987). Communiqué de presse, Déclaration par Undersecretary Robert Ortner, octobre 30, 1987.

WOLTER, K.M. (1986a). Some coverage error models for census data, *Journal of the American Statistical Association*, 81, 338-346.

WOLTER, K.M. (1986b). A combined coverage error model for individuals and housing units. Statistical Research Division Report Series RR 86/27, U.S. Bureau of the Census, Washington, D.C.

WOLTER, K.M. (1986c). Capture-recapture estimation in the presence of a known sex ratio. Statistical Research Division Report Series RR 86/20, U.S. Bureau of the Census, Washington, D.C.

Techniques d'enquête, juin 1988
Vol. 14, n° 1, pp. 125–143
Statistique Canada

Modélisation de l'erreur d'appariement et son effet sur les estimations de l'erreur d'observation du recensement

PAUL P. BIEMER[1]

RÉSUMÉ

L'efficacité des estimateurs de système dual du sousdénombrement du recensement repose en grande partie sur l'hypothèse selon laquelle des personnes dénombrées lors de l'étude d'évaluation peuvent être appariées de façon précise aux mêmes personnes dénombrées lors du recensement. Or, les erreurs d'appariement et les non-appariements erronés, qui sont inévitables, diminuent l'exactitude des estimateurs. De fait, des études ont démontré que l'ampleur de l'erreur résultante peut être suffisamment importante par rapport à l'erreur d'observation du recensement pour que l'estimation devienne inutilisable. Le présent mémoire a pour objet d'exposer un modèle d'analyse de l'effet de l'erreur d'appariement sur les estimateurs du sous-dénombrement et d'illustrer son utilisation possible dans le cadre du programme d'évaluation du sousdénombrement du recensement de 1990. L'erreur quadratique moyenne de l'estimateur de système dual est d'abord dérivée du modèle proposé et les composantes de l'EQM résultant de l'erreur d'appariement sont définies et expliquées. Nous étudierons ensuite, à la lumière du modèle, l'incidence de l'erreur d'appariement sur l'EQM de l'estimateur du sous-dénombrement du recensement. Enfin, nous illustrerons une méthodologie permettant d'utiliser le modèle pour optimaliser la conception des études d'évaluation de l'erreur d'appariement et donnerons la forme des estimateurs.

MOTS CLÉS: Sous-dénombrement; estimation de système dual; saisie-resaisie; erreur non due à l'échantillonnage; erreur de traitement.

1. INTRODUCTION

Sekar et Deming (1949) ont été les premiers à suggérer l'utilisation des méthodes de saisie-resaisie aux fins de l'évaluation du recensement et de l'enregistrement des naissances et des décès. Pour permettre l'estimation de l'erreur d'observation du recensement, la méthode nécessite l'appariement des personnes dénombrées dans le cadre d'une enquête-échantillon de la population et des personnes dénombrées dans le cadre du recensement en vue de déterminer le nombre de personnes dénombrées dans les deux cas. La mise en oeuvre de la méthode de saisie-resaisie peut se buter à nombre de difficultés susceptibles d'introduire un biais considérable dans l'estimation de la taille de la population totale, N [voir, par exemple, Burnham et collaborateurs (1987) et Wolter (1986)]. Un problème qui se pose assez souvent est l'incapacité d'apparier de façon précise les personnes dénombrées dans le cadre de l'enquêteéchantillon et les personnes dénombrées dans le cadre du recensement. Seltzer et Adlakta (1974) ont démontré que l'erreur d'appariement peut entraîner des biais relatifs de l'ordre de 33 % et qu'elle peut être positive ou négative selon que le nombre des faux non-appariements soit supérieur à celui des faux appariements ou l'inverse (voir aussi, Scheuren et Oh, 1985). Wolter (1983) souligne que l'une des raisons justifiant la non-correction de données du recensement des États-Unis de 1980 était la présomption d'erreurs d'appariement dans le cadre du programme de post-dénombrement de 1980.

Le présent mémoire fournit un cadre de base pour l'évaluation de l'erreur d'appariement afférente aux études de saisie-resaisie (en particulier pour leur application à des populations humaines) et pour l'évaluation de l'incidence de ces erreurs sur l'exactitude de l'estimation de

[1] Paul P. Biemer est chef du Department of Experimental Statistics et directeur du University Statistics Center de l'Université du Nouveau-Mexique, Las Cruces, Nouveau-Mexique, États-Unis.

N. En vue d'asseoir l'étude de l'erreur d'appariement sur une base simple et connue, nous adopterons le modèle original de saisie-resaisie de Sekar-Deming. On trouvera d'autres développements de la technique de Sekar-Deming dans les ouvrages de Marks, Seltzer et Krotki (1974), et Wolter (1986).

Considérons une population U dont la taille est de N. Un recensement est réalisé dans le cadre duquel N_c personnes sont dénombrées. Afin de permettre l'estimation de N-N_c (désigné sous le nom d'erreur d'observation du recensement), laquelle équivaut à l'estimation de N, on réalise une enquête postcensitaire (EP) dont la période de référence est identique à celle du recensement. Supposons : a) que ni le recensement ni l'EP ne comprennent de faux événements (par exemple, des doubles comptes, des falsifications, des personnes hors du champ de l'enquête ou non identifiables) ou que le nombre de ces événements peut être estimé avec exactitude et soustrait de N_c; et b) que le fait d'être dénombré dans le cadre du recensement et dans le cadre de l'enquête postcensitaire constituent deux événements indépendants.

Les personnes dénombrées dans l'EP et les personnes dénombrées dans le recensement sont appariées afin de déterminer le nombre de personnes dénombrées dans les deux cas. Soient x_{11} l'estimateur non biaisé de conception du nombre total des personnes comprises dans les populations de l'EP et du recensement, et N_p, l'estimateur non biaisé de conception de la taille de la population de l'EP. L'estimateur Sekar-Deming (désigné depuis peu sous le nom d'estimateur de système dual ou ESD) de N est

$$\hat{N} = \frac{N_p N_c}{x_{11}} \tag{1}$$

Comme nous le verrons plus loin, deux types d'erreurs peuvent avoir une incidence sur $\hat{N}$: l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage. Bien que l'erreur non due à l'échantillonnage puisse provenir de plusieurs sources, la source qui nous intéresse ici est l'erreur d'appariement; c'est-à-dire l'erreur découlant d'un classement incorrect des personnes dénombrées dans le cadre de l'EP comme ayant été recensées (erreurs positives fausses) ou comme n'ayant pas été recensées (erreurs négatives fausses).

En utilisant les développements en séries de Taylor, on peut dériver les formes générales des moments de $\hat{N}$. Il est possible de démontrer que, pour les termes de l'ordre de $1/n$, où n représente la taille de l'échantillon de l'EP,

$$\text{Biais } (\hat{N}) \doteq -N[\text{Biaisrel } (\hat{p}_{11}) - \text{Varrel } (\hat{p}_{11})] \tag{2}$$

$$\text{x } [1 + \text{Biaisrel } (\hat{p}_{11})]^{-1}$$

et

$$\text{Var } (\hat{N}) \doteq N^2 \text{ Varrel } (\hat{p}_{11}) \; [\; 1 + \text{Biaisrel } (\hat{p}_{11}) \;]^{-2} \tag{3}$$

où $\hat{p}_{11} = x_{11}/N_p$ est un estimateur de p_{11}, la proportion réelle de la population de l'EP dénombrée dans le recensement; Biaisrel $(\hat{p}_{11})$ = Biais $(\hat{p}_{11})/p_{11}$; et Varrel $(\hat{p}_{11})$ = Var$(\hat{p}_{11})\, E^{-2}(\hat{p}_{11})$. Ces équations supposent que N_c, le nombre de personnes recensées, a une variance égale à zéro. Il s'agit bien sûr d'une simplification puisque, comme il a déjà été mentionné, une estimation des faux événements du recensement a pu être soustraite du nombre de personnes recensées afin d'obtenir N_c, et que cette correction peut être sujette à des erreurs d'échantillonnage ou à d'autres erreurs. Néanmoins, l'hypothèse retenue est valable compte tenu de l'accent mis dans le présent mémoire sur l'erreur d'appariement et son incidence sur $\hat{N}$. Une application de cette méthodologie permettant la prise en compte de l'erreur dans l'estimateur N_c est présentée dans la dernière section du mémoire.

À partir des équations (2) et (3), on constate que l'erreur quadratique moyenne totale (EQM) de $\hat{N}$ est fonction de l'EQM totale de $\hat{p}_{11}$. La section suivante sera consacrée à l'étude de certains modèles d'évaluation de l'incidence de l'erreur d'appariement sur $\hat{p}_{11}$. En admettant que j $(j=1,\ldots,n)$ constitue l'indice de la $j^{ième}$ personne de l'échantillon EP, α_j peut être défini comme la probabilité qu'une erreur de classement se produise pour une personne j lors du processus d'appariement et nous pouvons étudier deux hypothèses formulées à l'égard des probabilités α_j.

2. MODELES D'ERREUR D'APPARIEMENT

2.1 Erreur d'appariement non corrélée

Supposons que:

1. L'événement {élément j est mal classé} est indépendant de l'événement {élément j' est mal classé} chaque fois que $j \neq j'$.
2. $\alpha_j = \theta$ si l'élément j a vraiment été recensé, cette probabilité est désignée comme étant la probabilité d'une erreur négative fausse, et $\alpha_j = \phi$ si l'élément j n'a pas vraiment été recensé, cette probabilité est désignée comme étant la probabilité d'une erreur positive fausse.

Afin de fixer les idées, supposons que l'EP a fait l'objet d'un sondage aléatoire simple et que n est petit comparé à N, ainsi

$$E(\hat{p}_{11}) = p_{11}(1-\theta) + (1-p_{11})\phi, \tag{4}$$

$$\text{Biais}(\hat{p}_{11}) = -p_{11}\theta + (1-p_{11})\phi \tag{5}$$

$$\text{Var}(\hat{p}_{11}) = n^{-1} E(\hat{p}_{11})\,(1-E(\hat{p}_{11})) \tag{6}$$

$$= n^{-1}\,(VE + VAS),$$

où VE, la *variance d'échantillonnage*, est obtenue par

$$VE = p_{11}(1-p_{11})\,(1-\theta-\phi)^2 \tag{7}$$

et où VAS, la *variance d'appariement simple*, est obtenue par

$$VAS = p_{11}\theta(1-\theta) + (1-p_{11})\,\phi(1-\phi) \tag{8}$$

(se reporter à l'appendice pour la preuve).

Les lecteurs qui connaissent le modèle d'erreur de réponse de Hansen, Hurwitz, et Pritzker (1964) reconnaîtront la correspondance entre leur variance de réponse simple et la VAS du présent modèle. Hansen et collaborateurs ont défini une mesure I, désignée sous le nom d'"indice d'incohérence [réponse]", comme le rapport entre la variance de réponse simple et la variance totale d'une seule réponse, c'est-à-dire la proportion de la variance qui constitue une variance de réponse. Dans le cas des réponses d'enquête, I est un indicateur de la fiabilité de réponse des informations recueillies lors de l'enquête. Il est possible d'obtenir une mesure analogue pour l'erreur d'appariement afin de déterminer l'incidence du manque de fiabilité de l'appariement sur la variance de $\hat{p}_{11}$. Cette mesure, notée I_M, est calculée à l'aide de la formule suivante:

$$I_M = \frac{VAS}{VE + VAS}. \tag{9}$$

Il est possible que les hypothèses (1) et (2) se révèlent trop restrictives pour certaines applications. Par exemple, l'hypothèse d'indépendance (1) est infirmée lorsque l'élément B de l'EP et l'élément A du recensement sont appariés de façon erronée, ce qui entraîne le classement erroné de l'appariement correct, l'élément A du recensement et l'élément A de l'EP, comme un non-appariement. Étant donné que cette situation implique que les erreurs relatives aux éléments A et B sont corrélées de façon négative, il s'ensuit que $\text{Var}(\hat{p}_{11})$ aura une valeur plus petite que celle calculée selon la formule (6). Cependant, les erreurs corrélées n'ont aucune incidence sur $E(\hat{p}_{11})$. Un autre type d'erreur d'appariement corrélée se produit lorsque l'appariement est réalisé par des commis faisant preuve d'une tendance variable à commettre des erreurs positives fausses et des erreurs négatives fausses. Un modèle décrivant ce type d'erreurs est présenté à la section suivante.

Il est stipulé dans l'hypothèse 2 que les probabilités d'erreur de classement α_j sont les mêmes pour toute la population de l'EP. Il peut s'agir d'une autre simplification étant donné qu'il est possible de classer certaines personnes, peut-être la majorité, avec un risque d'erreur relativement petit, tandis que l'appariement des autres personnes est plus difficile à réaliser. Fondamentalement, les erreurs d'appariement sont attribuables à l'inexactitude ou à la nature incomplète des informations concernant les caractéristiques de chaque personne dans l'un ou l'autre, ou encore dans les deux systèmes. Par conséquent, si l'échantillon EP peut être poststratifié selon l'exhaustivité des renseignements devant être utilisés aux fins de l'appariement, l'hypothèse peut être plausible (du moins approximativement) pour chaque sous-population. Dans un tel cas, l'erreur d'appariement totale représente une agrégation des taux d'erreur des sous-populations. Ce modèle est étudié dans la dernière sous-section du présent mémoire.

Enfin, l'hypothèse d'un sondage aléatoire simple simplifie de beaucoup l'équation de $\text{Var}(\hat{p}_{11})$. Comme les échantillons EP sont des échantillons complexes, l'hypothèse constitue une simplification qui permet néanmoins d'élaborer des formules utiles pour: a) déterminer quelles sont les composantes de l'erreur d'appariement susceptibles d'avoir l'incidence la plus marquée sur l'EQM totale de $\hat{N}$; b) concevoir des études d'évaluation de l'erreur d'appariement et affecter les ressources nécessaires à la réalisation de ces études. Dans bien des cas, l'ajustement de VE à l'aide d'une constante «d'effet de plan» permettra de compenser pour la majeure partie l'incidence de l'échantillonnage complexe sur $\text{Var}(\hat{p}_{11})$. En outre, pour l'essentiel, les formes d'échantillonnage plus complexes que le sondage aléatoire simple n'ont aucune incidence sur $E(\hat{p}_{11})$, pour autant que l'estimateur $\hat{p}_{11}$ soit pondéré de façon appropriée. En conséquence, la forme de $B(\hat{p}_{11})$ ne dépend pas de cette hypothèse.

2.2 Modélisation des erreurs de commis

Supposons que l'appariement des personnes dénombrées dans l'EP et dans le recensement est effectué par k commis. Soient m_i, le nombre de personnes appartenant à l'EP classées par le commis i, $i=1,\ldots,k$, et l'indice double (i,j), la $j^{ième}$ personne faisant partie de la tâche du $i^{ième}$ commis.

Supposons:

1. L'événement {élément (i,j) est mal classé} et l'événement {élément (i',j') est mal classé} sont indépendants lorsque $i \neq i'$ et conditionnellement indépendants pour tout commis i, $i=i'$; $j \neq j'$; $i=l, \ldots, k$; $j,j' = 1,\ldots,m_i$.
2. $\alpha_{ij} = \theta_i$ si la personne (i,j) a vraiment été recensée, et $= \phi_i$ si la personne (i,j) n'a pas vraiment été recensée.
3. $E(\theta_i) = \theta$; $E(\phi_i) = \phi$; $\text{Var}(\phi_i) = \sigma_\phi^2$; $V(\phi_i) = \sigma_\phi^2$; et $\text{Cov}(\theta_i, \phi_i) = \sigma_{\phi\theta}$.

En ce qui concerne le sous-ensemble de personnes faisant partie de la tâche du $i^{ième}$ commis, les hypothèses 1 et 2 sont analogues aux hypothèses 1 et 2 du modèle exposé à la section précédente. L'hypothèse 3 précise que les probabilités d'erreur d'appariement de commis sont indépendantes et qu'elles constituent des variables aléatoires distribuées de façon identique.

Cette hypothèse est analogue aux hypothèses formulées à l'égard des erreurs d'intervieweur dans le cadre des modèles d'analyse de l'incidence des erreurs d'intervieweur [se reporter, par exemple, à Kish (1962), Hartley et Rao (1978) ainsi que Biemer et Stokes (1985)]. Il convient d'utiliser cette hypothèse si l'objectif visé est d'estimer les paramètres d'un groupe de commis beaucoup plus important dont les k commis de l'EP constituent un échantillon représentatif.

Il est démontré dans l'appendice que $E(\hat{p}_{11})$ est quand même obtenu par l'équation (4) en admettant un sondage aléatoire simple. La formule générale utilisée pour déterminer $\text{Var}(\hat{p}_{11})$ correspond à la formule (A.3) de l'appendice; cependant, une simplification utile est obtenue si on peut supposer que la taille des tâches m_i est approximativement égale à m, la taille moyenne, et que le nombre d'appariements prévus est le même pour chaque tâche de commis (c'est-à-dire que les tâches de commis s'interpénètrent). Alors,

$$\text{Var}(\hat{p}_{11}) = \frac{1}{n}(SV + SMV) + \frac{m-1}{m}\frac{1}{k}CC \tag{11}$$

où CC, représentant la *composante corrélée de la variance d'appariement*, est

$$CC = p_{11}^2\,\sigma_\theta^2 + (1-p_{11})^2\sigma_\phi^2 - 2p_{11}(1-p_{11})\,\sigma_{\phi\theta} \tag{12}$$

et VE, VAS sont obtenues respectivement à l'aide des équations (7) et (8).

Il est à noter que CC résulte de la variabilité des probabilités d'erreur de classement θ_i et ϕ_i. De plus, si on considère que CC constitue la variance de $-p_{11}\,\theta_i + (1-p_{11})\phi_i$ et que les termes de cette équation sont similaires à ceux de l'équation (5), nous constatons que CC représente la variance des biais *nets* parmi les commis. Ce dernier facteur constitue la preuve que la valeur de CC doit être positive. La variance due aux commis a donc pour effet d'accroître la valeur de la variance de $\hat{p}_{11}$.

En se reportant encore aux ouvrages consacrés à la variance de réponse, on peut définir un paramètre ρ_M analogue au coefficnet de corrélation intra-intervieweur, ρ, defini par Kish (1962). Aux fins du présent mémoire, le symbole ρ_M note la corrélation intra-commis puisqu'il représente la corrélation entre les classements d'appariement de toute paire d'éléments compris dans la même tâche de commis. Selon le modèle,

$$\rho_M = \frac{CC}{SV + SMV}$$

est le rapport de la composante corrélée de la variance à la variance totale associée à un seul classement. Ce rapport peut être interprété comme le degré selon lequel les commis «influent» sur les taux d'appariement à l'intérieur de leurs tâches respectives. Ces considérations nous permettent d'établir une autre équation, équivalente à (11), pour obtenir $\text{Var}(\hat{p}_{11})$, soit

$$\text{Var}(\hat{p}_{11}) = \frac{VE + VAS}{n}[1 + (m-1)\rho_M] \tag{13}$$

2.3 Post-stratification

Le modèle d'erreur non corrélée et le modèle d'erreur de commis tiennent tous deux pour acquis (essentiellement) que le degré de difficulté afférent à la détermination de la classe exacte d'appariement est le même pour toutes les personnes dénombrées dans le cadre de l'EP (hypothèse (2) dans le cas des 2 modèles). Dans le cas du modèle d'erreur de commis, par exemple, le vecteur de la probabilité d'erreur de classement (θ_i, ϕ_i) est le même pour toutes les unités de la tâche du $i^{ième}$ commis. Toutefois, dans la pratique, certaines personnes sont beaucoup plus difficiles à classer que d'autres. Ces différences sont attribuables à des facteurs comme

l'exhaustivité des renseignements d'appariement, le fait que la personne ait déménagé ou non, qu'elle habite une maison unifamiliale ou un appartement, etc.

Une façon simple de modéliser cette situation consiste à stratifier l'échantillon de l'EP en fonction d'une variable, disons Z, corrélée avec les probabilités d'erreur de classement α_j. La variable Z peut être un indicateur de l'exhaustivité des renseignements, du type d'élément, etc.

Supposons que L de ces sous-populations sont identifiées pas l'indice h. Soit (i,h,j) le $j^{ième}$ élément de la $h^{ième}$ sous-population dans la tâche du $i^{ième}$ commis où $i=1,\ldots,k$; $h=1,\ldots,L$, $j=0,\ldots,m_{ih}$; et m_{ih} représente le nombre d'éléments de la tâche du $i^{ième}$ commis compris dans la sous-population h. Comme dans le cas du modèle de l'erreur de commis, l'hypothèse (1) peut encore être retenue. Cependant, supposons en plus que:

2. $\alpha_{ihj} = \theta_{ih}$ si la personne (i,h,j) a vraiment été recensée.

   $= \phi_{ih}$ si la personne (i,h,j) n'a pas vraiment été recensée.

3. $E(\theta_{ih}) = \theta_h$; $E(\phi_{ih}) = \phi_h$

   $\text{Var}(\theta_{ih}) = \sigma^2_{\theta h}$; $Var(\phi_{ih}) = \sigma^2_{\phi h}$;

   $Cov(\theta_{ih'}, \theta_{ih}) = \sigma_{\phi\theta h}$ si $h=h'$

   $= 0$ si $h \neq h'$

Selon ces hypothèses, nous obtenons $\text{Biais}(\hat{p}_{11}) = \Sigma\pi_h \text{ Biais } (\hat{p}_{11})$ et $\text{Var } (\hat{p}_{11}) = \Sigma\pi_h^2 \text{ Var } (\hat{p}_{11h}) + \Sigma\pi_h [E(\hat{p}_{11h}) - E(\hat{p}_{11})]^2$ où Biais $(\hat{p}_{11h})$, $E(\hat{p}_{11h})$, et Var $(\hat{p}_{11h})$ sont respectivement obtenus à l'aide des formules (5), (4), et (6), indiquant les paramètres de l'erreur de commis et p_{11} par h et où $\pi_h = E(n_h/n)$, la proportion de la population comprise dans la $h^{ième}$ sous-population.

3. DÉMONSTRATION DE L'INCIDENCE DE L'ERREUR D'APPARIEMENT SUR L'ERREUR TOTALE

Les modèles définis à la section précédente peuvent être utilisés pour démontrer l'incidence de l'erreur d'appariement sur l'erreur quadratique moyenne totale de $\hat{N}$ et $\hat{p}_{11}$. Dans l'exemple qui suit, nous donnerons aux paramètres du modèle des valeurs hypothétiques qui, selon notre expérience, représentent des valeurs types et sont compatibles avec les paramètres de conception de l'EP de 1990.

Dans le cadre de l'EP, des estimations de N seront réalisées pour un certain nombre de sous-populations du recensement. Assumons que le coefficient de variation désiré des estimations est de 1%. Des équipes de commis effectueront l'appariement dans un certain nombre de centres de traitement. (On trouvera des renseignements plus détaillés sur le processus d'appariement dans la section suivante.) Afin d'illustrer l'effet de l'erreur d'appariement sur l'ESD, considérons une sous-population «type» de l'EP. Admettons, pour cette sous-population, que $p_{11} = .85$ et que k, le nombre de commis préposés à l'appariement dans un centre de traitement, est égal à 10. Dans notre analyse, nous avons considéré des valeurs de θ variant entre 0 et .10 ainsi qu'un certain nombre de valeurs types pour le rapport $\gamma = \theta/\phi$, c'est-à-dire, le rapport de la probabilité des erreurs d'appariement négatif faux à la probabilité des erreurs d'appariement positif faux. Étant donné que la mesure de ρ_M pour une erreur d'appariement n'a jamais été effectuée dans le cadre d'aucune étude, il n'existe que peu d'information susceptible d'indiquer l'étendue type de ρ_M. Cependant, si nous assumons que les probabilités d'erreur de commis θ_i et ϕ_i suivent une distribution Bêta unimodale et qu'elles sont corrélées, on peut obtenir une valeur maximale de ρ_M correspondant aux valeurs données des probabilités

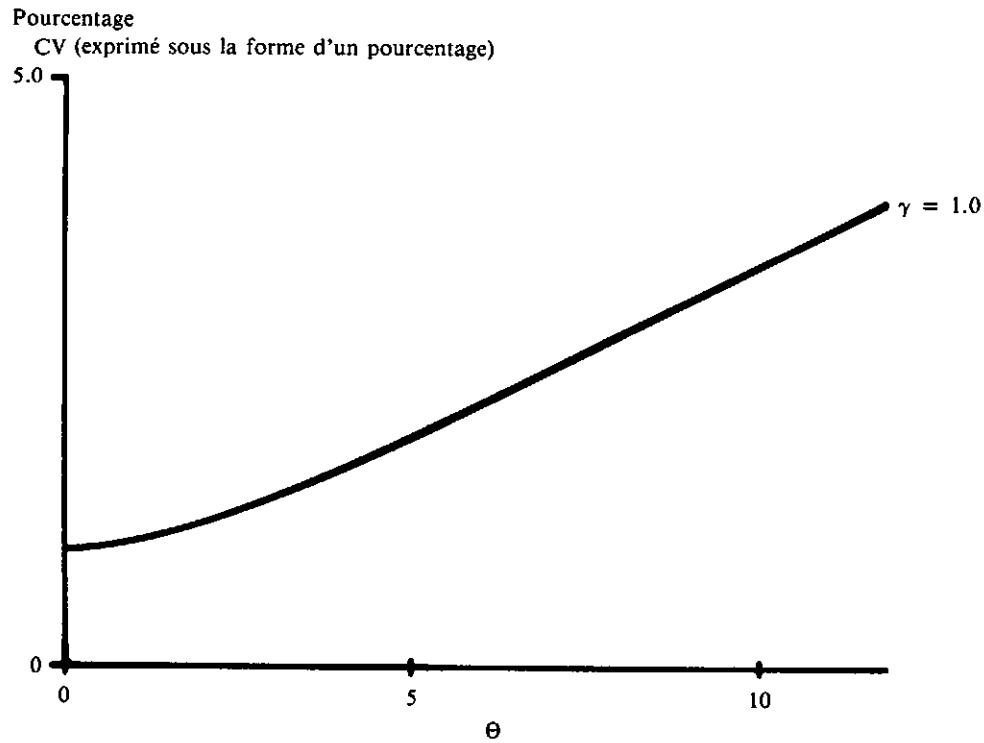

Figure 1. Coefficient de variation de $\hat{p}_{11}$ représenté comme une fonction de Θ pour $\gamma = 1$

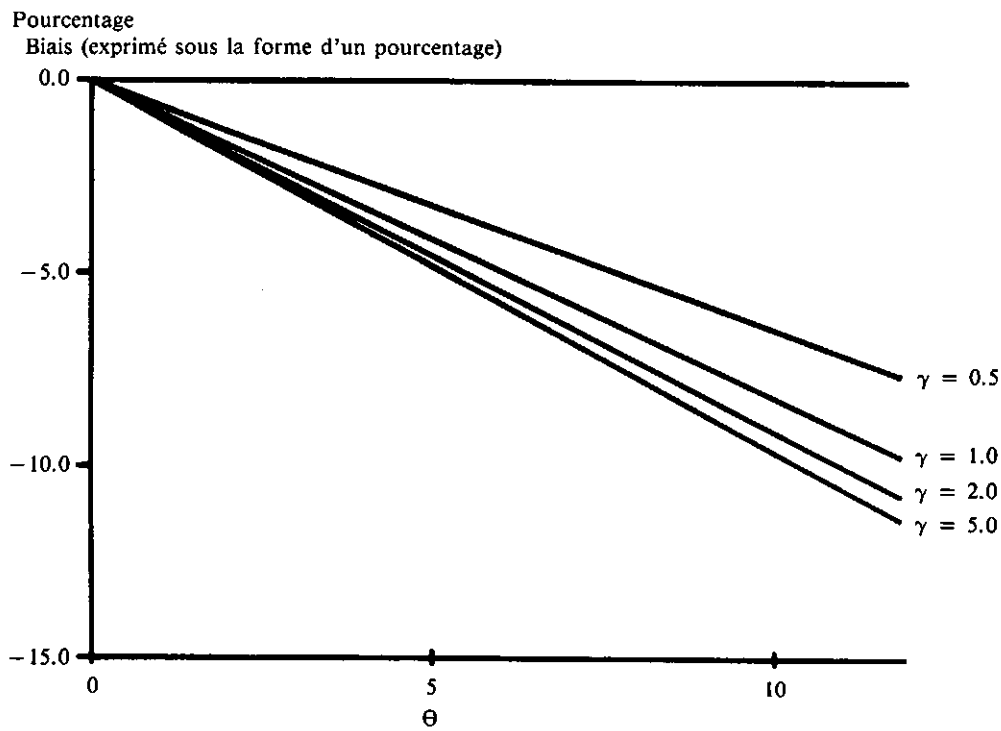

Figure 2. Biais relatif de $\hat{p}_{11}$ représenté comme une fonction de Θ pour des valeurs données de γ

d'erreur prévues θ et ϕ. Algébriquement, la valeur maximale de ρ_M est obtenue à l'aide de la formule

$$\rho_M^* = CC^* / (VAS + VE) \tag{14}$$

où $CC^* = p_{11}^2\,\theta^2\,(1-\theta)/(1+\theta) + (1-p_{11})^2\,\phi^2\,(1-\phi)/(1+\phi)$ (se reporter à Johnson et Kotz (1970) pour connaître la théorie sous-jacente). Si θ_i et ϕ_i sont corrélés de façon positive, l'hypothèse admettant une corrélation égale à zéro ne fait qu'accroître l'incidence de CC. En conséquence, les exemples suivants indiquent l'incidence maximale de la variance d'appariement sur les estimations.

Afin d'illustrer l'incidence maximale d'une variance corrélée sur l'exactitude de $\hat{p}_{11}$, on a tracé une courbe de la variation du coefficient de variation de $\hat{p}_{11}$, soit $CV(\hat{p}_{11})$, en fonction de θ pour différentes valeurs de γ. Aux fins de ces calculs, ρ_M^* a été substitué à ρ_M dans l'équation (13). L'étendue de θ était $0 \leq \theta \leq .10$ et celle deγ était $.5 \leq \gamma \leq 5$; c'est-à-dire $\phi = .2\theta$ à $\phi = 2\theta$. Cette étendue des valeurs de γ semble acceptable puisque, typiquement, ϕ est plus petit que θ. La figure 1 illustre la fonction lorsque $\gamma = 1$. Il n'existe pas de différence visible pour les autres valeurs de γ comprises dans l'étendue qui nous intéresse. Il semble donc que la grandeur de ϕ n'a qu'une influence négligeable sur le $CV(\hat{p}_{11})$. De fait, l'expression de CC^* démontre que, lorsque $p_{11} = .85$, pas plus de 3% de la variance corrélée est attribuable à la variance de ϕ_i, même lorsque la valeur de ϕ est égale à celle de θ. La figure 1 laisse aussi supposer que la valeur du $CV(\hat{p}_{11})$ peut passer à 2%, c'est-à-dire être doublée, pour les valeurs de θ de l'ordre de 5%.

La figure 2 illustre le biais relatif de $\hat{p}_{11}$, représenté par $BR(\hat{p}_{11})$, pour la même étendue de θ $(0 \leq \theta \leq .1)$ et de $\gamma(.5 \leq \gamma \leq 5)$. Le graphique indique clairement que le biais est moins marqué pour les valeurs plus petites de γ. De fait, le biais est égal à zéro lorsque $\gamma = (1 - p_{11})/p_{11}$ ou à .18 en supposant, comme dans le présent exemple, que $p_{11} = .85$. Lorsque θ prend une valeur aussi faible que 5%, le biais relatif se situe entre -2% et -4%, selon la grandeur de γ. Si on compare ce résultat à l'augmentation maximale de $CV(\hat{p}_{11})$ (un point de pourcentage), il appert que le biais peut avoir une incidence beaucoup plus marquée que la variance corrélée.

Afin de déterminer l'incidence possible de l'erreur d'appariement sur $\hat{N}$, on a calculé l'augmentation de l'erreur totale en fonction de θ et pour certaines valeurs de γ. Soit $M(\theta;\gamma)$, $V(\theta;\gamma)$, et $B(\theta;\gamma)$ représentant respectivement l'erreur quadratique moyenne, la variance et le biais de $\hat{N}$ pour des valeurs données de θ et γ. $M(0;\gamma)$ représente l'erreur quadratique moyenne de $\hat{N}$ sans erreur d'appariement (c'est-à-dire $\theta = \phi = 0$) et $M(0;\gamma)^{1/2}$ représente donc de façon approximative l'erreur type de $\hat{N}$. Définissons $RM(\theta;\gamma) = (M(\theta;\gamma)/M(0;\gamma) - 1)^{1/2}$; $RV(\theta;\gamma) = (V(\theta;\gamma)/M(0;\gamma) - 1)^{1/2}$; et $RB(\theta;\gamma) = (B^2(\theta;\gamma)/M(0;\gamma))^{1/2}$.

Ainsi, $RM(\theta;\gamma)$ est égal au rapport de la racine carrée de l'augmentation de l'erreur quadratique moyenne totale de $\hat{N}$ pour un θ et un γ donnés, à la racine carrée de EQM de $\hat{N}$ sans erreur d'appariement. $RV(\theta;\gamma)$ représente la proportion de cet accroissement attribuable à la variance d'appariement, tandis que $RB(\theta;\gamma)$ représente la proportion attribuable au biais d'appariement. Par conséquent, $RM(\theta;\gamma)^2 = RV(\theta;\gamma)^2 + RB(\theta;\gamma)^2$. Les figures 3 et 4 illustrent ces fonctions pour deux valeurs extrêmes de γ, $\gamma = .5$ et 5 et pour $0 \leq \theta \leq .1$. La valeur maximale de la variance corrélée, CC^*, a une fois de plus été utilisée aux fins du calcul de la variance. Aussi, il est probable que la contribution de la variance d'appariement à l'erreur totale soit exagérée de façon considérable.

Ces figures démontrent que, pour ces valeurs de θ et de γ, la majeure partie de l'erreur est attribuable au biais, bien que la contribution de la variance puisse être non négligeable. De plus, comme il a été mentionné pour les figures 1 et 2, le biais d'appariement compte pour la majeure partie de l'erreur d'appariement totale chaque fois que le pourcentage d'erreurs négatives fausses est supérieur à celui des erreurs positives fausses.

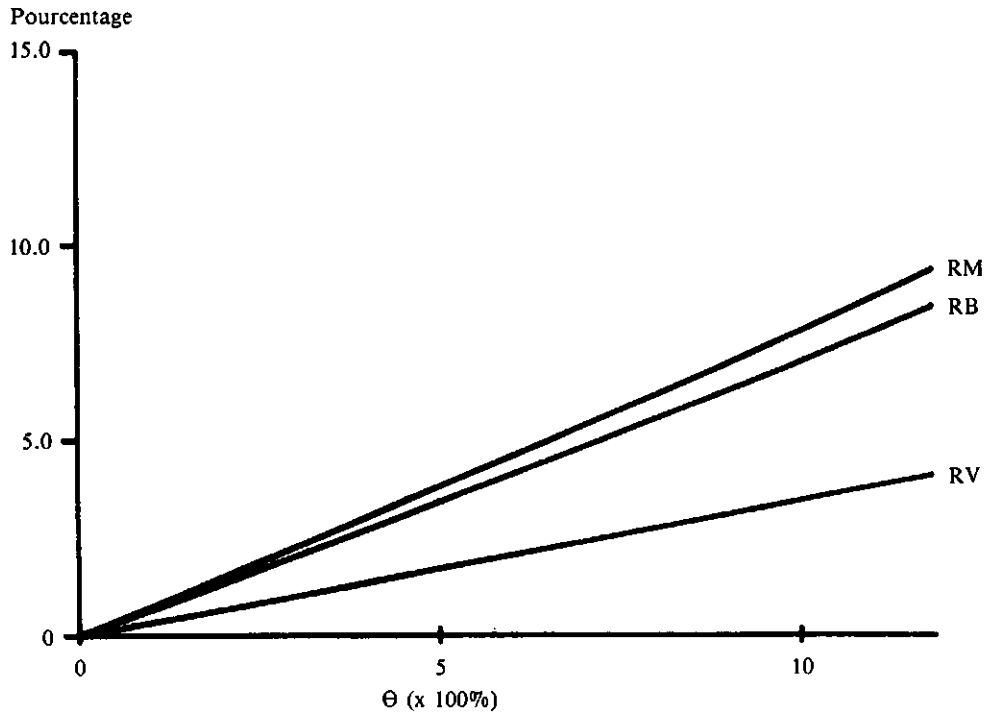

Figure 3. RM(Θ ; γ), RV(Θ ; γ), et RB(Θ ; γ), comme fonctions de Θ pour γ = .5

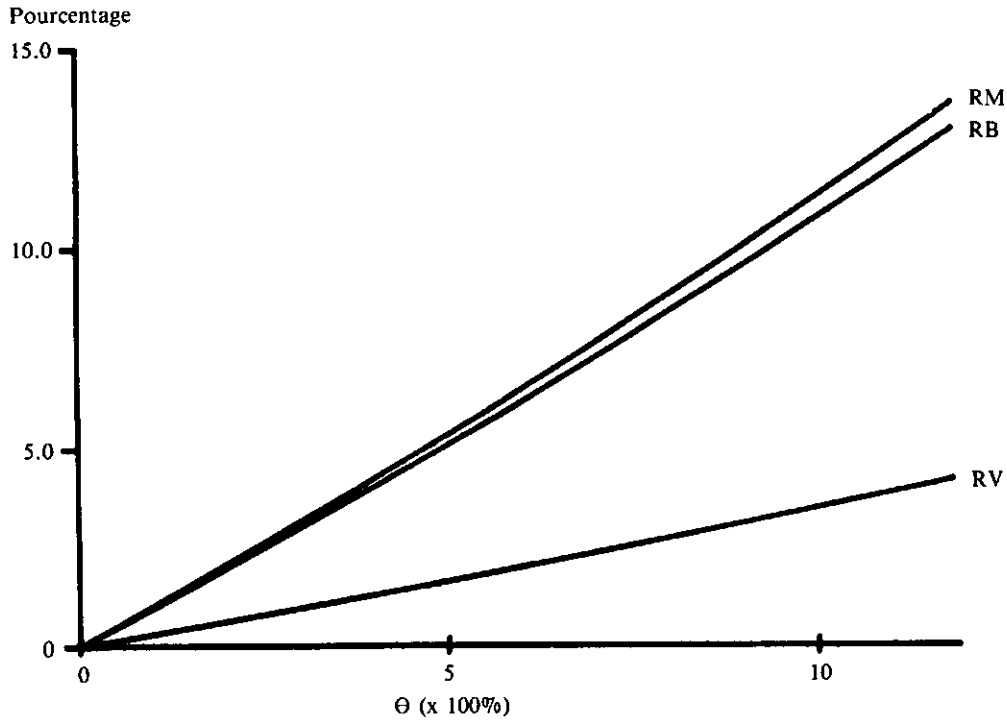

Figure 4. RM(Θ ; γ), RV(Θ ; γ), et RB(Θ ; γ), comme fonctions Θ pour γ = 5

4. ESTIMATION A PARTIR D'ÉTUDES DE RAPPARIEMENT

Plusieurs ouvrages traitent de façon exhaustive des méthodes utilisées pour estimer les composantes de l'erreur de réponse dans des enquêtes-échantillons [se reporter, par exemple, aux ouvrages de Hansen, Hurwitz, Pritzker (1964), Hansen, Hurwitz, Bershad (1961)]. Les méthodes mises en oeuvre pour estimer les composantes de l'erreur d'appariement sont essentiellement les mêmes. Par exemple, pour qu'il soit possible d'estimer la composante corrélée de la variance d'appariement, *CC*, les tâches des commis doivent «s'interpénétrer''. Cette méthode, décrite en détail dans l'ouvrage de Kish (1962), randomise l'attribution des cas d'EP aux commis de sorte que chaque tâche de commis comporte le même nombre prévu de personnes appariées. Ensuite, un estimateur de *CC* est obtenu en faisant la différence entre l'erreur quadratique moyenne des commis entre eux et l'erreur quadratique moyenne intra-commis tirées de l'analyse de la la variance due aux commis. Pour de plus amples renseignements concernant cette méthode, se reporter à l'ouvrage intitulé Bureau of the Census (1985).

La présente section porte surtout sur l'analyse des données relatives aux études de rappariement qui constitue la méthode d'évaluation de l'erreur d'appariement la plus utilisée. Il existe deux types d'étude de rappariement. La première tente de reproduire l'appariement initial pour un échantillon de cas en utilisant les mêmes méthodes et règles d'appariement, en donnant la même formation, etc. Ce type de rappariement a pour objet d'estimer VAS, la variance d'appariement simple ou, de manière équivalente, I_M, l'indice d'incohérence de l'appariement. Le second type de rappariement est utilisé en vue d'obtenir l'appariement le plus exact possible. En conséquence, sa réalisation nécessite la mise en oeuvre de méthodes plus détaillées de même que le recours à des commis experts hautement qualifiés et à l'arbitrage, c'est-à-dire à la résolution des problèmes de non-correspondance entre les classements originaux et ceux résultant du rappariement par un tiers, expert en matière d'appariement. Ce type de rappariement a pour objet d'estimer la valeur du biais d'appariement. De plus, comme il sera démontré plus loin, il est aussi possible d'obtenir une estimation de VAS à partir de ces données.

Les données (non pondérées) recueillies dans le cadre d'une étude de rappariement peuvent être présentées sous la forme illustrée au tableau 1.

Supposons que l'échantillon utilisé pour le rappariement constitue un échantillon aléatoire simple de r personnes de l'EP. De plus, il est possible de supposer que le modèle d'erreur non corrélée ou le modèle d'erreur de commis de la section précédente s'applique aussi bien à l'appariement qu'au rappariement. Soit μ_t ($t = a, b, c, d$) la proportion moyenne de la case correspondant à t dans le tableau 1.

Tableau 1

Données d'une étude de rappariement

Classement original	Classement de rappariement	
	Éléments appariés	Éléments non appariés
Éléments Appariés	a	b
Éléments non appariés	c	d

Alors,

$$\mu_a = p_{11}\,(1-\theta_A)\,(1-\theta_B) + (1-p_{11})\phi_A\phi_B \tag{15}$$

$$\mu_b = p_{11}\,(1-\theta_A)\,\theta_B + (1-p_{11})\phi_A\,(1-\phi_B) \tag{16}$$

$$\mu_c = p_{11}\,\theta_A\,(1-\theta_B) + (1-p_{11})\,(1-\phi_A)\phi_B \tag{17}$$

$$\mu_d = p_{11}\,\theta_A\,\theta_B + (1-p_{11})\,(1-\phi_A)(1-\phi_B) \tag{18}$$

où l'indice A représente l'appariement initial et l'indice B le rappariement.

Définissons,

$$\mu_A = E\left(\frac{a+b}{r}\right) = p_{11}(1-\theta_A) + (1-p_{11})\,\phi_A \tag{19}$$

et

$$\mu_B = E\left(\frac{a+c}{r}\right) = p_{11}(1-\theta_B) + (1-p_{11})\,\phi_B. \tag{20}$$

On notera que μ_A et μ_B sont les valeurs prévues des estimations de p_{11} réalisées respectivement à partir des classements originaux et des classements du rappariement. La différence entre ces deux estimations de p_{11}, c'est-à-dire $(b-c)/r$ est appelée *taux de différence net* (TDN). Sa valeur prévue est

$$E(NDR) = \mu_A - \mu_B = -p_{11}(\theta_A - \theta_B) + (1-p_{11})\,(\phi_A - \phi_B). \tag{21}$$

Enfin, la proportion des personnes de l'échantillon r dont le classement du rappariement ne correspond pas au classement du premier appariement est représentée par $(b+c)/r$, quotient désigné sous le nom de *taux de différence brut* (TDB). Sa valeur prévue est

$$E(TDB) = \mu_b + \mu_c$$

$$= p_{11}\,[\theta_A\,(1-\theta_B) + (1-\theta_A)\theta_B] + (1-p_{11})[(1-\phi_A)\phi_B + \phi_A\,(1-\phi_B)]. \tag{22}$$

Trois jeux d'hypothèses devront être pris en considération pour l'estimation des composantes de Var$(\hat{p}_{11})$ et du Biais$(\hat{p}_{11})$ dans l'étude du rappariement. Dans le premier cas, on suppose que l'étude de rappariement est réalisée dans les mêmes conditions générales que l'étude d'appariement initiale, de sorte que les paramètres d'erreur associés aux deux classements sont sensiblement les mêmes. Par exemple, les commis ont reçu la même formation, possèdent les mêmes aptitudes et utilisent les mêmes méthodes. Selon la deuxième hypothèse, le rappariement est parfait, c'est-àdire que le classement du rappariement peut être considéré comme le vrai classement. La troisième hypothèse se situe quelque part entre la première et la deuxième, elle suppose que des méthodes d'appariement améliorées et plus détaillées sont utilisées pour le rappariement; cependant, nous ne sommes pas prêts à admettre que les classements du rappariement ne comportent pas d'erreur. Nous supposerons plutôt qu'il se produit moins d'erreur lors du rappariement que lors de l'appariement initial.

Hypothèse 1. L'appariement et le rappariement ont été réalisés dans les mêmes conditions générales.

Supposons que $\theta_A = \theta_B = \theta$ et $\phi_A = \phi_B = \phi$, c'est-à-dire que le taux d'erreur de classement prévu est le même pour les deux appariements. Alors, selon la formule (21), $E(NDR) = 0$ et aucune estimation du Biais$(\hat{p}_{11})$ ne peut être calculée à l'aide des données. Cependant, selon les formules (22) et (8)

$$\tfrac{1}{2} E(TDB) = VAS \tag{23}$$

De plus, un estimateur de I_M dans la formule (9) est obtenu par

$$\hat{I}_M = TDB / [2\hat{p}_{11}(1-\hat{p}_{11})] \tag{24}$$

où $\hat{p}_{11}$ est l'estimateur EP de p_{11} comme il a été défini pour (2). Comme alternative, un estimateur de $E(\hat{p}_{11})$ peut être obtenu à partir du tableau 1; par exemple, se reporter aux estimateurs des formules (19) et (20).

Hypothèse 2. Rappariement parfait.

Supposons que $\theta_B = \phi_B = 0$; c'est-à-dire que le rappariement est réalisé sans erreur de classement. Alors, selon la formule (21),

$$\begin{aligned} E(TDN) &= -p_{11}\theta_A + (1-p_{11})\phi_A \\ &= \text{Biais } (\hat{p}_{11}). \end{aligned} \tag{25}$$

De plus, la probabilité d'erreur négative fausse, θ_A, est estimée par

$$\hat{\theta} = c/(a+c). \tag{26}$$

et la probabilité d'erreur positive fausse, ϕ_A, est estimée par

$$\hat{\phi} = b/(b+d). \tag{27}$$

Un estimateur de VAS est obtenu à l'aide de la formule

$$\widehat{VAS} = \frac{1}{r}\left(\frac{ac}{a+c} + \frac{bd}{b+d}\right) \tag{28}$$

et on obtient un estimateur de I_M grâce à la formule

$$\hat{I}_M = \widehat{VAS}/\hat{p}_{11}(1-\hat{p}_{11}) \tag{29}$$

où $\hat{p}_{11}$ représente un estimateur de $E(\hat{p}_{11})$ obtenue à partir de l'EP ou du tableau 1.

Hypothèse 3. Le rappariement comprend moins d'erreurs mais n'est pas parfait.

Supposons que $0 < \theta_B < \theta_A$ et $0 < \phi_B < \phi_A$; c'est-à-dire que les probabilités d'erreur de classement sont plus faibles dans le cas du rappariement que dans celui de l'appariement initial, sans être nulles. Il n'existe alors pas d'estimateur non biaisé du Biais$(\hat{p}_{11})$. Cependant, $|E(TDN)|$ sera plus petite que |Biais $(\hat{p}_{11})$| si les différences $\mu_A - p_{11}$ et $\mu_B - p_{11}$ ont toutes deux le même signe; c'est-à-dire que les estimateurs définis à partir de l'appariement initial et du rappariement sont biaisés dans la même direction. Ainsi, dans ces conditions, $|TDN|$ constitue l'estimateur minorant de |Bias $(\hat{p}_{11})$|.

De plus, il n'existe pas d'estimateur non biaisé de VAS. Il est toutefois possible de démontrer à partir de la formule (22) que

$$E(TDB) - 2VAS = p_{11}(\theta_B - \theta_A)(1 - 2\theta_A) + (1 - p_{11})(\phi_B - \phi_A)(1 - 2\phi_A).$$

Ainsi, chaque fois que θ_A et ϕ_A ont tous deux une valeur inférieure à .5, ce qui est vrai pour la plupart des applications pratiques, il s'ensuit que

$$E(TDB) < 2VAS$$

et que $\hat{I}_M$ défini par la formule (24) constitue une sous-estimation de I_M.

5. APPLICATION AU RECENSEMENT DE 1990

Dans le cadre du recensement de 1990, l'échantillon de l'EP sera constitué d'environ 5 000 «îlots» ou groupes d'environ 30 logements contigus et on tentera de procéder à l'appariement de chaque personne comprise dans les îlots avec les personnes recensées. Les variables utilisées aux fins de l'appariement seront le nom, l'adresse, le lien avec le chef de ménage, le sexe, la date de naissance, l'état matrimonial, la race et l'origine hispanique. Le processus d'appariement comportera les quatre étapes suivantes.

Étape 1. *Un appariement sera réalisé par ordinateur selon la technique de Fellegi et Sunter (1969).* Chaque personne faisant partie de l'EP sera classée par ordinateur soit dans la catégorie des personnes appariées à une personne recensée, soit dans celle des personnes non appariées ou encore dans la catégorie des personnes qu'il serait peut être possible d'apparier (c'est-à-dire dont le classement doit être révisé par un commis).

Étape 2. *Des commis effectueront une première révision afin de corriger toutes les erreurs d'appariement ou de non-appariement erroné introduites par l'ordinateur.* De plus, un ensemble normalisé de règles d'appariement sera appliqué à chaque appariement possible. Ainsi, chaque personne dénombrée dans le cadre de l'EP sera classée sous une des rubriques appariement, non-appariement, appariement possible ou cas non résolu.

Étape 3. *Des commis effectueront avec discernement une deuxième révision afin de vérifier de nouveau les classements réalisés aux deux étapes précédentes.* A cette étape-ci, les commis, désignés sous le nom d'équipe d'appariement spéciale (EAS), peuvent aussi décider du bien fondé d'un suivi sur place auprès de certains ménages.

Étape 4. *Une révision «subséquente au suivi sur place».* Les cas sont examinés de nouveau à la lumière de toute information supplémentaire obtenue dans le cadre du suivi sur place. Les codes de classement définitifs sont: apparié (dénombré), non apparié (non dénombré) ou non résolu (statut d'appariement devant être déterminé à l'étape de traitement final).

Les méthodes utilisées pour attribuer les codes «apparié» ou «non apparié» aux cas non résolus sont décrites dans le rapport de Schenker (1987). Ces cas, qui représentent environ 1% de l'échantillon de l'EP, ne sont pas pris en compte dans les tableaux qui suivent puisque les statuts d'appariement n'avaient pas été attribués aux cas non résolus, au moment de ce test. Néanmoins, l'erreur d'imputation peut constituer une source importante d'erreur d'appariement et poser des problèmes d'évaluation particuliers. Il est probable, par exemple, que certains cas d'EP non résolus resteront non résolus lors du rappariement et aucune estimation directe de l'erreur de classement ne peut être calculée pour ces cas. Dans le test décrit plus loin, 83% des cas non résolus dans le cadre de l'EP sont demeurés non résolus au rappariement. Inversement, 41% des cas qui n'ont pas été résolus au rappariement, ont été résolus à l'appariement de l'EP. Si on suppose que les imputations pour les cas qui n'ont pas été résolues lors du rappariement sont erronées, on peut obtenir un majorant de l'erreur d'imputation. De même, il est possible d'obtenir un minorant en supposant que toutes ces imputations sont justes. Cependant, à moins que la proportion représentée par les imputations soit très faible, cette analyse des «meilleurs cas, pires cas» peut produire des bornes trop élevées pour être utiles.

En 1986, un test d'essai de ces méthodes d'appariement d'EP a été effectué à Los Angeles. Un échantillon d'environ 4 000 personnes a été apparié avec le recensement d'essai de Los Angeles, puis apparié de nouveau par des spécialistes du recensement en vue d'évaluer le biais d'appariement. Des méthodes spéciales on été utilisées lors du rappariement afin d'assurer un classement d'appariement très exact. Le tableau 2 illustre les taux de désaccord entre les quatre étapes d'appariement et de rappariement. Ce tableau met en lumière l'amélioration des classements à chaque étape supérieure, amélioration indiquée par un taux de désaccord décroissant dans la colonne réservée au rappariement. Les données indiquent également que peu de classements sont modifiés lors de l'étape subséquente au suivi sur place (.68% en désaccord avec l'étape 3). De plus, le TDB de l'étape finale (par rapport au rappariement) est très bas, soit de moins de 1%.

En supposant que le processus de rappariement conduit à l'attribution de vrais statuts d'appariement, le tableau 3 indique les estimations de θ, la probabilité d'erreur d'appariement négatif faux, et de ϕ, la probabilité d'erreur d'appariement positif faux, pour chaque étape d'appariement. Il ressort de ce tableau que le taux de non-appariement faux est plus élevé dans le cas de l'appariement par ordinateur et dans le cas du premier appariement effectué par des commis. Cependant, le contraire est vrai pour les deux dernières étapes d'appariement.

Tableau 2

Comparaison des taux de désaccord entre les étapes d'appariement (%)

	Étape 2	Étape 3	Étape 4	Rappariement
Étape 1	2.9	4.4	4.7	5.5
Étape 2	0	3.3	4.0	4.8
Étape 3	3.3	0	.68	1.6
Étape 4	4.0	.68	0	.87

Tableau 3

Estimations de θ et ϕ pour les étapes d'appariement

Étape d'appariement	Estimation de θ (x100%) (taux de non-appariement faux)	Estimation de ϕ (x100%) (taux d'appariement faux)
1	6.2	2.3
2	5.1	3.3
3	1.5	2.1
4	.1	.3

Tableau 4

Resultats de l'étude de rappariement (pondérés).

Classement de l'appariement initial	Classement du rappariement	
	Appariés	Non appariés
Appariés	16 690	9
Non appariés	85	2 178

Tableau 5a

Resultats du rappariement pour les cas qui ont conservé leur statut jusqu'à la fin du procédé de rappariement.

Classement de l'appariement initial	Classement du rappariement	
	Appariés	Non appariés
Appariés	14 458	0
Non appariés	640	1 775

Tableau 5b

Resultats du rappariement pour les cas dont le classement a été modifié au cours d'au moins une des quatre étapes.

Classement de l'appariement initial	Classement du rappariement	
	Appariés	Non appariés
Appariés	2 223	9
Non appariés	21	403

En utilisant la méthodologie de la section précédente, il est possible d'estimer Biaisrel$(\hat{p}_{11})$, Biaisrel$(\hat{N})$ et I_M, l'indice d'incohérence de l'appariement. Le tableau 4 révèle les résultats de l'étude du rappariement, pondérés en fonction des probabilités de sélection de l'échantillon du rappariement. Pour ce tableau, le Biaisrel$(\hat{p}_{11})$ est estimé à $-.4$ %, et Biaisrel$(\hat{N})$ est donc estimé à .4 %. Ces calculs sont effectués à l'aide de la formule (2) en supposant que le coefficient de variation de $\hat{p}_{11}$ est de 1 % et en remplaçant le Biaisrel$(\hat{p}_{11})$ par son estimation. L'indice d'incohérence de l'appariement est estimé à .49 %, ce qui représente un très faible pourcentage. Le taux d'appariement positif faux, $\hat{\phi}$, est égal à .004 et le taux d'appariement négatif faux, $\hat{\theta}$, est égal à .005.

Comme il a été mentionné à la deuxième section, la probabilité d'erreur d'appariement peut, entre autres choses, être fonction de l'exhaustivité des données de l'EP ou du recensement. L'échantillon du rappariement à été divisé en deux sous-échantillons afin d'illustrer dans quelle mesure les taux d'erreur d'appariement peuvent varier. Le premier sous-échantillon était composé des cas qui avaient été classés comme «appariés» ou «non appariés» à toutes les étapes d'appariement sans exception. Le second sous-échantillon était constitué du reste de l'échantillon, c'est-à-dire des cas dont le classement a été modifié au cours d'au moins une des quatre étapes. Cette division de l'échantillon correspond à peu près à une division reposant sur l'exhaustivité des données d'appariement étant donné que la plupart des cas classés dans la même

catégorie au cours des quatre étapes sont ceux pour lesquels les données recueillies sont les plus complètes. Les tableaux 5a (cas pour lesquels les données sont complètes) et 5b (cas pour lesquels les données sont incomplètes) illustrent les résultats pondérés.

Dans le premier cas, le taux d'appariement négatif faux est de .44% tandis que le taux d'appariement positif faux est nul. Donc, aucun de ces cas n'a fait l'objet d'un appariement erroné, bien qu'un certain nombre d'entre eux aient été faussement désignés comme nonappariements. Ces données peuvent attester de la plus grande habileté du personnel responsable du rappariement à apparier les cas de l'EP. I_M est estimé à .39%, ce qui représente un indice très bas. En ce qui concerne les cas pour lesquels les données sont incomplètes, le taux d'appariement négatif faux est de .93% tandis que le taux d'appariement positif faux est de 2.18%. I_M est estimé à 1.1%, ce qui représente encore un indice assez bas. Cependant, ces données indiquent que le risque de produire de faux appariements est plus grand en ce qui concerne les cas pour lesquels l'information est incomplète.

Les données recueillies dans le cadre de cette étude indiquent que l'erreur d'appariement induit un faible biais négatif ($-.4\%$) dans l'estimateur $\hat{N}$, qui correspond à une sous-estimation de la valeur de N d'environ un million de personnes (en supposant que $N = 250$ millions de personnes). Le biais est seulement de $-.7\%$ même pour les cas les plus difficiles. Il serait intéressant d'analyser certains sous-groupes démographiques de la population (personnes ayant déménagé, enquêtés-substituts et personnes habitant un appartement) afin de déterminer l'importance de l'erreur d'appariement pour ces sous-groupes.

6. RÉSUMÉ

Les modèles et les formules de l'EQM qui ont été développés dans le présent mémoire peuvent être utilisés pour évaluer l'incidence de l'erreur d'appariement sur des estimations de l'erreur d'observation du recensement. Dans le cadre du recensement des États-Unis de 1990, le biais de l'erreur d'appariement semble être l'élément le plus important de l'EQM($\hat{N}$). Des études préliminaires ont démontré que le biais de l'erreur d'appariement pour le recensement de 1990 est de peu d'importance, sa valeur étant inférieure à .5%. Cette estimation ne tient pas compte de l'erreur d'imputation qui s'applique à environ 1% des cas de l'EP. De plus, l'exactitude des estimations de biais repose dans une large mesure sur la véracité de l'hypothèse selon laquelle le rappariement permet d'obtenir le vrai classement d'appariement. D'autres études seront nécessaires pour déterminer la validité de cette hypothèse.

Lors de l'élaboration des formules de l'erreur quadratique moyenne totale de $\hat{N}$, il a été supposé qu'il était peu probable que N_c comporte une erreur. Cependant, dans la pratique, il convient de soustraire de N_c une estimation du nombre d'événements faux du recensement (ou des dénombrements erronés), représentée par EE. Comme cet estimateur est obtenu à partir d'un appariement des éléments d'un échantillon de la population du recensement aux éléments de la population de l'EP, EE est aussi sujet à une erreur d'échantillonnage et à une erreur d'appariement. Par exemple, une personne peut être classée comme dénombrée de façon erronée alors qu'elle l'avait été correctement (erreur de dénombrement positif faux), ou peut être classée comme dénombrée correctement alors qu'elle l'avait été de façon erronée (erreur de dénombrement négatif faux). La méthodologie et le modèle présentés pour évaluer l'incidence des erreurs d'appariement positif faux et d'appariement négatif faux pour x_{11} peuvent facilement s'appliquer à l'estimateur des dénombrements erronés. On notera que les termes des formules d'approximation de Taylor, (2) et (3), utilisées pour calculer le biais et la variance de $\hat{N}$ seront remplacés par les termes correspondant au biais et à la variance de EE.

En ce qui concerne les recherches à venir, des études portant sur la variance corrélée de l'erreur d'appariement seront nécessaires pour déterminer dans quelle mesure la variance due aux commis contribue à l'erreur totale de $\hat{N}$. Il est possible que $CC*$, l'incidence maximale de l'erreur corrélée, se traduise par une surestimation considérable de l'incidence de l'erreur de commis.

REMERCIEMENTS

La rédaction du présent mémoire a été rendue possible grâce à un Joint Statistical Agreement avec le Bureau of the Census des États-Unis. L'auteur souhaite remercier Aref Dejani du Bureau of the Census pour avoir mis à sa disposition le soutien informatique nécessaire. Il remercie également Bernice Garrett qui a dactylographié le mémoire et s'est chargée de sa relecture.

APPENDICE

Dérivation des formules de l'EQM

Soit U la population de taille N à dénombrer, U_c le sous-ensemble tiré de la population U S l'échantillon tiré de l'EP, et S_c l'ensemble des personnes de l'EP faisant partie du recensement, représentant $S \cap U_c$. Enfin, soit les n éléments de l'échantillon S $u_1, \ldots, u_n$. Définissons les variables

$$\eta_i = 1 \text{ si } u_i \in S_c$$

$$= 0 \text{ si } u_i \notin S_c$$

et

$$y_i = 1 \text{ si } u_i \text{ a été classé (par appariement) dans } S_c.$$

$$= 0 \text{ si } u_i \text{ n'a pas été classé dans } S_c.$$

Modèle d'évaluation de l'erreur de commis

Supposons que: 1) y_i est une variable aléatoire, que les probabilités que $P(y_i = 1 \mid \eta_i = 0) = \phi$ et $P(y_i = 0 \mid \eta_i = 1) = \theta$, et 2) que y_i et y_j sont des variables indépendantes étant donné η_i et η_j pour $i \neq j$. Soit $E(\cdot \mid S)$ l'espérance conditionnelle et $V(\cdot \mid S)$, la variance conditionnelle étant donné S. Alors, $\hat{p}_{11} = \Sigma y_i / n$ et $E(\hat{p}_{11} \mid S) = (1-\theta)\tilde{p}_{11} + \phi(1-\tilde{p}_{11})$ où $\tilde{p}_{11} = \Sigma \eta_i / n$. Le résultat de l'expression (5) est obtenu en calculant l'espérance par rapport S.

De plus, $V(y_i \mid \eta_i = 0) = \phi(1-\phi)$ et $V(y_i \mid \eta_i = 1) = \theta(1-\theta)$. Par conséquent, $V(\hat{p}_{11} \mid S) = \phi(1-\phi)(1-\tilde{p}_{11})/n + \theta(1-\theta)\tilde{p}_{11}/n$.

VAS (8) est obtenu en calculant l'espérance par rapport à S.

Enfin, le résultat de l'expression (6) est obtenu en combinant $VE(\hat{p}_{11} \mid S)$ et $EV(\hat{p}_{11} \mid S)$.

Modèle d'évaluation de l'erreur de commis

Soit (i,j) la $j^{ième}$ personne de la tâche du $i^{ième}$ commis et soit y_{ij} *et* η_{ij} définis en fonction de y_i et η_i. Supposons (1) — (3) pour le modèle d'erreur de commis. E_2, V_2, et C_2 respectivement, l'espérance, la variance et la covariance conditionnelles par rapport aux distributions de l'erreur de commis maintenant l'échantillon de commis fixé. Les valeurs correspondantes E_1 l'espérance, V_1 la variance, C_1 la covariance par rapport à la sélection aléatoire des vecteurs des paramètres des commis k, comme pour l'hypothèse (3), maintenant l'échantillon S fixé. Donc,

$$E_1 E_2 (\hat{p}_{11}) = E_1 \left\{ \sum_i \left[(1-\theta_i) \frac{n_{1i}}{n} + \phi_i \frac{n_{0i}}{n} \right] \right\}$$

$$= (1-\theta)\tilde{p}_{11} + \phi(1-\tilde{p}_{11})$$

où $n_{1i} = \sum_j \eta_{ij}$ et $n_{0i} = \sum_j (1 - \eta_{ij})$. Donc, (4) est obtenu en calculant l'espérence de (A.1) par rapport à S.

Considérons la variance de $\hat{p}_{11}$. On a $\text{Var}(\hat{p}_{11}) = \text{VE}(\hat{p}_{11}|\, S) + \text{EV}(\hat{p}_{11}|\, S)$ où $\text{E}(\hat{p}_{11}|\, S)$ provient de (A.1). De plus $n^2 V(\hat{p}_{11}|\, S) = \sum_i \sum_j V(y_{ij}|\, S) + \sum_i \sum_{j \neq j'} \text{Cov}(y_{ij}, y_{ij'}|\, S)$ où $V(y_{ij}|\, S) = V_2(y_{ij}) + V_1E_2(y_{ij})$ et $\text{Cov}(y_{ij}, y_{ij'}|\, S) = C_1\,[E_2(y_{ij}), E_2(y_{ij'})], E_1C_2(y_{ij}, y_{ij'})$ étant égal à zéro. Étant donné que $E_2(y_{ij}) = \phi_i$, pour $\eta_{ij} = 0$, et que $E(y_{ij}) = 1 - \theta_i$ pour $\eta_{ij} = 1$, $V_1E_2(y_{ij}) = \sigma_\phi^2$, si $\eta_{ij} = 0$, et $V_1E_2(y_{ij}) = \sigma_\theta^2$ si $\eta_{ij} = 1$. De plus, $V_2(y_{ij}) = \phi_i(1-\phi_i)$ pour $\eta_{ij} = 0$ et $V_2(y_{ij}) = \theta_i(1-\theta_i)$ pour $\eta_{ij} = 1$. Ainsi

$$E_1V_2(y_{ij}) = \phi(1-\phi) - \sigma_\phi^2 \text{ si } \eta_{ij} = 0$$

$$= \theta(1-\theta) - \sigma_\theta^2 \text{ si } \eta_{ij} = 1$$

De la même façon, on peut démontrer que, pour $j \neq j'$,

$$C_1\,[E_2(y_{ij}), E_2(y_{ij'})] = \sigma_\theta^2 \text{ si } (\eta_{ij}, \eta_{ij'}) = (1,1)$$

$$= -\sigma_{\phi\theta} \text{ if } (\eta_{ij}, \eta_{ij'}) = (1,0)$$

$$= \sigma_\phi^2 \text{ if } (\eta_{ij}, \eta_{ij'}) = (0,0).$$

Par conséquent,

$$V(\hat{p}_{11}|\, S) = (\Sigma m_i^2 - n)/n^2\, CC + VAS/n. \tag{A.2}$$

Finalement, en combinant (A.1) et (A.2) dans l'expression

$$V(\hat{p}_{11}) = VE(\hat{p}_{11}|\, S) + EV(\hat{p}_{11}|\, S), \text{ nous obtenons}$$

$$V(\hat{p}_{11}) = 1/n(VE + VAS) + (\Sigma m_i^2 - n)/n^2\, CC. \tag{A.3}$$

Si de plus, nous supposons que $m_i = m$ pour tout i, nous obtenons la forme de l'expression (11).

BIBLIOGRAPHIE

BIEMER, P. P., et STOKES, S. L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of American Statistical Association* 80, 158-166.

BUREAU OF THE CENSUS (1985). *Evaluating censuses of population et housing*, Statistical Training Document. United States Bureau of the Census, Washington, D.C.

BURNHAM, K. P., ANDERSON, D. R., WHITE, G. C., BROWNIE, C., et POLLOCK, K. H. (1987). *Design et Analysis Methods for Fish Survival Experiments Based on Release — Recapture.* American Fisheries Society Monograph 5.

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

HANSEN, M. H., HURWITZ, W. N., et BERSHAD, M. A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.

HANSEN, M. H., HURWITZ, W. N., et PRITZKER, L. (1964). The estimation and interpretation of cross differences and the simple response variance. Dans *Contributions to Statistics* (éd. C. R. Rao), Oxford: Pergamon Press, 111-136.

HARTLEY, H. O., et RAO, J. N. K. Estimation of nonsampling variance components in sample surveys. Dans *Survey Sampling and Measurement*, (éd. N.K. Namboodiri), New York: Academic Press.

JOHNSON, N. L., et KOTZ S. (1969). *Continuous Univariate Distributions II*. Boston: Houghton Mifflen.

KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

MARKS, E. S., SELTZER, W., et KROTKI, K. J. (1974). *Population Growth Estimation*. New York: Population Council.

RICKER, W. E. (1958). *Handbook of Computations for Biological Statistics of Fish Populations*. Fisheries Research Board of Canada. Ottawa: Queen's Printer and Controller of Stationery.

SEBER, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*, (3^{e} éd.). New York: MacMillian.

SEKAR, C. C., et DEMING, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

SELTZER, W., et ADLAKTA A. (1974). On the effect of errors in the application of the Chandrasekaran — Deming techniques, Reprint 14. Laboratory for Population Statistics, University of North Carolina.

SCHENKER, N. (1987). Report on missing data in the 1986 test of adjustment related operations. Survey Research Division Report Series, Bureau of the Census, RR-87/09.

SCHEUREN, F., et OH, H. L. (1985). Fiddling around with nonmatches and mismatches. *Proceedings of the Workshop on Exact Matching Methodologies*. Arlington, Virginia.

WOLTER, K. M. (1983). Affidavit, Mario Cuomo et coll. *vs.* Malcolm Baldrige et coll. U. S. District Court, Southern District of New York, 80 Civ. 4550.

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.