Novembre 1992

SYMPOSIUM 92

Conception et analyse des enquêtes longitudinales

RECUEIL

Statistique Canada
Statistics Canada

Canada

SYMPOSIUM 92

Conception et analyse des enquêtes longitudinales

2 au 4 novembre 1992

Ottawa (Ontario) Canada

RECUEIL

Août 1993

Comité organisateur du Symposium 92

John Armstrong Nancy Darcovich Pierre Lavallée

Publication autorisée par le ministre responsable de Statistique Canada

AVANT-PROPOS

Le Symposium 92 a été le neuvième d'une série de conférences annuelles parrainées par Statistique Canada, portant sur des questions méthodologiques. Chaque symposium a porté sur un sujet particulier; celui de 1992 avait pour thème la conception et l'analyse des enquêtes longitudinales.

Le Symposium 92 a réuni plus de 400 personnes venant de neuf pays. Rassemblées pendant trois jours au Centre de conférences Simon Goldberg à Ottawa, ces personnes ont écouté le point de vue d'universitaires et de spécialistes d'organismes gouvernementaux et du secteur privé. En tout, 31 communications ont été présentées. Il s'agissait uniquement de conférenciers invités. Nous nous sommes chargés de la traduction et de la mise en page des communications reproduites dans ce recueil.

Les organisateurs du Symposium 92 tiennent à exprimer leurs remerciements aux nombreuses personnes qui ont contribué à la réalisation de cet ouvrage.

Il convient, naturellement, de remercier les conférenciers pour avoir pris le temps de mettre leurs idées par écrit. La publication de ce recueil a aussi nécessité le concours de nombreuses autres personnes. Le traitement des manuscrits a été exécuté habilement par Christine Larabie et Carmen Lacroix, aidées de Myra Kent. La traduction des communications a été confiée à Michel Charuest, Pierre Desautels, Maryse Montpetit et Josée René de Cotret. La correction des épreuves a été effectuée par de nombreux méthodologistes, dont Yanick Beaucage, Jean-René Boudreau, René Boyer, Marcel Bureau, James Croal, André Cyr, Sylvie DeBlois, Johanne Denis, Johane Dufour, Gerrit Faber, Lyne Guertin, Michel Hidiroglou, Wisner Jocelyn, Guy Laflamme, Danielle Lalande, Normand Laniel, Danielle Lebrasseur, Josée Morel, Christian Nadeau, Stephen Rathwell, Martin Renaud, Laurent Roy, Craig Seko, Michelle Simard, Hélène St-Jean, Pierre St-Martin, Alain Théberge, Brad Thomas, Jocelyn Tourigny et Johanne Tremblay. Christine Larabie a vu à la coordination de la production.

Le dixième symposium annuel de Statistique Canada, qui a pour thème les enquêtes auprès des établissements, s'est tenu du 27 au 30 juin 1993 à Buffalo, New York. Le onzième symposium aura lieu à l'automne de 1994 et portera sur la restructuration pour les organismes statistiques.

Comité organisateur du Symposium 92

Août 1993

LA SÉRIE DES SYMPOSIUMS DE STATISTIQUE CANADA

1984 - L'analyse des données d'enquête

1985 - Les statistiques sur les petites régions

1986 - Les données manquantes dans les enquêtes

1987 - Les utilisations statistiques des données administratives

1988 - Les répercussions de la technologie de pointe sur les enquêtes

1989 - L'analyse des données dans le temps

1990 - Mesure et amélioration de la qualité des données

1991 - Questions spatiales liées aux statistiques

1992 - Conception et analyse des enquêtes longitudinales

1993 - International Conference on Establishment Surveys

LA SÉRIE DES SYMPOSIUMS INTERNATIONAUX DE STATISTIQUE CANADA
RENSEIGNEMENTS CONCERNANT LA COMMANDE DES RECUEILS

Utilisez le bon de commande sur cette page pour commander des copies additionnelles du recueil du Symposium 92: "Conception et analyse des enquêtes longitudinales". Vous pouvez aussi commander les recueils des derniers symposiums. Une fois complétée, retournez la formule à:

RECUEIL DU SYMPOSIUM 92
STATISTIQUE CANADA
DIVISION DES OPÉRATIONS FINANCIÈRES
IMMEUBLE R.H. COATS, 6e ÉTAGE
PARC TUNNEY
OTTAWA (ONTARIO)
K1A 0T6
CANADA

Veuillez inclure le paiement avec votre commande (chèque ou mandat, en dollars canadiens ou l'équivalent, à l'ordre du "Receveur général du Canada - Recueil du Symposium 92").

RECUEIL DU SYMPOSIUM: NUMÉROS DISPONIBLES

1987 -	Les utilisations statistiques des données administratives - FRANÇAIS	____ @ $10 CHACUN
1987 -	Les utilisations statistiques des données administratives - ANGLAIS	____ @ $10 CHACUN
1987 -	ENSEMBLE DE 1 ANGLAIS ET 1 FRANÇAIS	____ @ $12 L'ENSEMBLE
1988 -	Les répercussions de la technologie de pointe sur les enquêtes - BILINGUE	____ @ $10 CHACUN
1989 -	L'analyse des données dans le temps - BILINGUE	____ @ $15 CHACUN
1990 -	Mesure et amélioration de la qualité des données - FRANÇAIS	____ @ $18 CHACUN
1990 -	Mesure et amélioration de la qualité des données - ANGLAIS	____ @ $18 CHACUN
1991 -	Questions spatiales liées aux statistiques - FRANÇAIS	____ @ $20 CHACUN
1991 -	Questions spatiales liées aux statistiques - ANGLAIS	____ @ $20 CHACUN
1992 -	Conception et analyse des enquêtes longitudinales - FRANÇAIS	____ @ $22 CHACUN
1992 -	Conception et analyse des enquêtes longitudinales - ANGLAIS	____ @ $22 CHACUN

S.V.P. AJOUTEZ LA TAXE SUR LES PRODUITS ET SERVICES (7%)
(Résidents du Canada seulement) $ ____

MONTANT TOTAL DE LA COMMANDE $ ____

S.V.P. INCLURE VOTRE ADRESSE COMPLÈTE AVEC VOTRE COMMANDE !

NOM _____

ADRESSE _____

_____

VILLE _____ PROV/ÉTAT _____ PAYS _____

CODE POSTAL _____ TÉLÉPHONE (____)_____ FAX _____

S.V.P. Notez: Chaque participant au Symposium 92 qui n'est pas un employé de Statistique Canada recevra une copie gratuite du recueil du Symposium 92.

CONCEPTION ET ANALYSE DES ENQUÊTES LONGITUDINALES

TABLE DES MATIÈRES[1]

[1] Dans le cas de co-auteurs, le nom de l'orateur est imprimé en caractères gras.

ALLOCUTION D'OUVERTURE

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

ALLOCUTION D'OUVERTURE

G.J. Brackstone[1]

Au nom de Statistique Canada, je vous souhaite la bienvenue au Symposium 92. Bienvenue à Ottawa et, pour plusieurs d'entre vous, bienvenue au Canada. Il s'agit du neuvième de la série de symposiums annuels de Statistique Canada traitant de questions méthodologiques. Dans la préparation de ces symposiums, nous avons généralement bénéficié de l'aide et de l'appui d'autres organisations. Cette année, nous avons accueilli avec grand plaisir la participation du Laboratoire de recherche en statistique et probabilité de l'Université Carleton et de l'Université d'Ottawa, ainsi que de la Direction d'hygiène du milieu de Santé et Bien-être social Canada, qui parrainent avec nous ce symposium.

Comme je l'ai mentionné, c'est le neuvième d'une série de symposiums qui ont traité d'une vaste gamme de sujets qui intéressent les méthodologistes d'enquête, les analystes statistiques et les statisticiens des gouvernements. La liste complète des sujets des symposiums précédents est incluse dans la brochure; j'omettrai donc de vous en faire la lecture. Le sujet du symposium de cette année est "la conception et l'analyse des enquêtes longitudinales". Avant de faire quelques observations sur les raisons du choix de ce sujet, je voudrais vous expliquer brièvement les objectifs que nous visons en parrainant de telles conférences.

Ce que nous voulons, c'est offrir une tribune permettant la discussion et l'échange d'idées; contribuer à la discipline de la méthodologie d'enquête en favorisant la coopération professionnelle; réunir des praticiens et des théoriciens de milieux variés, qui abordent des problèmes d'intérêt commun selon des expériences et des perspectives différentes. Et encore bien d'autres choses. En bref, notre but est l'avancement de la science statistique et de la pratique des enquêtes. Tout cela est parfaitement vrai; ce sont là des objectifs admirables et nous souhaitons que ce symposium nous aidera à les atteindre. Mais soyons honnêtes au sujet de nos motivations: nous n'agissons pas par pur altruisme; nous voulons aussi, dans une grande mesure, servir nos propres intérêts.

Premièrement, nous faisons face à des problèmes et à des défis concrets dans la réalisation des programmes de Statistique Canada, et nous voulons obtenir l'aide des experts les meilleurs et les plus expérimentés en la matière. Et je constate avec joie que bon nombre des spécialistes les meilleurs et les plus expérimentés du domaine des enquêtes longitudinales sont avec nous cette semaine. Nous sommes très reconnaissants envers ceux d'entre vous qui, venant parfois de très loin, se sont déplacés jusqu'ici pour partager leurs expériences avec nous. Je sais que nous allons en profiter; et j'espère qu'il en sera de même pour vous.

Deuxièmement, ces symposiums servent aussi nos intérêts du fait qu'ils permettent à une grande partie de nos employés d'améliorer et d'élargir leur compréhension des questions reliées à leur travail. Nos spécialistes peuvent venir ici écouter des experts de classe mondiale en bien plus grand nombre qu'ils ne le pourraient s'ils devaient assister à des conférences professionnelles tenues ailleurs.

C'est donc en grande partie dans notre propre intérêt que nous organisons ces symposiums, mais à en juger par le soutien et la participation que nous obtenons toujours, je suis convaincu que de nombreux autres participants en tirent avantage. Le temps est maintenant venu d'aborder le sujet de cette conférence, soit "la conception et l'analyse des enquêtes longitudinales".

---

[1] G.J. Brackstone, Statisticien en chef adjoint, Secteur de l'informatique et de la méthodologie, 26-J, Immeuble R.H. Coats, Parc Tunney, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

Le sujet de l'an dernier était les "questions spatiales liées aux statistiques". Nous sommes donc passés de la dimension spatiale à la dimension temporelle. L'an dernier, nous nous sommes intéressés aux structures géographiques, aux aspects touchant les distributions dans l'espace. Nous avons centré notre attention sur les latitudes et les longitudes. Cette année nous ne nous concentrons que sur les longitudes - ou du moins sur les enquêtes longitudinales. Nous nous donnons toutefois beaucoup de latitude dans l'interprétation du mot "longitudinal".

D'une certaine façon, le présent symposium peut être considéré comme le troisième volet d'une série de conférences sur des sujets connexes. La série s'est amorcée en 1986 à Washington avec la conférence sur les enquêtes par panel. Un livre de la série Wiley sur les enquêtes par panel fut produit suite à cette conférence. Trois ans plus tard, en 1989, le symposium tenu ici portait sur "l'analyse des données dans le temps" et présentait beaucoup de travaux reliés aux enquêtes longitudinales. Maintenant, trois ans plus tard, nous revenons aux enquêtes longitudinales. Nous avons perdu certains participants en cours de route, mais je constate que de nombreuses personnes qui assistaient à la conférence de 1986 sont encore présentes parmi nous, et nous feront encore profiter de leur expérience.

Les plans d'enquête consistant à effectuer des mesures sur le même échantillon à des moments différents ne sont pas nouveaux. Les échantillons chevauchants font partie du plan des enquêtes répétées depuis de nombreuses années. L'utilisation des échantillons chevauchants a été motivée par le besoin de disposer d'estimations efficaces du changement et par des considérations opérationnelles telles que le coût de l'ajout de nouveaux répondants à une enquête.

Ces dernières années, un intérêt de plus en plus marqué s'est manifesté à l'égard des plans d'enquête dans lesquels la mesure répétée de caractéristiques du même ensemble d'unités est un objectif de l'enquête en soi, plutôt qu'un choix exercé seulement pour les besoins du plan. Si un tel intérêt a été soulevé, c'est notamment parce qu'on s'est rendu compte que de nombreuses questions importantes en matière de politique économique et sociale exigent la cueillette de renseignements sur les conséquences de processus dynamiques pour des personnes ou des entreprises individuelles, et pas seulement sur les effets généraux ou globaux de ces processus. Parmi les principaux exemples, sur lesquels nous en apprendrons davantage au cours des prochains jours, notons la pauvreté et l'état de santé, pour lesquels, aux fins de l'élaboration des politiques et des programmes, il est probablement beaucoup plus important de savoir comment et pourquoi les gens entrent dans ces états ou en sortent que de simplement connaître les variations du niveau global de pauvreté ou de mauvaise santé de la population. La compréhension de telles dynamiques exige que des mesures soient faites sur les mêmes unités pendant de longues périodes.

De telles mesures exigent, semble-t-il, soit une enquête longitudinale qui permettra d'obtenir des réponses des mêmes unités à différentes occasions, soit une enquête rétrospective. En fait, toute enquête longitudinale comporte un certain élément rétrospectif, si bien qu'en pratique, les plans d'enquête doivent trouver un compromis entre ces deux méthodes, par le choix d'une période optimale pour laquelle les enquêtés doivent se rappeler des données. Mais il ne faut pas oublier qu'il existe aussi un troisième moyen important d'obtenir des données longitudinales: le couplage longitudinal de dossiers administratifs. Nous avons une certaine expérience de cette méthode à Statistique Canada.

Au cours des dix dernières années, des enquêtes longitudinales de vaste envergure, à plusieurs vagues, portant sur les conditions socio-économiques et le comportement des ménages ont été mises sur pied dans plusieurs pays européens. Aux États-Unis, on a commencé en 1983, dans le cadre de l'enquête «Survey of Income and Program Participation» (SIPP), à recueillir des données relatives au revenu, ainsi qu'à l'admissibilité et à la participation à divers programmes de soutien du revenu.

Ayant eu le privilège de siéger à un comité consultatif du Census Bureau sur l'enquête SIPP, en compagnie de notre conférencier qui prononcera le discours-programme, j'étais en bonne position pour apprécier les énormes défis se dressant devant le Census Bureau dans cette entreprise. Présomptueusement, j'osais affirmer que nous n'étions pas confrontés à ce genre de problèmes à Statistique Canada. Mais maintenant nous le sommes, et c'est en partie pourquoi le présent symposium a lieu. J'espère que nous avons appris beaucoup des travaux d'avant-garde accomplis par le Census Bureau pour les besoins de l'enquête SIPP.

Nos premiers efforts, modestes, dans le domaine des enquêtes longitudinales, se sont amorcés avec l'Enquête sur l'activité à la fin des années 1980. Une enquête plus ambitieuse, l'Enquête sur la dynamique du travail et du revenu (EDTR), sera lancée l'an prochain. Dans cette enquête, nous suivrons des personnes et des familles pendant cinq ou six ans, en recueillant des renseignements sur l'expérience du marché du travail, le revenu et la situation familiale.

Nous planifions également une importante enquête longitudinale sur la santé, dont la première vague de collecte de données aura lieu sur le terrain en 1994. L'enquête vise à fournir de l'information sur les effets, dans le temps, des conditions socio-économiques et du mode de vie sur le bien-être des individus et le recours au système de soins de santé.

Les enquêtes longitudinales soulèvent de nombreuses questions méthodologiques. Le programme qui a été élaboré pour les trois prochains jours reflète la diversité de ces questions, ainsi que l'intérêt manifesté, à l'échelle internationale, envers les enquêtes longitudinales. Les communications aborderont les questions de la collecte des données, de la sélection et de la pondération des échantillons, de la non-réponse, de l'analyse des données longitudinales (domaine qui ouvre la voie à des méthodes qui ne sont pas traditionnellement utilisées dans les enquêtes), et de la qualité des données provenant des enquêtes longitudinales. Nous assisterons aussi à des présentations sur des applications des méthodes propres aux enquêtes longitudinales.

Au programme, nous avons des conférenciers du Royaume-Uni, de la France, de l'Allemagne et du Luxembourg, ainsi que du Canada et des États-Unis. Le secteur privé, les milieux universitaires et les services gouvernementaux sont tous représentés.

Bref, je vous remercie encore d'être venus partager vos connaissances et votre expérience avec nous. J'espère qu'à la fin du symposium, vous aurez l'impression d'en avoir tiré le même profit que nous, à Statistique Canada, allons certainement en tirer.

DISCOURS - PROGRAMME

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

ENQUÊTES PAR PANEL: AJOUT D'UNE QUATRIÈME DIMENSION

G. Kalton[1]

RÉSUMÉ

Les enquêtes qui consistent à recueillir des données dans le temps peuvent viser de nombreux objectifs. Dans la première moitié de la présente communication, nous examinons diverses options de plans d'enquête - enquêtes à passages répétés, enquêtes par panel, enquêtes par panel avec renouvellement et enquêtes à panel fractionné - pouvant permettre d'atteindre ces objectifs. La deuxième moitié est axée sur les enquêtes par panel. Nous y traitons des décisions qui doivent être prises au moment de la conception d'une enquête par panel, des problèmes posés par la non-réponse aux différentes vagues, du biais de conditionnement et de l'effet de lisière, ainsi que de certaines méthodes permettant l'analyse longitudinale des données d'enquête par panel.

MOTS-CLÉS: Enquêtes par panel; enquêtes par panel avec renouvellement; enquêtes à passages répétés; érosion du panel; biais de conditionnement; effet de lisière; analyse longitudinale.

1. INTRODUCTION

Les populations visées par les enquêtes changent constamment dans le temps, aussi bien parce que leur composition évolue qu'en raison de modifications des caractéristiques de leurs membres. Des changements de composition surviennent quand des membres entrent dans la population au moment de la naissance (ou de l'atteinte de l'âge adulte), ou encore lorsqu'ils immigrent ou qu'ils sortent d'une institution (si la population étudiée ne comprend pas les personnes en institution); de même, les décès, l'émigration et l'entrée en institution sont des causes de changement. Les caractéristiques changent, par exemple, lorsque des personnes mariées divorcent, ou lorsque le revenu mensuel d'une personne passe de $2 000 à $2 500. Ces changements touchant la population sont à l'origine de toute une série d'objectifs des analyses de l'évolution des données d'enquête en fonction du temps. La présente communication passe en revue les plans d'enquête qui produisent les données nécessaires à l'atteinte de ces divers objectifs.

Cette communication se divise en deux parties. La première présente les aspects généraux de la réalisation d'enquêtes longitudinales, notamment les objectifs de ces enquêtes et les types de plans d'enquête possibles. Cet examen fait l'objet de la section 2. La deuxième partie, qui est le corps principal de la présente communication, s'intéresse au cas particulier des enquêtes par panel, qui consistent à suivre le même échantillon d'unités au fil du temps. Les différents aspects de la conception, de l'exécution et de l'analyse d'une enquête par panel sont exposés à la section 3. La section 4 présente quelques remarques en guise de conclusion.

2. ENQUÊTES LONGITUDINALES

La présente section décrit de façon générale les objectifs analytiques propres aux données qui évoluent dans le temps, les plans d'enquête qui conviennent à ces données et la mesure dans laquelle les différents plans peuvent satisfaire aux divers objectifs. L'exposé s'inspire fortement de celui de Duncan et Kalton (1987), qui contient un traitement plus détaillé de ces questions.

L'évolution des caractéristiques et de la composition de la population dans le temps est à la source de toute une gamme d'objectifs ayant trait aux enquêtes longitudinales. Ces objectifs sont, notamment:

---

[1] G. Kalton, Westat, 1650 Research Blvd., Rockville (Maryland), É.-U. 20850.

(a) l'estimation des paramètres de la population (p. ex. la proportion de la population vivant sous le seuil de la pauvreté) à divers points dans le temps;

(b) l'estimation des valeurs moyennes de paramètres de la population dans le temps (p. ex. moyenne annuelle de la consommation quotidienne de fer);

(c) l'estimation de variations nettes, c'est-à-dire de variations à des niveaux d'agrégation élevés (p. ex. variation de la proportion de chômeurs d'un mois à l'autre);

(d) l'estimation de variations brutes et d'autres facettes de changements touchant les individus (p. ex. proportion de personnes qui étaient sous le seuil de la pauvreté une année et qui ne l'étaient plus l'année suivante);

(e) l'agrégation sur une certaine période de données relatives aux individus (p. ex. sommation des revenus de douze mois pour obtenir le revenu annuel);

(f) l'obtention de données sur des événements qui se produisent au cours d'une période donnée (p. ex. tomber en chômage) et sur leurs caractéristiques (p. ex. durée des périodes de chômage);

(g) l'accumulation progressive d'échantillons, notamment des échantillons de populations rares (p. ex. les femmes qui deviennent veuves);

(h) la tenue à jour d'un échantillon de membres d'une population rare observé à un moment particulier (p. ex. scientifiques et ingénieurs relevés dans une enquête à grande échelle à un moment donné).

Un certain nombre de plans d'enquête ont été mis au point pour la collecte des données nécessaires à l'atteinte de ces objectifs. En voici une liste:

- *Enquêtes à passages répétés*. Une enquête à passages répétés est une série d'enquêtes transversales distinctes réalisées à différents moments. On ne cherche aucunement à s'assurer que les éléments de l'échantillon soient les mêmes d'un passage à l'autre. Les éléments de l'échantillon sont tirés d'une population définie de la même manière à chaque enquête (p. ex. mêmes frontières géographiques et limites d'âge) et les questions, pour une bonne part, sont identiques d'un passage à l'autre.

- *Enquêtes par panel*. Une enquête par panel consiste à recueillir des données, à différents moments, auprès du même échantillon de répondants.

- *Enquête par panel à passages répétés*. Une enquête par panel à passages répétés est formée d'une série d'enquêtes par panel ayant chacune une durée fixe. Il peut n'y avoir aucun chevauchement des périodes visées par les différents panels, par exemple si un panel ne commence qu'au moment où le précédent se termine (ou plus tard), ou encore il peut y avoir un chevauchement, si deux panels ou plus couvrent partiellement la même période.

- *Enquête par panel avec renouvellement*. À strictement parler, une enquête par panel avec renouvellement est équivalente à une enquête par panel à passages répétés dans laquelle il y a un chevauchement. Dans les deux cas, la durée d'un panel est limitée, et deux panels ou plus sont en cours d'examen au même moment. Toutefois, il est utile de faire une distinction entre ces deux plans, car ils ne visent pas les mêmes objectifs. Les enquêtes par panel avec renouvellement sont largement utilisées pour produire une série d'estimations transversales et d'estimations de variations nettes (p. ex. taux de chômage et évolution de ces taux), tandis que les enquêtes par panel à passages répétés avec chevauchement accordent en outre beaucoup d'importance à des mesures longitudinales (p. ex. durée des périodes de chômage). Par conséquent les enquêtes par panel à passages répétés ont tendance à être de plus longue durée et à comporter moins de panels en cours de traitement à un moment quelconque que les enquêtes par panel avec renouvellement.

- *Enquête à panel fractionné*. Une enquête à panel fractionné est la combinaison d'une enquête par panel et d'une enquête à passages répétés ou d'une enquête par panel avec renouvellement.

Le choix du plan dans une situation particulière dépend des objectifs que l'on veut atteindre. Certains plans sont supérieurs vis-à-vis de certains objectifs, mais moins performants à d'autres égards. Il y a des plans qui ne satisfont aucunement à certains objectifs. Pour un examen détaillé, voir Duncan et Kalton (1987).

Le principal atout d'une enquête à passages répétés est qu'elle comporte le prélèvement d'un nouvel échantillon à chaque passage, de sorte que chaque enquête transversale se fonde sur un échantillon probabiliste de la population qui existe à ce moment précis. Une enquête par panel se fonde sur un échantillon tiré de la population qui existait au début du panel. Bien que, parfois, des efforts soient faits pour ajouter à un panel des échantillons de nouveaux membres à des stades ultérieurs, une telle mise à jour se révèle généralement difficile et est empreinte d'imperfections. De plus, les pertes dues à la non-réponse qui sont enregistrées à mesure qu'un panel vieillit accentuent le problème du biais de non-réponse dont sont entachées, à des points ultérieurs, les estimations de paramètres tirées du panel. Pour ces raisons, les enquêtes à passages répétés sont supérieures aux enquêtes par panel pour produire des estimations transversales et des moyennes de ces dernières (objectifs (a) et (b)). Dans le cas des moyennes d'estimations transversales, un autre facteur à prendre en considération est la corrélation qui existe entre les valeurs des variables de l'enquête pour la même personne à différents moments. Si cette corrélation est positive, comme c'est le cas généralement, il en résulte un accroissement des erreurs-types des moyennes des estimations transversales provenant d'une enquête par panel. Ce facteur joue donc aussi en faveur des enquêtes à passages répétés par rapport aux enquêtes par panel, lorsqu'on s'intéresse à des moyennes d'estimations transversales.

La représentation plus fidèle des échantillons d'une enquête à passages répétés à des stades ultérieurs semblerait aussi favoriser ce type d'enquêtes par rapport aux enquêtes par panel pour l'estimation de la variation nette (en supposant qu'on veut mesurer cette dernière en tenant compte des changements touchant aussi bien la composition que les caractéristiques de la population). Toutefois, dans ce cas, les corrélations positives des valeurs des variables de l'enquête pour la même personne au fil du temps ont pour effet de diminuer les erreurs-types des estimations de la variation nette établie d'après une enquête par panel. Par conséquent, la présence de ces corrélations joue en faveur de l'enquête par panel pour la mesure de la variation nette.

Les avantages principaux des enquêtes par panel résident dans leur capacité de mesurer la variation brute, ainsi que de permettre l'agrégation, au fil du temps, de données sur les membres de l'échantillon (objectifs (d) et (e)). Les enquêtes à passages répétés ne peuvent répondre à ces objectifs. L'énorme potentiel d'analyse offert par la mesure de variations touchant les mêmes personnes est la principale raison incitant à recourir à une enquête par panel.

Les enquêtes à passages répétés permettent de recueillir des données sur des événements qui surviennent au cours d'une période donnée, ou encore sur la durée d'événements (p. ex. des périodes de maladie) au moyen de questions rétrospectives. Toutefois, les questions portant sur le passé, parce que les répondants éprouvent de la difficulté à se souvenir des dates, engendrent souvent un important problème d'erreur de réponse et font naître le risque d'un biais de télescopage. Une enquête par panel utilisant, pour les événements à étudier, une période de référence correspondant à l'intervalle entre les vagues de collecte des données peut éliminer le problème du télescopage, en utilisant l'interview précédente pour délimiter les événements (ainsi, on ne tiendra pas compte d'une maladie déclarée au cours de la présente interview si la même maladie avait été déclarée à l'interview précédente). De même, une enquête par panel peut permettre de déterminer la durée d'un événement d'après les vagues successives de collecte des données, en limitant la période de référence à l'intervalle entre les vagues.

Des collectes de données répétées au fil du temps peuvent être l'occasion de constituer progressivement un échantillon de membres d'une population rare, par exemple des personnes ayant une maladie chronique rare ou des personnes ayant récemment vécu un deuil. Les enquêtes à passages répétés peuvent être utilisées à cette fin et permettre de créer des échantillons de toutes sortes de populations rares. Les enquêtes par panel, toutefois, ne peuvent permettre d'accumuler que des événements rares nouveaux (p. ex. des deuils) et non des caractéristiques rares stables (p. ex. des personnes ayant une maladie chronique). Si un échantillon de membres ayant une caractéristique rare stable (p. ex. personnes ayant un doctorat) a déjà été établi, une enquête par panel peut être utile à la tenue à jour de cet échantillon, grâce à un apport approprié de nouveaux membres au moment de vagues ultérieures (voir, par exemple, Citro et Kalton 1989).

Les enquêtes par panel avec renouvellement visent principalement l'estimation des niveaux courants et de la variation nette (objectifs (a) et (c)). Dans de telles enquêtes, les éléments font généralement partie du panel pendant une courte période seulement. Par exemple, les membres de l'échantillon de l'enquête mensuelle sur la population active du Canada demeurent dans l'échantillon pendant une période de six mois seulement. Par conséquent, la mesure dans laquelle on peut évaluer les changements individuels et obtenir une agrégation des données dans le temps est limitée par la courte durée des panels. Une caractéristique spéciale des enquêtes par panel avec renouvellement réside dans la possibilité de recourir à l'estimation composite pour améliorer la précision tant des estimations transversales que des estimations de la variation nette (voir Binder et Hidiroglou 1988).

Une enquête à panel fractionné, parce qu'elle est une combinaison d'une enquête par panel et d'une enquête à passages répétés ou d'une enquête par panel avec renouvellement, peut offrir les avantages de chacune. Toutefois, si un plafond est imposé aux ressources globales disponibles, la taille de l'échantillon de chaque composante sera nécessairement plus faible que si une seule composante était utilisée. En particulier, les estimations des variations brutes et d'autres mesures des variations individuelles provenant d'une enquête à panel fractionné se fonderont sur un échantillon plus petit que ce n'aurait été le cas si toutes les ressources avaient été consacrées uniquement à la composante «enquête par panel».

Lorsqu'on compare différents plans d'enquêtes longitudinales, les coûts doivent être pris en considération. Par exemple, les enquêtes par panel permettent d'éviter les coûts de la sélection répétée d'échantillons qu'exigent les enquêtes à passages répétés, mais comportent des coûts pour le dépistage et le suivi des déplacements des membres de l'échantillon, et parfois aussi pour l'offre d'incitatifs destinés à s'assurer la fidélité des membres du panel (voir la section 3). Si deux plans peuvent l'un et l'autre répondre aux objectifs de l'enquête, les coûts relatifs pour des niveaux de précision donnés des estimations de l'enquête doivent être examinés.

3. ENQUÊTES PAR PANEL

Les mesures répétées portant sur le même échantillon que permettent les enquêtes par panel confèrent à ces enquêtes un avantage clé, sur le plan de l'analyse, par rapport aux enquêtes à passages répétés. Les mesures de la variation brute et d'autres composantes de la variation individuelle qu'on peut extraire des données d'une enquête par panel sont la source d'une bien meilleure compréhension des processus sociaux que celle que peut procurer une série de clichés transversaux indépendants. Le potentiel des données longitudinales tirées des enquêtes par panel est depuis longtemps reconnu (voir, par exemple, Lazarsfeld et Fiske 1938; Lazarsfeld 1948) et, depuis de nombreuses années, des enquêtes par panel sont réalisées dans de nombreux domaines. Les sujets suivants, par exemple, ont fait l'objet d'enquêtes par panel: croissance et développement humains, délinquance juvénile, consommation de drogues, victimisation, comportement des électeurs, études de marketing sur les dépenses de consommation, choix d'études et de carrières, retraite, santé, frais médicaux. (Voir Wall et Williams (1970) pour un examen d'études par panel réalisées il y a longtemps sur la croissance et le développement humains, Boruch et Pearson (1988) pour des descriptions de certaines enquêtes par panel menées aux État-Unis, et le Subcommittee on Federal Longitudinal Surveys (1986) pour des descriptions d'enquêtes par panel fédérales américaines.) On a pu constater ces dernières années un important gain d'intérêt pour les enquêtes par panel dans de nombreux secteurs, notamment dans le domaine de l'information économique sur les ménages. L'enquête permanente américaine «Panel Study of Income Dynamics» a commencé en 1968 (voir Hill 1992, pour une description de la PSID) et des enquêtes par panel de longue durée du même genre ont été amorcées dans de nombreux pays européens au cours de la dernière décennie. Le U.S. Bureau of the Census a lancé l'enquête «Survey of Income and Program Participation» (SIPP) en 1983 (Nelson et coll. 1985; Kasprzyk 1988; Jabine et coll. 1990), et Statistique Canada a introduit l'Enquête sur la dynamique du travail et du revenu (EDTR) en 1993. L'intérêt accru manifesté à l'égard des enquêtes par panel s'est aussi accompagné de la parution d'un nombre croissant d'ouvrages sur la méthodologie de telles enquêtes; citons, à titre d'exemples récents, Kasprzyk et coll. (1989), Magnusson et Bergman (1990) et Van de Pol (1989).

La présente section décrit les principaux aspects qui interviennent dans la conception et l'analyse des enquêtes par panel. Nous nous intéressons principalement aux enquêtes par panel à passages répétés de durée fixe, comme l'enquête SIPP et l'enquête EDRT, mais l'essentiel de notre examen s'applique, de façon générale, à toutes les formes d'enquêtes par panel.

3.1 Plan d'une enquête par panel: les choix à faire

La dimension «temps» est une dimension de complexité additionnelle que possède une enquête par panel par rapport à une enquête transversale. En sus de toutes les décisions qui doivent être prises dans le cadre de la planification d'une enquête transversale, une vaste gamme de choix additionnels doivent être faits pour une enquête par panel. Voici les principales décisions à prendre:

- *Durée du panel.* Plus le panel est de longue durée, plus grande est la richesse des données pour les fins de l'analyse longitudinale. Par exemple, plus le panel durera longtemps, plus il y aura de périodes de chômage commençant pendant la durée du panel qui se termineront avant la fin du panel, et donc plus l'estimation de la fonction de survie pour de telles périodes sera précise. En revanche, plus le panel est de longue durée, plus il est difficile de garder un échantillon transversal représentatif aux vagues finales du panel, en raison aussi bien de l'érosion de l'échantillon que des difficultés de mise à jour de l'échantillon en fonction des nouveaux venus dans la population.

  Il peut parfois être bénéfique de faire varier la durée du panel selon divers types de membres. Ainsi, lorsque les objectifs de l'analyse l'exigent, les membres du panel ayant certaines caractéristiques (p. ex. les membres d'une minorité) ou qui vivent certaines événements au cours de la durée du panel normal (p. ex. un divorce) peuvent être gardés dans le panel pour des périodes d'observation plus longues.

- *Durée de la période de référence.* La fréquence de collecte des données dépend de la capacité des répondants de se souvenir de l'information demandée. Ainsi, la PSID, qui comporte des vagues annuelles de collecte des données, exige des répondants qu'ils se souviennent des événements survenus au cours de l'année civile précédente, tandis que l'enquête SIPP, avec des vagues de collecte des données aux quatre mois, exige que l'on se souvienne de ce qui s'est passé au cours des quatre mois précédents. Plus la période de référence est longue, plus le risque d'erreur de mémoire est élevé.

- *Nombre de vagues.* Dans la plupart des cas, le nombre de vagues de collecte des données est déterminé par la durée du panel et la durée de la période de référence. Plus il y a de vagues, plus on risque de subir une érosion du panel et des effets de conditionnement, et plus lourd est le fardeau imposé aux répondants.

- *Panels chevauchants ou non chevauchants.* Dans le cas d'une enquête par panel à passages répétés de durée fixe, une décision doit être prise quant au chevauchement des panels. Prenons le cas, par exemple, d'une proposition d'un groupe d'étude du National Research Council selon lequel l'enquête SIPP devrait comporter un panel de quatre ans (Citro et Kalton 1993). Une possibilité est d'avoir un panel de quatre ans, et de commencer un nouveau panel dès que le précédent prend fin. On pourrait aussi établir des panels de quatre ans dont un nouveau serait amorcé tous les deux ans. Autre possibilité, on pourrait avoir des panels de quatre ans, mais en commencer un nouveau chaque année.

  Un plan comportant des panels non chevauchants a l'avantage d'être simple, car un seul panel à la fois est en cours de traitement. Il produit aussi un vaste échantillon pour l'analyse longitudinale; par exemple, les panels non chevauchants peuvent être approximativement deux fois plus gros que ceux d'un plan comportant en tout temps deux panels chevauchants. Toutefois, cet avantage d'une taille d'échantillon plus grande qu'offrent les panels non chevauchants ne vaut pas pour les estimations transversales, car s'il y a plusieurs panels simultanés, les données qui concernent un moment précis peuvent être combinées pour les besoins de l'estimation transversale. En outre, les estimations visant des données transversales observées vers la fin d'un panel, dans le cas d'un plan sans chevauchement, sont davantage susceptible de souffrir d'un biais dû à l'érosion ou d'un biais de conditionnement, ou encore de ne pas tenir compte pleinement des nouveaux éléments de la population, que ce n'est le cas pour un plan avec chevauchement, dans lequel l'un des panels est d'origine plus récente. En outre, le plan avec chevauchement permet l'examen de tels biais grâce à une comparaison des résultats des deux panels au cours d'une période donnée, ce que ne permet pas le plan sans chevauchement. Une autre limite du plan sans chevauchement est qu'il est susceptible de ne pas mesurer adéquatement l'effet d'événements comme des modifications législatives. Par exemple, si de nouvelles mesures législatives entrent en vigueur au cours de la dernière année d'un panel non chevauchant, il y aura peu de possibilité d'évaluer leur effet en comparant les situations des mêmes personnes avant, et pendant un

certain temps après, l'instauration des nouvelles mesures. Avec des panels chevauchants, l'un des panels offrira une fenêtre d'observation plus large.

- *Taille du panel*. Pour un niveau annuel donné de ressources, la taille d'échantillon de chaque panel est déterminée par les facteurs qui précèdent. On peut obtenir un panel de grande taille pour l'analyse longitudinale en allongeant la période de référence et en utilisant un plan sans chevauchement. Pour des estimations transversales, l'allongement de la période de référence peut permettre d'augmenter la taille de l'échantillon, mais le recours à un plan sans chevauchement n'a aucun effet semblable.

La liste ci-dessus détermine les principaux paramètres d'un plan d'enquête par panel, mais il existe un certain nombre d'autres facteurs qui doivent aussi être examinés:

- *Mode de collecte des données*. Comme pour n'importe quelle enquête, il faut décider si les données seront recueillies par des interviews sur place, par téléphone ou par un questionnaire rempli par le répondant, et si l'on utilisera une méthode assistée par ordinateur (IPAO -- interview sur place assistée par ordinateur ou ITAO -- interview téléphonique assistée par ordinateur). Dans le cas d'une enquête par panel, une telle décision doit être prise à l'égard de chaque vague de collecte des données, car on peut vouloir faire varier les modes de collecte (par exemple, interview sur place à la première vague pour établir un contact et créer un lien, et interviews par téléphone ou par questionnaire postal à certaines vagues subséquentes). Si le mode de collecte est susceptible de changer d'une vague à l'autre, il faut examiner la question de la comparabilité des données entre les vagues. Parfois, un changement de mode peut entraîner un changement d'interviewer, par exemple si l'on passe d'une interview sur place à une interview téléphonique assistée par ordinateur, effectuée à partir d'un endroit central. Dans un tel cas, l'effet du changement d'interviewer sur la volonté du répondant de continuer de participer au panel et sur la comparabilité des réponses entre les vagues doit aussi faire l'objet d'un examen attentif.

- *Interview avec rétro-information*. Dans les enquêtes par panel, il est possible de rappeler aux participants les réponses données à des vagues antérieures. Cette technique d'interview avec rétro-information permet d'obtenir des réponses plus cohérentes entre les vagues, mais risque de produire un niveau de cohérence exagérément élevé. La facilité d'application de l'interview avec rétro-information dépend de la durée de l'intervalle entre les vagues et du mode de collecte des données. Le traitement des réponses d'une vague de manière à pouvoir les rappeler à la vague suivante est plus facile à accomplir si l'intervalle entre les vagues est long et si des techniques d'interview assistée par ordinateur sont employées.

- *Incitatifs*. De l'argent, ou encore d'autres types d'incitatifs (p. ex. tasses, calculatrices, sacs-repas) peuvent être offerts aux membres de l'échantillon pour qu'ils demeurent fidèles à l'enquête. Dans le cas d'une enquête par panel, les incitatifs peuvent servir non seulement à obtenir la participation initiale du répondant, mais aussi à s'assurer sa coopération pendant toute la durée du panel. Le moment qui convient le mieux à l'offre d'incitatifs dans une enquête par panel (p. ex. à la première vague, à une vague intermédiaire ou à la dernière vague du panel) est une question qui reste à résoudre. Les chercheurs utilisant les enquêtes par panel envoient souvent aux répondants un bulletin périodique relatif à l'enquête, donnant les plus récents faits saillants des résultats, à la fois pour susciter leur bonne volonté à l'égard de l'enquête et pour maintenir un contact avec eux (voir ci-dessous). L'envoi de cartes de souhaits au moment de l'anniversaire de naissance des répondants est un autre moyen fréquemment employé.

- *Règles relatives aux répondants*. Les données des enquêtes sont souvent recueillies auprès de répondants substituts lorsque les répondants devant être interviewés ne sont pas disponibles. Dans le cas d'une enquête par panel, il est alors possible que les répondants ne soient pas les mêmes d'une vague à l'autre, ce qui nuit à la comparabilité des données entre les vagues. Les règles relatives aux répondants, dans une enquête par panel, doivent tenir compte de ce facteur.

- *Plan d'échantillonnage*. La nature longitudinale d'une enquête par panel doit être prise en considération au moment de construire le plan d'échantillonnage visant la première vague. Un échantillonnage par grappes est souvent employé dans les enquêtes transversales avec interviews sur place, de façon à réduire les frais de déplacement et à limiter le travail de construction de la base de sondage à l'établissement de listes d'unités de logement pour certains segments seulement. Ces avantages sont obtenus au prix d'un d'accroissement de

la variance des estimations de l'enquête, attribuable à l'utilisation de grappes. Le composition optimale des grappes dépend des divers coûts en cause et de l'homogénéité des variables de l'enquête à l'intérieur des grappes (voir, par exemple, Kish 1965). Dans le cas d'une enquête par panel, l'utilisation et l'ampleur des grappes devrait être déterminée en tenant compte du panel dans son ensemble et de toutes les vagues de collecte de données. En particulier, l'avantage d'une réduction des frais du travail sur place disparaît pour les vagues au cours desquelles l'interview téléphonique ou le questionnaire postal est utilisé. Par ailleurs, la migration de membres du panel vers des endroits se trouvant à l'extérieur des grappes originales réduit, aux vagues ultérieures, l'avantage que procurait la formation initiale des grappes en termes de réduction des coûts du travail sur place. (Toutefois, certains avantages des grappes initiales demeurent valables dans le cas de la proportion élevée des personnes mobiles qui déménagent à l'intérieur de leur propre quartier.)

Le suréchantillonnage de certains sous-groupes de la population est largement utilisé dans les enquêtes transversales, afin de fournir un nombre suffisant de membres de ces sous-groupes pour permettre des analyses séparées. Voici des exemples de tels sous-groupes: personnes à faible revenu, minorités, groupe d'âge particulier, résidents d'un certain secteur géographique. Un tel suréchantillonnage peut aussi être utile aux enquêtes par panel, mais il faut se montrer prudent dans son application. Dans le cas des panels de longue durée, la prudence se justifie notamment par le fait que les objectifs de l'enquête peuvent changer avec le temps. Le suréchantillonnage visant à répondre à un objectif énoncé au début d'un panel peut se révéler nuisible aux objectifs qui s'imposeront plus tard. Un autre motif de prudence vient du fait que bon nombre des sous-groupes étudiés sont de nature transitoire (p. ex. personnes à faible revenu, personnes vivant dans un secteur géographique donné). Le suréchantillonnage de tels sous-groupes au début d'un panel pourrait avoir une valeur limitée lors des vagues ultérieures: certaines personnes incluses dans un tel suréchantillonnage sortiront du sous-groupe, tandis que d'autre exclues du suréchantillonnage y entreront. Troisièmement, la définition du sous-groupe désiré dans le contexte d'une analyse longitudinale doit être prise en considération. Par exemple, les données de l'enquête SIPP servent à estimer la durée des périodes de participation à divers programmes sociaux. Puisque ces estimations se fondent habituellement sur de nouvelles périodes qui commencent pendant la durée du panel, il pourrait être inutile de suréchantillonner les personnes déjà bénéficiaires de programmes sociaux. Voir Citro et Kalton (1993) pour une analyse du suréchantillonnage dans le contexte de l'enquête SIPP.

- *Mise à jour de l'échantillon*. Lorsque le seul objectif d'une enquête par panel est d'effectuer une analyse longitudinale, il peut être suffisant de procéder comme pour une cohorte et de suivre simplement l'échantillon sélectionné au moment de la vague initiale. Cependant, si l'on veut aussi établir des estimations transversales, il peut être nécessaire de mettre à jour l'échantillon à chaque vague afin de tenir compte des nouveaux venus dans la population. Une mise à jour reflétant tous les types de nouveaux venus est souvent difficile, mais il est parfois possible d'établir une méthode assez simple pour incorporer certains types de nouveaux venus. Par exemple, dans un panel de personnes de tous âges, les bébés nés de femmes membres du panel une fois ce dernier amorcé peuvent être inclus comme membres du panel. La population d'inférence de l'enquête SIPP est formée des personnes âgées de 15 ans ou plus. En déterminant, dans l'échantillon initial, les membres des ménages qui n'ont pas encore 15 ans mais qui l'auront avant la fin du panel, en suivant ces derniers pendant la durée du panel et en les interviewant dès qu'ils atteignent 15 ans, on peut effectuer une mise jour d'un panel de l'enquête SIPP tenant compte de cette catégorie de nouveaux venus (Kalton et Lepkowski 1985).

Il importe également de prendre en considération les membres du panel qui sortent de la population à l'étude. Dans certains cas, le départ est irréversible (p. ex. décès), mais dans d'autres cas, il pourrait n'être que temporaire (p. ex. déménagement à l'étranger ou entrée en institution). Si l'on prend des mesures pour garder la trace des participants qui sortent du panel temporairement, ceux-ci peuvent être réadmis dès qu'ils réintègrent la population d'inférence.

Dans des enquêtes par panel comme l'enquête SIPP et la PSID, des données sont recueillies non seulement sur les membres des ménages de l'échantillon initial, mais également sur d'autres personnes - non membres de l'échantillon - avec lesquelles ils vivent lors de vagues ultérieures. L'objectif principal de la collecte de données d'enquête sur des personnes ne faisant pas partie de l'échantillon est de permettre de décrire la situation économique et sociale des membres de l'échantillon. Il convient, toutefois, de se demander si l'on devrait garder dans le panel une partie ou la totalité de ces personnes non membres de l'échantillon

lorsqu'elles ne vivent plus avec les ménages du panel. Dans certains types d'analyse, il est utile de continuer de les suivre, mais cela oblige à leur consacrer une part importante des ressources de l'enquête.

- *Dépistage et suivi*. Une difficulté qui entrave la plupart des enquêtes par panel vient du fait que certains membres du panel ont déménagé depuis la dernière vague et ne peuvent pas être retrouvés. Il y a deux façons de traiter ce problème. Premièrement, on peut tenter d'éviter qu'il ne survienne en mettant sur place un processus de suivi des membres du panel entre les vagues. Une méthode largement utilisée, lorsqu'il y a un long intervalle entre les vagues, consiste à faire des envois postaux aux répondants entre les vagues, par exemple à leur envoyer des cartes d'anniversaire ou des bulletins sur l'enquête, et à demander au service de la poste de fournir des avis de changement d'adresse, le cas échéant. Un autre moyen consiste à demander aux répondants d'indiquer les noms, les adresses et les numéros de téléphone de personnes qui leur sont proches (p. ex. des parents), qui sont peu susceptibles de déménager et qui pourront fournir leur adresse s'ils élisent domicile ailleurs.

  La deuxième façon de traiter les cas de répondants introuvables est de mettre en oeuvre diverses méthodes de dépistage pour tenter de les retrouver. Avec des efforts et de l'ingéniosité, on peut atteindre des taux de réussite élevés. Certaines méthodes de dépistage peuvent être propres à la population à l'étude (p. ex. sociétés professionnelles pour les personnes faisant partie de groupes professionnels), tandis que d'autres peuvent être plus générales, par exemple: annuaire téléphonique, recherche informatisée de numéros de téléphones, annuaires téléphoniques inversés permettant d'obtenir les numéros de téléphone de voisins, suivi du courrier, registres des actes de mariage, immatriculation des véhicules, employeurs et services d'information financière. Il peut se révéler utile d'examiner les registres des décès, notamment dans le cas des enquêtes par panel de longue durée. Les membres du panel qui sont décédés peuvent ainsi être classés de façon appropriée, plutôt que d'être considérés comme des non-répondants. Les méthodes de dépistage sont examinées par Burgess (1989), Clarridge et coll. (1978), Crider et coll. (1971) et Eckland (1968).

3.2 Problèmes des enquêtes par panel

Les enquêtes par panel sont confrontées, comme toutes les autres enquêtes, à une vaste gamme de sources d'erreur non due à l'échantillonnage. La présente section n'examine pas toutes ces sources, mais en présente trois qui sont propres aux enquêtes par panel: la non-réponse de vague, le biais de conditionnement et l'effet de lisière.

3.2.1 Non-réponse de vague

La non-réponse enregistrée à la première vague des enquêtes par panel correspond à celle qui s'observe dans les enquêtes transversales. Ce qui distingue les enquêtes par panel, c'est qu'elles font face à une non-réponse additionnelle au moment de vagues subséquentes. Certains membres du panel qui deviennent des non-répondants à une vague particulière ne répondent plus ensuite à aucune autre vague, tandis que d'autres répondent à nouveau à certaines ou à la totalité des vagues subséquentes. Les premiers sont souvent appelés des cas d'érosion, tandis que les seconds sont appelés cas de non-érosion. Les taux globaux de non-réponse de vague des enquêtes par panel s'accroissent au fur et à mesure des vagues, mais dans le cas des enquêtes bien gérées, le taux d'augmentation s'atténue nettement avec le temps. Par exemple, dans le cas du panel de l'enquête SIPP de 1987, la perte était de 6.7% à la vague 1, de 12.6% à la vague 2, puis elle a augmenté lentement jusqu'à 19.0% à la vague 7 (Jabine et coll. 1990). La tendance de la non-réponse à s'atténuer aux vagues ultérieures est rassurante, mais le cumul de la non-réponse sur de nombreuses vagues produit néanmoins des taux de non-réponse élevés aux dernières vagues d'un panel de longue durée. Par exemple, en 1988, après 21 rondes annuelles de collecte des données, le taux de non-réponse de l'enquête PSID pour les personnes qui vivaient dans les ménages de 1968 avait atteint 43.9% (Hill 1992).

Le choix entre les deux méthodes d'usage général qui permettent de traiter les données d'enquête manquantes - pondération et imputation - n'est pas immédiat dans le cas de la non-réponse de vague touchant les enquêtes par panel. Pour l'analyse longitudinale, la méthode de pondération consiste à laisser de côté tous les enregistrements ayant une ou plusieurs vagues manquantes et à essayer de compenser pour leur absence par des rajustements de pondération appliqués aux enregistrements restants. Cette méthode peut entraîner une importante perte d'information lorsque le fichier de données porte sur plusieurs vagues. Dans la méthode

d'imputation, en revanche, toutes les données déclarées sont conservées, mais on doit faire des imputations à vaste échelle à titre de compensation pour les vagues manquantes. Un compromis est aussi possible: on peut utiliser l'imputation pour certains profils de non-réponse de vague (p. ex. ceux qui ne comportent qu'une seule vague manquante, et pour lesquels on dispose des données des deux vagues adjacentes), et la pondération pour d'autres (voir, par exemple, Singh et coll. 1990). Pour l'analyse transversale, des fichiers de données distincts peuvent être créés à chaque vague. Ces fichiers peuvent comprendre tous les répondants à cette vague et utiliser soit des rajustements de pondération, soit des imputations, pour tenir compte des non-répondants à cette vague. Les méthodes de traitement de la non-réponse de vague sont examinées par Kalton (1986) et Lepkowski (1989).

3.2.2 Effet de conditionnement

Le biais dû au conditionnement, ou effet lié au temps passé dans l'échantillon, est l'effet qui s'exerce sur les réponses d'un membre du panel à une vague donnée de collecte des données en raison de sa participation aux vagues précédentes. L'effet peut refléter simplement un changement de comportement à l'égard de la déclaration. Par exemple, un répondant peut avoir remarqué dans une interview précédente que le fait de répondre «oui» à une question entraîne une série de sous-questions qui, en revanche, ne sont pas posées si la réponse est «non». Dans ce cas, le répondant peut répondre «non» pour éviter d'avoir à répondre à ces questions additionnelles. Du côté positif, un répondant peut avoir constaté à des interviews précédentes qu'il fallait fournir des renseignements détaillés sur le revenu et se préparer aux interviews subséquentes en amassant au préalable l'information nécessaire. L'effet de conditionnement peut aussi se traduire par un changement du comportement réel. Par exemple, un répondant peut s'inscrire au programme des coupons alimentaires après avoir appris l'existence de ce programme grâce aux questions posées à ce sujet à des vagues précédentes de collecte des données.

Une récente étude expérimentale sur l'effet de conditionnement dans une enquête par panel de quatre ans sur les nouveaux mariés a permis de déceler un lien entre la participation à l'enquête et l'harmonie du mariage (Veroff et coll. 1992). Toutefois, cette étude s'est fondée sur des techniques d'interview en profondeur qui vont plus loin que celles utilisées dans la plupart des enquêtes. Un certain nombre d'études sur le conditionnement des panels ayant été effectuées dans des contextes d'enquête plus courants ont révélé que des effets de conditionnement se produisent parfois, mais qu'ils ne sont pas généralisés (Traugott et Katosh 1979; Ferber 1964; Mooney 1962; Waterton et Lievesley 1989).

Les enquêtes par panel avec renouvellement et chevauchement ont comme avantage de permettre la comparaison entre des estimations relatives à la même période obtenues de panels différents. De telles comparaison ont permis de détecter clairement la présence de ce qu'on appelle un «biais lié au groupe de renouvellement» dans les enquêtes sur la population active des États-Unis et du Canada (p. ex. Bailar 1975, 1989 et U.S. Bureau of the Census 1978 pour l'enquête «Current Population Survey» américaine; Ghangurde 1982 pour l'enquête sur la population active canadienne). Le biais lié au groupe de renouvellement peut être le fruit d'un biais de non-réponse et d'effets de conditionnement. Dans des analyses comparant les panels chevauchants de 1985, 1986 et 1987 de l'enquête SIPP, Pennell et Lepkowski (1992) ont observé peu d'écarts entre les résultats des différents panels.

3.2.3 Effet de lisière

Dans de nombreuses enquêtes par panel, des données sont recueillies relativement à des sous-périodes de la période de référence, à partir de la dernière vague de collecte de données. Dans l'enquête SIPP, par exemple, des données sont recueillies sur une base mensuelle à l'intérieur de la période de référence de quatre mois qui sépare les vagues. L'effet de lisière correspond à ce qu'on observe souvent avec cette forme de collecte de données, c'est-à-dire que les niveaux de changements déclarés entre des sous-périodes adjacentes (p. ex. adhérer à un programme social, puis cesser d'y participer d'un mois à l'autre) sont beaucoup plus élevés lorsque les données des deux sous-périodes proviennent de vagues différentes que lorsqu'elles viennent de la même vague. On a constaté dans l'enquête SIPP un très vaste effet de lisière, ayant trait aussi bien à l'état de prestataire qu'aux montants des prestations (voir, par exemple, Jabine et coll. 1990; Kalton et Miller 1991). L'effet de lisière s'est également manifesté dans l'enquête PSID (Hill 1987). Murray et coll. (1991) décrivent des méthodes permettant de réduire l'effet de lisière dans l'enquête sur la population active du Canada.

3.3 Analyse longitudinale

Il existe une documentation abondante, qui ne cesse de s'accroître, dans le domaine de l'analyse des données longitudinales, notamment plusieurs manuels qui en exposent la matière (p. ex. Goldstein 1979; Hsiao 1986; Kessler et Greenberg 1981; Markus 1979). La présente description ne peut être complète, et doit se limiter à quelques thèmes généraux.

- *Mesure de la variation brute.* Comme nous l'avons déjà signalé, un avantage essentiel de l'enquête par panel par rapport à l'enquête à passages répétés vient de la possibilité de mesurer la variation brute, c'est-à-dire la variation enregistrée au niveau individuel. La méthode de base pour mesurer la variation brute est de dresser un tableau comparatif des réponses fournies à une vague par rapport à celles fournies à la même question, lors d'une autre vague. Cette sorte d'analyse souffre toutefois d'une contrainte importante; en effet, les variations de l'erreur de mesure d'une vague à l'autre peuvent engendrer un biais prononcé dans l'estimation de la variation brute (voir Kalton et coll. 1989, et Rodgers 1989, pour un examen plus détaillé).

- *Relation entre les variables dans le temps.* Les enquêtes par panel permettent de recueillir les données nécessaires à l'étude des liens qui existent entre des variables mesurées à différents moments. Par exemple, à partir de données recueillies sur la cohorte britannique des naissances de 1946 dans le cadre de la National Survey of Health and Development, Douglas (1975) a constaté que les enfants qui étaient hospitalisés pendant plus d'une semaine ou qui subissaient des hospitalisations répétées entre les âges de 6 mois et de 3 1/2 ans affichaient un comportement plus problématique à l'école et avaient des notes de lecture plus faibles à 15 ans. En principe, les enquêtes transversales peuvent utiliser des questions rétrospectives pour recueillir les données nécessaires à ce genre d'analyse. Cependant, les réponses à de telles questions sont souvent entachées par un niveau élevé d'erreur de mémoire et sont parfois empreintes de distorsions systématiques influant sur la relation étudiée.

- *Régression avec termes de différence.* Une régression dont les coefficients sont exprimés sous forme de différences peut permettre d'éviter un certain type d'erreur dans la spécification d'un modèle. Supposons que le modèle de régression approprié pour la personne i au temps t soit

$$Y_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \epsilon_{it}$$

où x_{it} est une variable explicative dont la valeur change avec le temps et z_{it} est une variable explicative qui est constante dans le temps (p. ex. sexe, race). Supposons en outre que z_{it} ne soit pas observée, par exemple qu'elle soit inconnue. Alors ß peut encore être estimé, mais d'après une régression avec termes de différence:

$$Y_{i(t+1)} - Y_{i(t)} = \beta (x_{i(t+1)} - x_{it}) + \epsilon_{i(t+1)} - \epsilon_{it}$$

(Rodgers 1989; Duncan et Kalton 1987).

- *Estimation des durées des périodes.* Les données recueillies dans les enquêtes par panel peuvent servir à estimer la distribution des durées d'événements comme le fait de participer à un programme de protection sociale. Dans des enquêtes par panel comme l'enquête SIPP, certaines personnes se trouvent déjà dans une période de participation au début du panel (périodes tronquées au départ), certaines personnes amorcent une période de participation durant le panel, et certaines périodes se poursuivent après la fin du panel (périodes tronquées à droite). Par conséquent, les périodes ne sont pas toutes observées au complet. La distribution des durées des périodes peut être estimée en appliquant des méthodes d'analyse de survie, par exemple la méthode d'estimation de Kaplan-Meier fondée sur une limite de produit, à toutes les nouvelles périodes (y compris celles qui sont tronquées à droite) qui commencent pendant la durée du panel (p. ex. Ruggles et Williams 1989).

- *Modèles à équations structurelles avec erreurs de mesure.* La séquence de collecte des données dans une enquête par panel se traduit par une mise en ordre claire des variables de l'enquête, dont l'analyse se prête bien à l'utilisation de modèles à équations structurelles. Cette forme d'analyse peut prendre en considération

les erreurs de mesure et, grâce à plusieurs mesures répétées, peut traiter des structures d'erreur corrélées (p. ex. Jöreskog et Sörbom 1979).

4. CONCLUSION

Les ensembles de données construits à partir d'enquêtes par panels offrent habituellement une grande richesse de contenu pour l'analyse. Ils contiennent des mesures répétées pour certaines variables observées plusieurs fois, et des mesures d'autres variables qui sont examinées à une seule vague. Les interviews répétées du même échantillon donnent l'occasion de recueillir des données sur de nouvelles variables à chaque vague, ce qui produit des données sur toute une gamme de variables observées sur plusieurs vagues. Les données d'un panel peuvent être soumises aussi bien à une analyse longitudinale qu'à une analyse transversale. Les mesures répétées peuvent permettre d'examiner les profils de réponse individuels dans le temps et peuvent aussi être reliées à d'autres variables. Les variables mesurées à une vague unique peuvent être analysées à la lumière d'autres variables observées à cette vague, ainsi que de variables mesurées à d'autres vagues.

La richesse des données d'un panel n'est mise à contribution que si les données sont effectivement analysées, et si elles le sont dans les délais les plus brefs possibles. Mettre en oeuvre une enquête par panel, c'est comme faire tourner une immense roue: les étapes de la conception du questionnaire, de la collecte des données, du traitement et de l'analyse doivent être reprises à chaque vague successive. Il y a un réel danger que l'équipe de l'enquête soit submergée par ce processus, et que les données ne soient pas pleinement analysées. Pour éviter un tel écueil, il faut mettre en place le personnel adéquat et construire une organisation bien intégrée.

Il est souhaitable, par ailleurs, de préserver la simplicité du plan de l'enquête par panel. Le plan devrait être élaboré à la lumière d'objectifs clairement définis. Un bon sens critique doit être exercé au moment d'ajouter au plan des éléments de complexité visant à accroître la richesse des données du panel pour en élargir les applications. Bien que des arguments persuasifs puissent souvent être invoqués en faveur de tels ajouts, ceux-ci doivent être rejetés s'ils menacent l'exécution ordonnée de n'importe quelle étape du processus d'enquête.

Comme il a été indiqué plus haut, les erreurs de mesure ont des effets particulièrement néfastes sur l'analyse des variations individuelles révélées par les données d'une enquête par panel. L'affectation d'une partie des ressources de l'enquête à la mesure de l'ampleur de ces erreurs est donc une décision judicieuse (Fuller 1989). Les erreurs de mesure peuvent être étudiées soit par des études de validité (c.-à-d. par une comparaison des réponses de l'enquête avec des valeurs «vraies» tirées d'une source externe), soit par des études de fiabilité (p. ex. études fondées sur des réinterviews). Les résultats de telles études peuvent ensuite être utilisés dans le processus d'estimation et permettre d'effectuer des rajustements tenant compte des erreurs de mesure.

BIBLIOGRAPHIE

Bailar, B.A. (1975). The effects of rotation group bias on estimates from Panel Surveys. *Journal of the American Statistical Association*, 70, 23-30.

Bailar, B.A. (1989). Information needs, surveys, and measurement errors. *Panel Surveys*, Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh, New York: John Wiley, 1-24.

Binder, D.A., et Hidiroglou, M.A. (1988). Sampling in time. *Handbook of Statistics, Vol. 6*, Éds. P.R. Krishnaiah et C.R. Rao, New York: North Holland, 187-211.

Boruch, R.F., et Pearson, R.W. (1988). Assessing the Quality of Longitudinal Surveys, *Evaluation Review*, 12, 3-58.

Burgess, R.D. (1989). Major Issues and Implications of Tracing Survey Respondents. *Panel Surveys*, Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh, New York: John Wiley, 52-74.

Citro, C.F., et Kalton, G. (1989). *Surveying the Nation's Scientists and Engineers*, Washington DC: National Academy Press.

Citro, C.F., et Kalton, G. (1993). *The Future of the Survey of Income and Program Participation*, Washington DC: National Academy Press.

Clarridge, B.R., Sheehy, L.L., et Hauser, T.S. (1978). Tracing Members of a Panel: a 17-year Follow-up. *Sociological Methodology*, Éd. K.F. Schuessler, San Francisco: Jossey-Bass, 389-437.

Crider, D.M., Willits, F.K., et Bealer, R.C. (1971). Tracking Respondents in Longitudinal Surveys. *Public Opinion Quarterly*, 35, 613-620.

Douglas, J.W.B. (1975). Early Hospital Admissions and Later Disturbances of Behaviour and Learning. *Developmental Medicine and Child Neurology*, 17, 456-480.

Duncan, G.J., et Kalton, G. (1987). Issues of Design and Analysis of Surveys Across Time. *International Statistical Review*, 55, 97-117.

Eckland, B.K. (1968). Retrieving Mobile Cases in Longitudinal Surveys. *Public Opinion Quarterly*, 32, 51-64.

Subcommittee on Federal Longitudinal Surveys (1986). *Federal Longitudinal Surveys*, Statistical Policy Working Paper 13, Washington DC: Office of Management and Budget.

Ferber, R. (1964). Does a Panel Operation Increase the Reliability of Survey Data: the Case of Consumer Savings. *Proceedings of the Social Statistics Section, American Statistical Association*, 210-216.

Fuller, W.A. (1989). Estimation of Cross-Sectional et Change Parameters: Discussion, *Panel Surveys*, Eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh, New York: John Wiley, 480-485.

Ghangurde, P.D. (1982). Rotation Group Bias in the LFS Estimates. *Survey Methodology*, 8, 86-101.

Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies*, New York: Academic Press.

Hill, D. (1987). Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 210-215.

Hill, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*, Newbury Park, CA: Sage Publications.

Hsiao, C. (1986). *Analysis of Panel Data*, New York: Cambridge University Press.

Jabine, T.B., King, K.E., et Petroni, R.J. (1990). *Survey of Income and Program Participation: Quality Profile*, Bureau of the Census, Washington DC: U.S. Department of Commerce.

Jöreskog, K.G., et Sörbom, D. (1979). *Advances in Factor Analysis and Structural Equation Models*, Lanham MD: University Press of America.

Kalton, G. (1986). Handling Wave Nonresponse in Panel Surveys. *Journal of Official Statistics*, 2, 303-314.

Kalton, G., Kasprzyk, D., et McMillen, D.B. (1989). Nonsampling Errors in Panel Surveys. *Panel Surveys*, Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh, New York: John Wiley, 249-270.

Kalton G., et Lepkowski, J.M. (1985). Following Rules in SIPP. *Journal of Economic and Social Measurement*, 13, 319-329.

Kalton, G., et Miller, M.E. (1991). The Seam Effect with Social Security Income in the Survey of Income and Program Participation. *Journal of Official Statistics*, 7, 235-245.

Kasprzyk, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues*. Document de travail du SIPP n° 8830, Washington DC: U.S. Bureau of the Census.

Kasprzyk, D., Duncan G., Kalton, G., et Singh, M.P. (Éds.) (1989). *Panel Surveys*, New York: John Wiley.

Kessler, R.C., et Greenberg, D.F. (1981). *Linear Panel Analysis*, New York: Academic Press.

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

Lazarsfeld, P.F. (1948). The Use of Panels in Social Research. *Proceedings of the American Philosophical Society*, 42, 405-410.

Lazarsfeld, P.F., et Fiske, M. (1938). The Panel as a New Tool for Measuring Opinion. *Public Opinion Quarterly*, 2, 596-612.

Lepkowski, J.M. (1989). Treatment of Wave Nonresponse in Panel Surveys. *Panel Surveys*, Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh, New York: John Wiley, 348-374.

Magnusson, D., et Bergman, L.R. (Éds.) (1990). *Data Quality in Longitudinal Research*, New York: Cambridge University Press.

Markus, G.B. (1979). *Analyzing Panel Data*, Beverly Hills, CA: Sage Publications.

Mooney, H.W. (1962). *Methodology in Two California Health Surveys*, Public Health Monograph No. 70, Washington DC: U.S. Department of Health, Education, and Welfare.

Murray, T.S., Michaud, S., Egan, M., et Lemaitre, G. (1991). Invisible Seams? The Experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Bureau of the Census Annual Research Conference*, Washington DC: U.S. Department of Commerce, 715-730.

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

Pennell, S.G., et Lepkowski, J.M. (1992). Panel Conditioning Effects in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, à venir.

Rodgers, W.L. (1989). Comparisons of Alternative Approaches to the Estimation of Simple Causal Models from Panel Data. *Panel Surveys*, Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh, New York: John Wiley, 432-456.

Singh, R., Huggins, V., et Kasprzyk, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys*. Document de travail du SIPP n° 9009, Bureau of the Census, Washington DC: U.S. Department of Commerce.

Traugott, M., et Katosh, K. (1979). Response Validity in Surveys of Voting Behavior. *Public Opinion Quarterly*, 79, 359-377.

U.S. Bureau of the Census (1978). *The Current Population Survey: Design and Methodology*, Bureau of the Census Technical Paper n°. 40, Washington DC: U.S. Government Printing Office.

Van de Pol, F.J.R. (1989). *Issues of Design and Analysis of Panels*, Amsterdam: Sociometric Research Foundation.

Veroff, J., Hatchett, S., et Douvan, E. (1992). Consequences of Participating in a Longitudinal Study of Marriage. *Public Opinion Quarterly*, 56, 315-327.

Wall, W.D., et Williams, H.L. (1970). *Longitudinal Studies and the Social Sciences*, London: Heinemann.

Waterton, J., et Lievesley, D. (1989). Evidence of Conditioning Effects in the British Social Attitudes Panel. *Panel Surveys*, Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh, New York: John Wiley, 319-339.

SESSION 1

Conception de questionnaires et questions liées à la collecte

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

UTILISATION DE L'IPAO AU COURS D'UNE ENQUÊTE LONGITUDINALE: UN RAPPORT SUR L'ENQUÊTE «MEDICARE CURRENT BENEFICIARY SURVEY»

W.S. Edwards, S. Sperry et B. Edwards[1]

RÉSUMÉ

L'enquête «Medicare Current Beneficiary Survey (MCBS)» est une enquête permanente par panel portant sur les bénéficiaires du programme «Medicare» aux États-Unis. Au rythme de trois interviews par année, on recueille des données sur l'utilisation et le coût des services de santé. Toutes les interviews à domicile sont effectuées en personne avec l'aide d'un système d'interviews sur place assistées par ordinateur (IPAO). Outre les éléments habituels d'une interview assistée par ordinateur, le système IPAO de la MCBS utilise largement des listes, d'où les interviewers extraient de l'information, auxquelles ils ajoutent des éléments ou dont ils se servent pour corriger les réponses de l'interview en cours ou d'une interview précédente. Le système IPAO de la MCBS utilise également l'information antérieure à diverses autres fins: (1) présenter ces renseignements pour aider à délimiter la période de référence ou rappeler des détails à la mémoire du répondant, (2) offrir ces renseignements comme données de base et s'enquérir des changements survenus entre-temps et (3) demander des précisions sur des détails jusque-là non disponibles. Bien que cette méthode n'ait pas échappé à certains problèmes, elle s'est révélée fructueuse pour la collecte de données longitudinales relatives à un comportement complexe.

MOTS-CLÉS: IPAO; interview avec rétro-information; enquête sur les soins de santé.

1. INTRODUCTION

1.1 Questions méthodologiques propres aux enquêtes longitudinales

Les enquêtes longitudinales, ou enquêtes «par panel», se caractérisent par toute une gamme de problèmes méthodologiques liés au fait que plusieurs interviews sont réalisées avec les mêmes répondants. Parce qu'il n'y a pas toujours uniformité des réponses d'une étape ou d'un volet à l'autre d'une enquête longitudinale, la correction d'erreurs et l'analyse doivent surmonter des difficultés comme celle de déterminer laquelle de deux valeurs d'une variable statique obtenues dans deux interviews différentes est exacte, ou de savoir si une modification apparente d'une variable dans le temps représente un réel changement ou seulement une façon différente de dire la même chose. Les enquêtes longitudinales portant sur le comportement obligent aussi à déterminer si un événement particulier s'est produit au cours de la période de référence courante de l'enquête, ou s'il s'agit d'un événement déjà signalé au cours d'une interview précédente.

Neter et Waksberg (1964), dans une recherche méthodologique relative à la «Consumer Expenditure Survey» des É.-U., ont constaté que la «délimitation» d'une période de référence avec des renseignements déjà déclarés réduisait l'erreur due au «télescopage», c.-à-d. l'attribution à la période de référence d'événements qui n'en font pas partie. D'autres enquêtes longitudinales ont utilisé les renseignements déjà déclarés pour réduire l'«effet de lisière», ou la tendance des enquêtes longitudinales à surestimer les changements survenant à la «lisière», ou limite, des périodes de référence entre les interviews du panel. (L'«effet de lisière» est décrit, par exemple, dans Moore et Kasprzyk 1984). Lorsque des données recueillies précédemment sont utilisées dans une interview, on parle souvent d'«interview avec rétro-information».

---

[1] W.S. Edwards, S. Sperry et B. Edwards, Westat, Inc., Rockville (MD), É.-U.

L'usage d'information recueillie au cours d'une interview précédente a généralement été entravé, jusqu'ici, par les obstacles techniques de conception du questionnaire et la difficulté pour les interviewers de faire une interview avec rétro-information dans le cadre d'un questionnaire structuré. Les interviews avec rétro-information exigent habituellement au moins deux documents distincts: une liste, ou un guide, énonçant les questions, et un relevé de données recueillies précédemment. Les références de l'un à l'autre de ces documents sont souvent peu commodes et exigent une certaine créativité de la part de l'interviewer pour que des phrases ou des questions compréhensibles puissent être formulées.

1.2 Avantages possibles des interviews assistées par ordinateur dans les enquêtes longitudinales

Comme le signale Saris (1991), l'interview assistée par ordinateur (IAO) offre la possibilité de faire un usage beaucoup plus grand de renseignements recueillis précédemment que ne l'ont permis jusqu'ici les interviews menées avec crayon et papier. Saris énonce les avantages suivants de l'IAO:

- des vérifications d'étendue et de cohérence peuvent être incorporées à une interview (la cohérence pouvant être vérifiée aussi bien à l'intérieur de l'interview en cours que d'une interview à l'autre);

- une correction d'erreur peut être effectuée, grâce à l'affichage d'information cumulative sur des écrans «de sommaire et de correction», comme les désigne Saris, comprenant des données aussi bien de l'interview en cours que d'interviews précédentes;

- l'IAO permet d'inclure des questions au sujet d'événements périodiques d'après l'information relative aux intervalles prévus;

- l'IAO permet, dans des enquêtes portant sur la vie entière de la personne, d'utiliser des données antérieures pour rappeler des détails à la mémoire du répondant;

- l'IAO peut servir à rappeler aux répondants des réponses précédentes, afin de délimiter la période de référence.

1.3 But de la présente communication

De nombreux avantages de l'interview assistée par ordinateur dans les enquêtes longitudinales ont été mis à contribution dans la conception de la «Medicare Current Beneficiary Survey» (MCBS), une enquête par panel menée auprès de bénéficiaires de «Medicare» aux États-Unis, dans laquelle on utilise les interviews sur place assistées par ordinateur (IPAO) pour recueillir la majorité des données. La présente communication vise à décrire certains aspects du plan de la MCBS qui mettent en évidence l'utilisation de renseignements recueillis antérieurement.

2. VUE D'ENSEMBLE DE LA «MEDICARE CURRENT BENEFICIARY SURVEY»

La MCBS est une enquête par panel permanente, menée auprès de bénéficiaires de «Medicare» par Westat, Inc., pour le compte de la HCFA (Health Care Financing Administration), l'organisme du gouvernement fédéral responsable de l'administration de Medicare. Cette enquête permet de recueillir des renseignements sur l'utilisation et le coût des soins de santé, sur les régimes d'assurance-maladie et sur d'autres questions liées à la santé. L'échantillon comprend en tout temps environ 12 000 personnes, sélectionnées à partir des dossiers de Medicare, dont environ 11 000 vivent à domicile et 1 000 dans des centres d'hébergement ou d'autres établissements. Toutes les interviews à domicile sont effectuées sur place par des interviewers utilisant le système IPAO de Westat.

2.1 Le programme Medicare

Medicare est un programme d'assurance-maladie fédéral couvrant la plupart des personnes de 65 ans et plus et certaines personnes handicapées aux États-Unis. Le programme comprend à la fois une assurance-hospitalisation

et une assurance médicale. Il couvre une grande partie, mais pas la totalité, des frais de soins de santé des bénéficiaires. Ceux-ci doivent payer des franchises et une co-assurance pour de nombreux services couverts, et doivent assumer le coût entier des services non couverts. La plupart de médicaments d'ordonnance, les soins dentaires et les soins à long terme en centre d'hébergement ne sont pas couverts par le programme Medicare. Les bénéficiaires peuvent avoir d'autres assurances publiques ou privées couvrant une partie ou la totalité des frais non couverts par Medicare.

2.2 Objectif de la MCBS

Les volumineux dossiers de réclamations tenus à jour dans le cadre du programme Medicare peuvent servir à analyser l'ampleur de l'utilisation des services par les bénéficiaires et l'importance des montants versés en prestations. Les dossiers de Medicare, toutefois, contiennent des données limitées sur les caractéristiques démographiques et l'état de santé, comportent très peu d'information sur la façon dont les bénéficiaires paient les franchises, la co-assurance et d'autres frais relatifs aux services couverts, et ne renferment aucune donnée sur l'utilisation et le coût des services non couverts par Medicare. La MCBS vise à fournir aux analystes cette information manquante afin de permettre une meilleure gestion du programme, à évaluer les effets sur les bénéficiaires de modifications du programme et à modéliser les effets de changements proposés.

2.3 Complexité du financement des soins de santé aux États-Unis

Aux États-Unis, l'obtention de soins médicaux et le paiement des frais connexes peuvent être des activités complexes. Les gens peuvent se rendre chez plusieurs médecins ou autres dispensateurs de soins différents, dont les services peuvent être organisés de façons très diverses. Un événement médical unique, notamment lors d'une hospitalisation, peut faire intervenir plusieurs services médicaux différents, plusieurs dispensateurs et plusieurs factures distinctes. À l'inverse, de nombreux services peuvent être reçus d'un même dispensateur et inclus sur une seule facture. Comme nous l'avons vu au paragraphe précédent, il se peut qu'en fin de compte le paiement d'un compte médical particulier soit partagé par plusieurs sources différentes. Toute cette complexité rend difficile la tâche de concevoir une enquête permettant de décrire l'utilisation, le coût et les modes de paiement des services de santé.

2.4 L'interview de la MCBS

La MCBS comprend trois interviews de chaque répondant par année. Chaque interview couvre la période écoulée entre l'interview actuelle et l'interview précédente. L'interview de base, effectuée à chaque reprise, comprend des questions sur la composition du ménage, sur les régimes d'assurance-maladie, sur l'utilisation d'une vaste gamme de services médicaux, et sur les frais ainsi que les sources et les montants payés pour ces services. Chaque interview comprend aussi des questions supplémentaires sur toute une variété de sujets.

3. ASPECTS TRANSVERSAUX DU SYSTÈME IPAO DE LA MCBS

La présente section décrit les caractéristiques du système IPAO de la MCBS qui ne sont pas directement liées à la nature longitudinale de l'enquête, mais qui sont des composantes essentielles des fonctions propres aux interviews menées auprès du panel.

3.1 Caractéristiques de base du système IPAO

Le questionnaire de la MCBS comporte les caractéristiques habituelles de l'interview assistée par ordinateur (IAO): branchements automatiques d'après les réponses aux questions précédentes; choix automatiques de mots comme l'insertion du nom de la personne, l'usage de pronoms adaptés à la personne, l'insertion de noms de dispensateurs de soins, etc.; vérifications d'étendue pour des données numériques ou d'autres valeurs afin de réduire les erreurs de saisie au clavier.

En outre, le système IPAO de la MCBS repose largement sur l'emploi de «listes» relatives à des catégories particulières traitées au cours des interviews. Il comprend notamment des listes de personnes (p. ex. personne interviewée, membres du ménage, personnes à contacter), de dispensateurs de soins (p. ex. médecins, hôpitaux,

agences de soins à domicile), de régimes d'assurance-maladie (p. ex. Medicare, Medicaid, Blue Cross/Blue Shield), de dates de visite à chacun des dispensateurs de soins, et de sources de paiement pour les soins médicaux (p. ex. régimes d'assurance, famille, bureau où travaille la personne). Les listes sont tenues à jour par l'interviewer au cours de l'interview. Elles sont affichées automatiquement au moment approprié d'une interview, ou encore elles peuvent être appelées en tout temps par l'interviewer pour consultation. En général, les interviewers peuvent sélectionner un élément figurant déjà dans la liste, ajouter des éléments à la liste ou encore y apporter des corrections.

3.2 Détails sur les listes de la MCBS

La figure 1 présente un écran du système IPAO de la MCBS -- une demande de précisions sur les régimes d'assurance-maladie privés. L'écran comprend des numéros d'identification au sommet, une question devant être lue par l'interviewer et des catégories de réponse à la partie inférieure. L'interviewer entre le numéro de la réponse appropriée (le curseur se trouve à l'intérieur des parenthèses situées au-dessus des catégories de réponse) et appuie sur la touche «Enter» pour enregistrer la réponse.

Figure 1: Écran du système IPAO de la MCBS -- Demande de précisions sur l'assurance-maladie.

3.17 H117B 10000027 9112181803 911219

J'ai une question au sujet des autres types d'assurance-maladie.

En tout temps depuis le 20 août 1991, avez-vous été protégé par un régime d'assurance-maladie privé, c'est-à-dire un régime couvrant les frais d'hospitalisation ou les services de médecins, ou encore le coût des médicaments d'ordonnance?

( )

1. OUI
2. NON

Si la réponse à la question de la figure 1 est «oui», le programme IPAO affiche la liste des régimes d'assurance-maladie présentée à la figure 2.

Figure 2: Écran IPAO de la MCBS -- Liste des régimes d'assurance-maladie.

3.200 H120 10000027 9112181803 911219

Quel est le nom de chacun des autres régimes qui vous offrent une assurance médicale?
[ENTRER TOUS LES RÉGIMES PRIVÉS]

POUR EFFACER UN X, APPUYER SUR LA BARRE D'ESPACEMENT.
POUR AJOUTER UN PLAN, APPUYER SUR CTRL/A.
POUR QUITTER L'ÉCRAN, APPUYER SUR ESC.

--- NOM DU RÉGIME----------	ÉTAT----------
MEDICARE	COURANT
MEDICAL ASSISTANCE	COURANT
MARYLAND PHARMACY ASST.	ARRÊTÉ 11/30/91
THE RAINBOW COVERAGE PLAN	AJOUTÉ

À la figure 2, on peut voir la liste des régimes d'assurance-maladie à l'intérieur du rectangle. Le texte de la question est affiché au-dessus du rectangle en minuscules, tandis que les instructions pour l'interviewer sont

affichées en majuscules. L'interviewer peut sélectionner un régime déjà entré en plaçant un «X» à gauche du nom du régime, ajouter un régime ou corriger le nom d'un régime existant. Dans l'exemple de la figure, l'interviewer ajoute le régime «The Rainbow Coverage Plan». Apparemment, le répondant a indiqué auparavant que la protection en vertu du régime «Maryland Pharmacy Assistance» avait pris fin. Après cet écran, le programme présentera des questions visant à obtenir des détails sur le «Rainbow Plan».

Les figures 3 à 5 présentent d'autres exemples de l'utilisation de listes pour la MCBS. La figure 3 illustre une demande de précisions au sujet des consultations de médecins au cours de la période de référence (du 20 août 1991 à la date de l'interview).

Figure 3: Écran IPAO de la MCBS -- Demande de précisions sur les consultations de médecins.

10.991 MP1 10000027 9112181803 911219

Depuis le 20 août 1991,
avez-vous fait des visites
chez des médecins?
[INCLURE TOUTE VISITE POUR DES TESTS OU RADIOGRAPHIES]

( )

1. OUI
2. NON

Si la réponse à la question de la figure 3 est «oui», le programme IPAO affiche la liste des dispensateurs de soins, comme à la figure 4. L'interviewer sélectionne un dispensateur figurant déjà dans la liste ou en introduit un nouveau, comme pour les régimes d'assurance-maladie à la figure 2. Une fois qu'un dispensateur de soins médicaux a été sélectionné, le programme IPAO présente la liste des visites faites à ce dispensateur, comme à la figure 5.

Figure 4: Écran IPAO de la MCBS -- Liste des dispensateurs de soins.

10.992 MP2 10000027 9112181803 911219

Qui avez-vous vu?
[INTRODUIRE UN SEUL DISPENSATEUR.]

POUR SÉLECTIONNER UN DISPENSATEUR, UTILISER LES TOUCHES FLÉCHÉES, APPUYER SUR X, APPUYER SUR ENTER.

POUR EFFACER UN X, APPUYER SUR LA BARRE D'ESPACEMENT.
POUR AJOUTER UN DISPENSATEUR, APPUYER SUR CTRL/A.
POUR QUITTER L'ÉCRAN, APPUYER SUR ESC.

|--- NOM DU DISPENSATEUR ---------- |
| SARA DALE |
| BOB DALE |
| ROCKVILLE CENTER |
DR JOE MARTIN

À la figure 5, l'interviewer a ajouté six dates auxquelles le Dr Joe Martin, le dispensateur sélectionné dans la liste précédente, a été visité au cours de la période de référence. Le programme IPAO affichera ensuite une série de questions sur chacune des visites. Les rectangles contenant les listes aux figures 2, 4 et 5 permettent un défilement vers le haut et vers le bas, afin de permettre l'entrée et la consultation d'un grand nombre d'éléments.

Figure 5: Écran IPAO de la MCBS -- Liste des visites.

10.997 MP6 10000027 9112181803 911219

Quand avez-vous vu
le DR JOE MARTIN?
Veuillez me donner toutes les dates
depuis le 20 août 1991. [ENTRER TOUTES LES DATES]

POUR EFFACER UN X, APPUYER SUR LA BARRE D'ESPACEMENT.
POUR AJOUTER UNE DATE, APPUYER SUR CTRL/A.
POUR QUITTER L'ÉCRAN, APPUYER SUR ESC.

---MM	--JJ--	AA --
X	10	1	91
X	10	5	91
X	10	10	91
X	10	15	91
X	10	20	91
X	10	25	91

3.3 Résumé des avantages de l'IPAO pour les aspects transversaux de la MCBS

La présente section a présenté quelques exemples de la façon dont l'IPAO permet à une structure de questionnaire complexe de recueillir des détails sur le comportement complexe consistant à obtenir des services médicaux et à en acquitter les frais. Comme dans la plupart des systèmes d'IAO, le programme IPAO de la MCBS comprend des options complexes de choix de mots à l'intérieur d'une question, permettant notamment l'insertion du nom de la personne interviewée, du nom des dispensateurs de soins et d'autres éléments tirés des listes, ainsi que l'utilisation d'une période de référence distincte pour chaque interview.

Au coeur du programme IPAO de la MCBS se trouve le système de listes, qui peuvent être mises à jour à jour pendant l'interview et dans lesquelles l'interviewer peut sélectionner des éléments qui réapparaîtront plus tard. Le programme IPAO de la MCBS permet de suivre des liens complexes entre les éléments d'une liste et ceux d'une autre liste. Dans les exemples montrés ci-dessus, un régime d'assurance-maladie se trouvant dans la liste de la figure 2 pourrait être une source de paiement pour une visite mentionnée dans la liste de la figure 5, faite à un dispensateur de soins faisant partie de la liste de la figure 4.

Enfin, un important avantage du système IPAO est la possibilité d'employer des schémas de branchement interdépendants complexes, grâce auxquels plusieurs éléments de données différents peuvent être examinés avant qu'on passe au point suivant de l'enquête. Par exemple, l'emploi des listes déclenche généralement un détour vers une série de questions. Dans l'exemple de la figure 5, le programme présenterait une série de questions sur chacune des visites inscrites dans la liste, demanderait ensuite s'il n'y aurait pas eu d'autres visites au même dispensateur de soins, puis demanderait s'il n'y aurait pas d'autres dispensateurs de soins à introduire. Si d'autres visites ou dispensateurs étaient mentionnés, les séries de questions recommenceraient.

4. CARACTÉRISTIQUES LONGITUDINALES DU SYSTÈME IPAO DE LA MCBS

4.1 Vue d'ensemble

Comme il a été mentionné plus haut, la MCBS est une enquête longitudinale. De nombreuses caractéristiques décrites à la section précédente ont des applications longitudinales. Ainsi, les listes sont tenues à jour aussi bien d'une interview à l'autre qu'à l'intérieur de la même interview, de façon que, par exemple, un dispensateur de soins mentionné à une interview apparaisse dans la liste des dispensateurs au moment de l'interview suivante. Certaines listes contiennent une colonne additionnelle dans laquelle est indiqué à quelle ronde d'interviews l'élément a été déclaré. Au nombre des autres éléments couramment reportés d'une interview à l'autre figurent

la composition du ménage, le nom du répondant et, s'il s'agit d'un remplaçant, son lien avec le membre de l'échantillon, ses coordonnées, etc.

D'une certaine manière, ces applications constituent des éléments d'une «interview avec rétro-information», c.-à-d. l'utilisation dans une interview de renseignements recueillis dans des interviews antérieures. Toutefois, l'interview de la MCBS comporte d'autres caractéristiques d'une «interview avec rétro-information», qui mettent à contribution de façon encore plus évidente les données obtenues précédemment. Ces caractéristiques sont décrites dans les sections qui suivent.

4.2 Interview avec rétro-information «passive»

Les données d'interviews précédentes peuvent être utilisées dans la MCBS d'une première façon, que nous avons appelée interview avec rétro-information «passive». Il s'agit d'une utilisation passive parce que l'information est soumise au répondant, mais que le questionnaire ne contient aucune question particulière au sujet des données, et que les données ne sont pas utilisées explicitement dans des questions visant à obtenir de l'information additionnelle. La présentation a la forme d'un calendrier annoté, avec des symboles représentant les visites chez le médecin, les séjours à l'hôpital et les achats de fournitures ou d'appareils médicaux et d'articles ou de services connexes.

En présentant les données ainsi, on vise à (1) réduire le «télescopage», c.-à-d. la déclaration d'événements médicaux survenus en dehors de la période de référence courante et (2) rappeler des détails à la mémoire du répondant -- il se peut que le membre de l'échantillon ait vu le même dispensateur de soins ou acheté le même médicament pendant la période courante que pendant la période précédente. Le répondant est invité à examiner attentivement le sommaire et à le consulter au cours de l'interview. Le questionnaire IPAO ne demande pas d'ajouter des données au sommaire ou d'y apporter des corrections, mais si le répondant signale des ajouts ou des modifications, l'interviewer peut les enregistrer. Sur plus de 110 000 événements médicaux déclarés à la première ronde de l'interview de base, tout juste plus de 1 000, ou environ un pour cent, ont été supprimés au cours de l'interview suivante. Par ailleurs, 768 événements médicaux additionnels, représentant 0.7 % du total, ont été ajoutés durant l'interview suivante.

4.3 Changements par rapport à des «données de base»

L'une des deux techniques d'interview avec rétro-information «active» utilisées dans la MCBS consiste à présenter des renseignements recueillis antérieurement sous forme de «données de base» pendant l'interview, et à demander s'il y a eu des changements. Cette méthode est appliquée à l'information à caractère permanent, comme la composition du ménage, les régimes d'assurance-maladie et les soins de santé à domicile.

Figure 6: Écran IPAO de la MCBS -- Vérification des données de base sur les régimes d'assurance-maladie.

3.01 H1S1 10000027 9112181803 911219

Vous étiez bénéficiaire de Medicare,
et vous étiez aussi couvert par (LIRE LES NOMS DES PLANS CI-DESSOUS)
le 19 février 1992.
Est-ce exact?

MEDICARE CONNPACE
BLUE CROSS BLUE SHIELD

( )

1. OUI, TOUT EST EXACT TEL QU'INDIQUÉ
2. NON, IL MANQUE UN RÉGIME
3. NON, UN NOM DE RÉGIME EST INEXACT
4. NON, UN RÉGIME DOIT ÊTRE SUPPRIMÉ

Dans le cas des régimes d'assurance-maladie, l'interviewer présente le sujet et remet au répondant un sommaire imprimé des régimes en vigueur au moment de l'interview précédente. L'interviewer vérifie ensuite la situation courante, à l'aide de l'écran présenté à la figure 6.

Contrairement à la présentation «passive» de l'information, l'interviewer demande ici explicitement si des corrections doivent être apportées à l'information antérieure. La raison d'une telle demande est d'empêcher que des corrections de l'information déjà recueillies soient confondues avec des changements. Des 11 804 régimes d'assurance-maladie autres que Medicare déclarés au cours de la première interview de la MCBS, 308 (2.6 %) ont été supprimés au cours de l'interview suivante. En outre, 298 régimes (2.5 %) ont été ajoutés dans le cadre de cette vérification. Dans d'autres conditions, ces corrections auraient pu être considérées comme des changements survenus entre les deux interviews.

La suite de l'interview consiste à demander, pour chacun des régimes (sauf Medicare) déclaré comme étant en vigueur au moment de l'interview précédente, si ce régime est encore en vigueur. Sur environ 10,950 plans en vigueur à la fin de la première ronde d'interviews, 358 (3.3 %) avaient pris fin au cours de la période de référence de la deuxième ronde. Par ailleurs, 849 régimes (7.8 % de ceux en vigueur à la première ronde) ont été ajoutés au moment de la deuxième ronde. Ces modifications devraient représenter des changements réels, contrairement aux corrections apportées à la réponse précédente.

4.4 Recherche de renseignements nouveaux

Bien que la plupart des événements médicaux aient une durée d'un jour ou moins (les séjours en centre d'hébergement et à l'hôpital constituant des exceptions notables), le processus de paiement relatif à un événement médical exige souvent plusieurs mois. Par conséquent, pour de nombreux événements médicaux déclarés dans l'interview de la MCBS pour la période de référence d'environ quatre mois, le répondant ne peut fournir tous les détails concernant la source et le montant du paiement. La structure de l'interview tient compte de ce délai. Par exemple, si le membre de l'échantillon n'a pas reçu de relevé de Medicare (habituellement la première source de paiement), mais en attend un, l'interviewer ne pose aucune autre question sur les frais ou le paiement. On peut ainsi gagner du temps et sans doute éviter d'ennuyer le répondant avec des questions dont il n'a pas la réponse. Dans d'autres cas, une source peut avoir fait le paiement, mais une autre non, ou encore le membre de l'échantillon peut avoir payé lui-même et attendre un remboursement de Medicare ou d'une assurance privée.

Dans de tels cas, le questionnaire de la MCBS reporte l'information d'une interview à l'autre, et reprend là où il y avait une information manquante la fois précédente. Dans ce genre de situations, la question de base est «Le savez-vous maintenant?» Si, par exemple, un relevé de Medicare relatif à un événement particulier était attendu au moment de la première interview, un indicateur est placé dans la base de données de manière à ce qu'on s'arrête de nouveau à cet événement dans la section «Sommaire des frais/paiements» de l'interview suivante. Toutefois, avant qu'on en arrive à cette section, l'interviewer demande au répondant si des relevés de Medicare ou d'une autre assurance ont été reçus depuis l'interview précédente, et introduit les données sur les frais et les paiements connexes. Il se peut qu'un événement portant un indicateur soit couvert de cette façon, auquel cas l'indicateur est enlevé avant qu'on passe à la section du sommaire. Pour les événements qui restent, un écran d'interrogation apparaît dans la section «Sommaire des frais/paiements»; dans l'exemple présenté à la figure 7, il s'agit d'un «événement» constitué de trois achats du médicament Percodan.

Si un relevé a été reçu, le programme demande ensuite l'information au sujet des frais et des paiements; des écrans différents peuvent apparaître selon que le relevé est disponible ou non. Si le répondant n'a pas reçu le relevé, la question suivante serait «Prévoyez-vous encore recevoir un relevé...?» Un événement pourrait être encore reporté à la ronde suivante si un relevé était toujours attendu.

Des 19 031 «relevés attendus» au moment de la première interview de base, 26 % ont été couverts par la question précédant le Sommaire des frais/paiements. Une proportion additionnelle de 11 % ont été récupérés (relevés reçus et disponibles) par la question montrée à la figure 7, et 32 % ont été enregistrés dans la catégorie «relevé reçu, non disponible». Il est clair que cette méthode permet de recueillir des renseignements additionnels qui, sans cela, ne seraient pas connus aussi précisément (si l'on suppose que les relevés fournissent des données plus exactes que celles dont pourraient se souvenir les répondants). Toutefois, il est évident que

certains des cas de la catégorie «relevé reçu, non disponible» n'étaient pas de «réels» cas de «relevés attendus» au moment de l'interview précédente. Pour le moment, il n'est pas clair s'il est préférable, d'une part, de gagner du temps et d'alléger le fardeau du répondant en retardant les questions sur les frais et les paiements tout en améliorant l'exactitude des données déclarées ou, d'autre part, de retarder des questions qui exigeront des répondants un effort pour se rappeler de faits passés.

Figure 7: Écran IPAO de la MCBS -- Question de suivi sur un relevé attendu.

15.01 CPS1 10000027 9112181803 911219
ÉVÉNEMENT: Vous avez obtenu trois fois du PERCODAN

Maintenant, j'aimerais vous demander des renseignements sur des soins médicaux dont nous avons parlé dans une interview précédente.

INTERVIEWER: IL Y A 19 ÉVÉNEMENTS OU GROUPES POUR EXAMEN SOMMAIRE

Voyons d'abord [LIRE L'ÉVÉNEMENT CI-DESSUS].

Au moment de la dernière interview, vous attendiez de recevoir un relevé de Medicare ou du régime d'assurance. Avez-vous reçu un relevé depuis ce temps?

( )

1. RELEVÉ REÇU ET DISPONIBLE
2. RELEVÉ REÇU, NON DISPONIBLE
3. RELEVÉ NON REÇU

5. ANALYSE -- LEÇONS TIRÉES DE LA MCBS

5.1 Obstacles

Cette communication n'a présenté qu'un aperçu de la complexité de la structure du questionnaire IPAO de la MCBS. Bien que nous soyons convaincu que l'observation d'un comportement complexe exige un plan complexe et que l'IPAO a grandement facilité la réalisation d'un tel plan pour la MCBS, des obstacles importants se sont dressés sur notre passage.

La conception de la MCBS a duré un an, depuis l'attribution du contrat jusqu'au début de l'essai préliminaire, soit une période beaucoup plus longue qu'on ne l'avait initialement prévu. Le travail de conception a aussi exigé l'intégration de compétences qui, traditionnellement, sont distinctes dans les activités d'enquête, c'est-à-dire la conception du questionnaire et la conception de la base de données. Les concepteurs du questionnaires devaient comprendre la structure complexe de la base de données, et les concepteurs de la base de données devaient comprendre les objectifs et les techniques de la conception du questionnaire. Des spécialistes de ces deux domaines ont travaillé en commun à la programmation du système IPAO.

Comme nous en avons fait état ailleurs (Edwards et coll. 1992), les interviewers se sont bien habitués à utiliser l'IPAO. Toutefois, les interviewers n'ont pas tous maîtrisé l'ensemble des facettes de la structure du questionnaire. La formation des interviewers a été longue (séances de sept jours sur place avant les rondes 1 et 2, et cinq jours avant la ronde 3, celle-ci étant la première à inclure les caractéristiques longitudinales complexes) et a été axée sur la résolution de problèmes ainsi que sur le déroulement de l'interview proprement dite. Contrairement à une enquête faite avec crayon et papier, dans laquelle les interviewers peuvent contourner les problèmes en écrivant de longues notes et en trouvant la prochaine question pertinente, ou aux interviews téléphoniques assistées par ordinateur (ITAO), au cours desquelles le surveillant du groupe peut être appelé en

cas d'impasse, une étude IPAO exige des interviewers qu'ils résolvent un problème de la bonne façon pour pouvoir continuer l'interview. La MCBS, comme nous l'avons signalé, tente de recueillir des données sur un comportement complexe souvent étayé par des documents sur papier qui sont compliqués. Bien que le recours à l'IPAO ait rendu plus facile la réalisation de l'interview, il n'a pas nécessairement levé l'obstacle de la difficulté du sujet, tant pour l'interviewer que pour le répondant.

Enfin, l'usage abondant de données recueillies antérieurement fait que la base de données est en constante modification pendant de longues périodes au moment des interviews en personne. La conception d'une telle application doit inclure des contrôles stricts quant aux données qu'un interviewer peut inclure ou non dans la base. Par exemple, les régimes d'assurance ou les événements médicaux que nous avons décrits comme étant «supprimés» dans la réalisation des interviews avec rétro-information étaient en fait seulement marqués d'un indicateur de suppression. L'information n'est pas réellement retranchée de la base de données.

5.2 Avantages de l'IPAO pour les enquêtes longitudinales

Malgré les obstacles mentionnés ci-dessus, l'IPAO offre de nombreux avantages dans la conception d'enquêtes longitudinales complexes, le plus manifeste étant l'emploi d'une rétro-information. D'abord, grâce à un programme IPAO, les interviewers peuvent examiner, corriger et enrichir l'information recueillie antérieurement avec une relative facilité, peu importe qu'il s'agisse de données venant de l'interview courante ou d'une interview antérieure. Dans la conception de ce genre d'application, d'importantes décisions doivent être prises, notamment pour déterminer si des corrections et des suppressions seront autorisées et, si oui, à quel moment, quel genre de contraintes imposer à ces interventions et comment manipuler les anciennes et les nouvelles entrées dans la base de données.

Une deuxième utilisation des données recueillies antérieurement dans une structure IPAO consiste à programmer des branchements précis et complexes d'après les réponses précédentes. Encore ici, il peut s'agir de réponses de l'interview courante ou d'une interview antérieure. Une troisième possibilité est de formuler des questions précises d'après les réponses antérieures, notamment en insérant des noms, des pronoms adaptés à la personne, des montants, etc. Cette caractéristique peut comprendre des calculs ou un autre traitement de réponses antérieures, comme (dans l'exemple de la MCBS) le calcul du montant d'une facture qui reste à payer une fois effectué le total de plusieurs paiements ou le classement d'événements en ordre chronologique.

Ces exemples n'épuisent certainement pas les possibilités offertes par l'utilisation de l'IPAO dans les enquêtes longitudinales. La structure de la MCBS montre que, grâce à l'IPAO, les interviewers peuvent manipuler des données d'une interview précédente presque aussi facilement que des données de l'interview courante.

BIBLIOGRAPHIE

Edwards, B., Edwards, W.S., Gay, N., et Sperry, S. (1992). CAPI on the medicare current beneficiary survey: A report on round 1. Discussion présentée au Annual Conference of the American Association of Public Opinion Research, St. Petersburg, FL.

Moore, J., et Kasprzyk, D. (1984). Month-to-month recipiency turnover in the ISDP. *Proceedings of the Survey Research Methods Section, American Statistical Association*, Washington, DC.

Neter, J., et Waksberg, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.

Saris, W.E. (1991). Computer-assisted interviewing. Sage University paper series on quantitative applications in the social sciences, series no. 07-080. Newbury Park, CA: Sage.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

UNE MÉTHODE «COGNITIVE» D'INTERVIEW POUR L'ENQUÊTE «SURVEY OF INCOME AND PROGRAM PARTICIPATION»: ÉLABORATION DES PROCÉDURES ET RÉSULTATS DES ESSAIS INITIAUX

J.C. Moore, K. Bogen et K.H. Marquis[1]

RÉSUMÉ

La présente communication décrit l'élaboration et l'essai initial de procédures expérimentales de collecte des données pour l'enquête «Survey of Income and Program Participation» (SIPP). Les nouvelles procédures découlent d'une recherche antérieure qui a révélé des niveaux graves d'erreur de mesure pour certaines statistiques de base de la SIPP, les implications importantes des erreurs pour les utilisations analytiques courantes des données et qui a laissé supposer que les erreurs avaient des fondements cognitifs. Les principales caractéristiques de ces nouvelles méthodes constituent un message clair et cohérent à *tous* les participants que l'exactitude des réponses est ce qui compte avant tout et un accent sur l'utilisation de dossiers pour aider à déclarer le revenu. Les résultats initiaux d'essais à petite échelle des nouvelles procédures montrent un taux élevé d'utilisation de dossiers pour déclarer les mouvements de revenus et une erreur de réponse moindre (telle qu'indiquée par une réduction dans les erreurs de sous-déclaration et dans le «biais dû à la lisière»); il y a eu des éléments négatifs, les taux de non-réponse ont été considérablement plus élevés pour les essais initiaux que pour la SIPP courante et il se peut que les coûts par cas aient augmenté.

MOTS CLÉS: Recherche cognitive; erreur de mesure; conception de questionnaires; utilisation de dossiers; biais dû à la lisière.

1. INTRODUCTION

1.1 L'enquête Survey of Income and Program Participation

L'enquête Survey of Income and Program Participation (SIPP) est un important programme d'enquête démographique permanente du Census Bureau des É.-U. ainsi qu'une source importante d'indicateurs sociaux et économiques clés pour les États-Unis. Cette enquête à grande échelle fournit les renseignements les plus complets jamais assemblés sur la situation financière des personnes et des familles aux États-Unis. Les données de l'enquête SIPP sont utilisées pour une gamme étendue de décisions stratégiques - assurance-maladie et protection en matière de pensions, réforme fiscale, coûts de la sécurité sociale, efficacité des programmes d'assistance fédéraux et des États, etc.

Dans la conception actuelle de l'enquête, un nouveau panel de l'enquête SIPP est introduit chaque année et il est conservé dans l'enquête pendant environ 2½ années; les ménages dans chaque panel sont interviewés huit fois à des intervalles de quatre mois. Tous les membres des ménages âgés de 15 ans et plus sont admissibles à l'interview. Au cours de chaque interview (ou «cycle») on recueille des données mensuelles pour les quatre mois civils qui précèdent le mois de l'interview. L'auto-déclaration est le mode de déclaration préféré, mais les réponses par personne interposée sont acceptées pour les personnes qui ne sont pas disponibles au moment de la visite de l'intervieweur. Jusqu'à récemment, toutes les interviews de l'enquête SIPP étaient réalisées par visite sur place. Depuis février 1992, toutefois, les interviews 3, 4, 5, 7 et 8 sont des interviews téléphoniques.

---

[1] J.C. Moore, K. Bogen et K.H. Marquis, Center for Survey Methods Research, U.S. Bureau of the Census, Washington (DC), É.-U. 20233-4700.
Les opinions exprimées sont celles des auteurs et ne représentent pas nécessairement les opinions officielles du Census Bureau.

1.2 Aperçu de la communication

Cette communication présente un rapport provisoire sur un programme de recherche qui se poursuit. Elle décrit le projet de recherche cognitive pour l'enquête SIPP (SIPP Cognitive Research Project (SIPP-CR)), dont le but est d'élaborer et de faire l'essai d'autres procédures de mesure pour la SIPP afin de réduire des erreurs de mesure importantes. Dans la section 2, on présente une brève description de la recherche précédente qui a mené à l'étude actuelle. Dans la section 3, on décrit les principales caractéristiques des nouvelles procédures et comment elles diffèrent de celles de la SIPP courante. Dans la section 4, on décrit sommairement le plan de recherche pour le projet SIPP-CR et dans la section 5, on résume les résultats des essais préliminaires qui utilisent les nouvelles procédures. La section 6 renferme quelques brèves conclusions et pensées à propos des prochaines étapes. Des renseignements additionnels sur ces sujets sont présentés dans Marquis, Moore et Bogen (1991).

2. RECHERCHE ANTÉRIEURE SUR LES ERREURS DE MESURE POUR L'ENQUÊTE SIPP

2.1 Le «biais dû à la lisière»

La recherche réalisée auparavant a révélé des problèmes importants d'erreur de mesure dans la SIPP. Dans une des premières études, Burkhead et Coder (1985) ont signalé un «biais dû à la lisière» dans la mesure du changement d'un mois à l'autre à l'aide des données de la SIPP. Le biais dû à la lisière est la tendance pour un nombre beaucoup plus considérable de changements (par exemple, du passage de la participation à la non-participation à un programme de transfert) à apparaître entre les mois adjacents à la «lisière» entre deux cycles d'interview qu'entre deux mois adjacents dans la période de référence d'un seul cycle d'interview. Aucun scénario raisonnable expliquant le changement véritable d'un mois à l'autre ne pouvait laisser prévoir ce genre de comportement; le biais dû à la lisière indique donc clairement que la mesure du changement dans le cadre de la SIPP comporte des problèmes.

2.2 L'étude de vérification de dossiers pour l'enquête SIPP (L'étude de vérification de dossiers pour l'enquête SIPP)

L'étude de vérification de dossiers pour l'enquête SIPP (Moore et Marquis (1989), Marquis et Moore (1990)) a été réalisée afin d'étudier la nature et la portée de l'erreur de réponse dans la SIPP et, plus particulièrement, pour mieux comprendre la nature du biais dû à la lisière et ses causes. Pour l'étude on a utilisé un processus complet de vérification de dossiers afin d'évaluer la qualité des mesures pour les rapports sur la participation à huit programmes de transfert gouvernementaux et sur les revenus découlant de ces programmes, dans quatre États, pour les deux premiers cycles du panel de 1984 de la SIPP.

L'étude de vérification de dossiers a montré que les erreurs de déclaration dans la SIPP sont très rares; globalement, on a trouvé que moins de 2% des déclarations relatives à la participation à des programmes ou à des changements dans la participation à des programmes étaient erronées. L'étude a aussi démontré, toutefois, que même de faibles niveaux d'erreur de réponse peuvent avoir des effets graves sur des estimations importantes, tant des estimations à une variable que des mesures d'association. Pour les taux de participation aux programmes, Marquis et Moore (1990) signalent des sous-estimations nettes allant de 10 à 40%. La sous-estimation des taux de changement dans la participation aux programmes est encore plus considérable dans un cycle (pas à la lisière), alors que les taux de changement à la lisière de l'interview sont considérablement *sur*estimés.

Bien que la recherche portant sur la vérification des dossiers ait permis d'obtenir des descriptions détaillées des erreurs de réponse à l'enquête SIPP, elle s'est révélée beaucoup moins utile pour en déterminer les causes. Marquis et Moore (1990) ont examiné plusieurs des hypothèses traditionnelles relatives aux causes des erreurs de réponse à des enquêtes - oubli, affaiblissement de la mémoire, confusion, biais dû à la réponse par personne interposée, etc. - et trouvé que les données disponibles n'apportaient un appui sérieux à aucune de ces hypothèses.

2.3 Recherche cognitive exploratoire pour l'enquête SIPP

En cherchant davantage pour obtenir une meilleure compréhension intuitive causale, le Census Bureau a réalisé un projet de recherche cognitive exploratoire à petite échelle afin de chercher des indices des difficultés liées aux erreurs de réponse à l'enquête dans la compréhension qu'ont les répondants des tâches et des questions relatives à l'enquête SIPP et dans leurs processus de pensée quand ils répondent à ces questions. Des employés de l'administration centrale du Census Bureau ont reçu une formation en techniques d'«interview cognitive» et accompagné des intervieweurs expérimentés de l'enquête SIPP alors que ces derniers réalisaient l'interview de l'enquête SIPP courante. Les observateurs pouvaient lancer des questions pour savoir comment le répondant interprétait une tâche générale ou une question particulière, comment il formulait une réponse, etc., ou ils pouvaient simplement observer l'interaction entre l'intervieweur et le répondant. Ce projet ainsi que ses résultats sont résumés dans Marquis (1990).

La recherche cognitive exploratoire a permis de comprendre plusieurs aspects importants de la dynamique de l'erreur de réponse lors de l'interview de l'enquête SIPP. Un de ces aspects clés était le rôle limité que la mémoire joue dans la soi-disant «évocation» que font les répondants de leur revenu pendant la période de référence. À la place d'une évocation détaillée directe de l'histoire des paiements pendant quatre mois, les répondants ont tendance à s'en remettre à des règles très simples, combinées avec quelques faits dont ils se souviennent pour construire une histoire plausible (bien que pas nécessairement exacte) à propos de leur revenu. De plus, cette stratégie abrégée de «fabulation» semble être non seulement tolérée par les procédures actuelles de l'enquête SIPP, mais encouragée de façons subtiles mais importantes. Par exemple, les évaluations du rendement des intervieweurs sont surtout basées sur leurs taux de réponse et sur leur efficience. Cette façon d'agir peut nuire à la qualité des réponses, parce qu'elle peut décourager les intervieweurs «d'insister davantage» pour obtenir des réponses exactes, soit par le recours à des dossiers, soit par des stratégies d'évocation difficiles et plus complexes (faire un suivi quand on répond «je ne sais pas» et lorsqu'on refuse de répondre à une question, ou réagir aux élaborations des répondants suite à des réponses trop simples ou même reconnaître ces situations). L'environnement dans lequel se fait l'évaluation peut aussi encourager les intervieweurs à se dépêcher pour réaliser l'interview et même à aider activement les répondants à trouver des approximations faciles à une vérité complexe.

La recherche cognitive a aussi attiré l'attention sur de nombreuses façons par lesquelles le questionnaire de l'enquête SIPP courante cause des problèmes au niveau de la qualité des réponses. Dans de nombreux domaines, le questionnaire présente des exigences du point de vue mémoire qui sont tout simplement déraisonnables (par exemple, le fait de demander des détails très peu réalistes ou des renseignements qui n'auront vraisemblablement jamais été mémorisés); dans d'autres domaines, on ne donne effectivement pas aux répondants l'occasion de déclarer des données avec exactitude en vue d'améliorer l'efficience du traitement (par exemple, l'obligation de déclarer tout le revenu en «blocs mensuels», même un revenu qui est versé selon un calendrier qui ne peut être exprimé facilement en mois). Ces lacunes empêchent les répondants de répondre avec exactitude et les incitent à «fabuler». De plus, le questionnaire de l'enquête SIPP n'est pas à la hauteur, car il ne fournit pas aux répondants des renseignements clairs et cohérents à propos de la nature de leur tâche. Les observateurs des interviews cognitives ont souvent trouvé que les répondants ne comprenaient pas l'objet d'une série complète de questions. Cela était parfois dû au manque d'énoncés explicatifs de transition entre des domaines importants; dans d'autres cas, le défaut est clairement imputable à la complexité de l'instrument, avec sa myriade de questions de sélection et de postes réservés à l'intervieweur (que ce dernier lit souvent à haute voix, perturbant davantage le cheminement des questions et le contexte du questionnaire). De plus, l'instrument ne fournit pas des renseignements adéquats ou cohérents à propos du niveau d'exactitude ou d'effort attendu de la part des répondants.

3. UNE AUTRE CONCEPTION DES MESURES POUR L'ENQUÊTE SIPP

La recherche cognitive exploratoire a permis de comprendre plusieurs aspects importants des causes probables des problèmes de l'enquête SIPP pour ce qui est des erreurs de réponse et a mené directement à un bon nombre des changements incorporés dans un ensemble d'autres procédures de mesure pour l'enquête. Voici quelles sont les principales composantes des nouvelles procédures:

- La pierre angulaire des nouvelles procédures de mesure est l'accent mis sur l'*utilisation de dossiers sur le revenu personnel* par les répondants pour aider ces derniers à déclarer leur revenu. Il est habituellement très difficile, et même virtuellement impossible, de se remémorer avec exactitude l'ensemble des revenus pendant une période de quatre mois; les procédures révisées reconnaissent ce fait explicitement. Elles tentent d'enlever complètement la tâche de déclaration de la tête des répondants (et ne sont donc pas réellement «cognitives») en insistant plutôt pour que pour les répondants utilisent leurs dossiers personnels pour déclarer leur revenu, afin de les empêcher d'employer des stratégies de réponse trop simples et d'assurer l'exactitude des réponses. Les intervieweurs sont aussi chargés de montrer aux répondants comment interpréter leurs dossiers et comment les tenir à jour pour la prochaine interview. Cela comprend le fait de remettre aux répondants une chemise pour conserver leurs documents entre les interviews et, pour un revenu qui n'est pas accompagné d'un document, une feuille sur laquelle il faut inscrire les détails pertinents à propos de ce revenu (date, montant, source et bénéficiaire).

- En l'absence de dossiers sur le revenu, les intervieweurs sont formés pour reconnaître les stratégies abrégées inacceptables et pour guider les répondants afin que ces derniers *utilisent des stratégies d'évocation plus réalistes*. Dans de telles circonstances, on demande tout d'abord aux répondants de décrire la distribution «habituelle» des dates et des montants des paiements puis de dresser la liste des facteurs qui peuvent avoir une incidence sur les dates ou les montants des paiements; ensuite, de considérer si l'un quelconque des facteurs de «changement» possibles s'est produit pendant la période de référence et, dans l'affirmative, quand; et finalement, à partir de cet ensemble complexe de renseignements, de reconstruire ce qui s'est effectivement produit pendant la période de référence.

- Pour éviter «de se faire conter des histoires», pour renforcer le message que l'exactitude est le but premier de l'enquête et pour faciliter l'utilisation des dossiers, la nouvelle procédure *recueille chaque paiement de revenu «au cent près»*, et non des totaux mensuels. Peu importe le nombre de fois où les répondants reçoivent un revenu d'une source particulière, les intervieweurs recueillent les dates et les montants pour chaque paiement. Les totaux mensuels sont produits par ordinateur, pas dans la tête des répondants. Même dans le cas des sources de revenu pour lesquelles les utilisateurs des données peuvent ne pas avoir besoin des montants exacts, on recueille les renseignements avec le même degré de précision afin de s'assurer qu'un message uniforme est transmis aux répondants à l'effet que l'exactitude est essentielle et que des estimations ne sont pas acceptables.

- Les nouvelles procédures utilisent des *techniques d'interview non normalisées pour la collecte de renseignements sur le revenu*. L'autre interview de l'enquête SIPP commence avec une section à «rappel libre» au cours de laquelle on tente d'établir clairement les buts de la section, on permet ensuite aux répondants de déterminer dans une large mesure comment ils déclareront leur revenu pour la période de référence. Cette partie de l'interview est structurée - les buts visés par la collecte des renseignements sont explicites et le mécanisme de saisie des données donne des indications claires à propos des données précises requises. C'est le texte inviolable, avec des questions établies à l'avance dans un ordre prédéterminé qui manque. Cette façon d'agir peut présenter de nombreux avantages. Le fait de permettre aux répondants de déclarer des renseignements importants à propos de leur revenu sans trop de retard, dans l'ordre le plus naturel, sans avoir à entendre de longues séries de questions inapplicables ou qui semblent hors de propos, laisse les répondants participer immédiatement à l'interview et à la production de bons renseignements. Cette façon d'agir donne aussi à l'interview la flexibilité nécessaire pour traiter la grande complexité et la diversité considérable de situations relatives au revenu des personnes. Moore, Bogen et Marquis (1992) fournissent une description complète de cet aspect des procédures révisées de l'enquête SIPP.

- Les nouvelles procédures tentent de *simplifier le plus possible les tâches de déclaration*, et *expliquent clairement aux répondants les desseins et les buts* de chaque section. L'ordre des questions a été modifié afin de rendre les sections du questionnaire plus cohérentes. Ce changement, ainsi que les procédures de «rappel libre» décrites plus haut, ont éliminé la nécessité d'employer un bon nombre d'instructions «passez à» complexes, ce qui permet aux intervieweurs de se concentrer sur leur tâche essentielle, qui n'est plus de lire des questions, mais de résoudre des problèmes. Dans certains cas, on a choisi pour le questionnaire révisé de poser des questions à un univers de répondants un peu plus grand que nécessaire, afin d'éliminer les questions de sélection qui précédaient ces questions. Un autre changement est l'ajout de brefs énoncés

de transition entre les principales sections du questionnaire, afin de fournir aux répondants une indication de ce qui va suivre.

- Pour la première interview, les nouvelles procédures insistent sur l'*autodéclaration, de préférence «de style familial», dans un cadre d'interview qui favorise la concentration*. Ces composantes de l'interview révisée ont pour but à la fois de renforcer le message que l'enquête cherche des renseignements de la plus haute qualité et de fournir un milieu qui est le plus propice à atteindre une qualité élevée. Dans les interviews ultérieures, si le ménage dispose de dossiers, on peut assouplir les règles relatives à l'autodéclaration et à l'interview de groupe; l'objet initial de ces règles est de s'assurer que tous les membres du ménage comprennent les buts et l'importance de l'enquête, de permettre à ceux-ci de s'entraider pour déterminer les sources de revenu et les détails pertinents et aussi de fournir une approbation implicite aux membres du ménage pour que ces derniers se partagent des renseignements sur le revenu, préparant ainsi la voie pour la collecte de renseignements de qualité par personne interposée (à l'aide de dossiers, bien entendu) au cours des interviews ultérieures.

- Pour s'attaquer directement au biais dû à la lisière, en particulier à la surdéclaration du changement à la lisière, les procédures révisées utilisent des *périodes de référence qui se chevauchent avec rapprochement* de renseignements divergents, une technique adaptée de Murray et coll. (1991). Contrairement à la situation pour l'enquête SIPP courante, la période de référence pour chaque cycle s'étend jusqu'à la date de l'interview, plutôt que de se terminer le dernier jour du mois complet précédant l'interview. Puisque la période de référence pour l'interview suivante commence au début du mois au cours duquel l'interview précédente a eu lieu, pour les interviews qui suivent la première il y a une période de chevauchement visée tant par l'interview courante que par l'interview précédente. Lors de la deuxième interview et de celles qui suivent, l'intervieweur recueille tout d'abord des renseignements sur le revenu indépendants de ceux fournis lors de l'interview précédente, puis il révise les renseignements avec les répondants en tenant compte des renseignements déjà fournis. Cette révision comprend deux étapes. Premièrement, l'intervieweur résout tout problème lié à des écarts dans les sources de revenu, vérifiant toutes les sources de revenu déclarées dans une interview mais pas dans l'autre au cas où il y aurait eu des omissions. Après cela, les intervieweurs révisent les données pour les deux cycles au cas où il y aurait des renseignements divergents sur le revenu déclaré pendant la période de chevauchement et ils résolvent, avec l'aide des répondants, tous les cas de divergences.

- Un ensemble de *critères révisés pour l'évaluation des intervieweurs* constitue un élément essentiel des nouvelles procédures, ces critères ont pour but d'encourager les intervieweurs à rechercher un rendement basé sur la qualité. Selon les procédures révisées, l'accent principal et presque exclusif n'est plus mis sur des taux de réponse élevés et une efficience élevée, mais on ajoute de nombreux indicateurs de la mesure dans laquelle le rendement des intervieweurs est compatible avec les buts principaux en matière de qualité et on accorde la priorité à ces indicateurs. C'est l'examen d'un échantillon d'interviews enregistrées sur bande magnétique (toutes les interviews sont supposées être enregistrées) dans des domaines tels que l'obtention d'interviews de groupe et d'autodéclarations, le fait de persuader des répondants d'utiliser des dossiers, de reconstituer les détails du revenu en l'absence de dossiers à l'aide de stratégies d'évocation complexes, de fournir de la rétroaction aux répondants, de reconnaître et de résoudre les difficultés des répondants, etc., qui constitue la principale forme de rétroaction.

4. LE PLAN DE RECHERCHE

Le Census Bureau a conçu un programme de recherche, dont la réalisation sur le terrain est présentement en cours, afin d'évaluer et d'améliorer les procédures «cognitives» révisées. Ce programme comprend deux petits essais préliminaires, une étude complète d'évaluation de la qualité des mesures (qui est actuellement réalisée sur le terrain) et un panel pour la recherche sur la mise en application afin de s'attaquer aux questions opérationnelles.

4.1 Essai préliminaire 1

Le premier essai préliminaire a été réalisé à Milwaukee, WI, d'août à novembre 1991. Les interviews du 1er cycle ont été réalisées en août et en septembre, avec une période de référence normale de quatre mois; les interviews du 2e cycle, avec une période de référence réduite de seulement deux mois[2], ont été réalisées en octobre et en novembre dans les ménages qui avaient participé à l'interview initiale deux mois auparavant. L'échantillon était composé de 130 adresses choisies au hasard. L'objet du premier essai préliminaire était d'évaluer la faisabilité des nouvelles procédures et des nouveaux instruments utilisés sur le terrain et de les améliorer, au besoin.

4.2 Essai préliminaire 2

Pour le second essai préliminaire, on a utilisé la même conception générale que pour le premier: deux mois d'interviews du 1er cycle en décembre 1991 et en janvier 1992, avec une période de référence de quatre mois et deux mois d'interviews du 2e cycle en février et en mars 1992, avec une période de référence de deux mois. L'échantillon utilisé pour le 2e essai préliminaire était composé de 130 particuliers (et des membres du ménage dont ils faisaient partie), qui habitaient Milwaukee et qui avaient été repérés dans les systèmes de dossiers officiels comme ayant reçu un revenu d'une des cinq sources suivantes: Aid to Families with Dependent Children (aide aux familles avec enfants à charge) (AFDC), Food Stamps (coupons alimentaires), assurance-chômage, Supplemental Security Income (revenu supplémentaire de sécurité sociale) (SSI) ou des gains provenant d'un employeur déterminé de la région de Milwaukee. L'objet du second essai préliminaire était de faire l'essai des procédures utilisées pour effectuer l'échantillonnage et l'appariement à l'aide de dossiers administratifs et d'employeurs; d'élaborer des stratégies et des programmes pour la saisie des données, la gestion des bases de données et l'analyse des données et de faire un essai plus poussé des procédures et instruments révisés et d'améliorer ces procédures et instruments.

4.3 L'étude d'évaluation

L'étude d'évaluation est actuellement en cours, aussi à Milwaukee. Quand elle sera terminée, elle comprendra deux cycles d'interviews, chacun avec une période de référence complète de quatre mois. Les interviews du 1er cycle ont commencé en septembre 1992 et se poursuivront jusqu'en janvier 1993; les interviews du 2e cycle seront réalisées de février à mai 1993. Comme pour le second essai préliminaire, les cas dans l'échantillon sont composés de particuliers (et des membres du ménage dont ils font partie) tirés des systèmes de dossiers d'une parmi cinq sources de revenus. Le but visé est de réaliser environ 350 interviews du 2e cycle selon deux méthodes attribuées aléatoirement: procédures de mesures de l'enquête SIPP courante et procédures révisées.

L'objet de l'étude d'évaluation est de fournir une comparaison directe de la qualité des mesures entre les deux méthodes, à l'aide de dossiers administratifs et d'employeurs utilisés comme principaux critères pour évaluer la qualité. La participation aux programmes (ainsi que l'emploi) et les montants tels que déclarés par les répondants seront comparés aux «vrais» renseignements figurant dans les dossiers. De plus, des comparaisons des éléments de coût (temps de déplacement, durée des interviews, temps de contrôle, etc.) seront faites entre les deux méthodes afin d'évaluer les coûts des nouvelles procédures et de déterminer les causes de tout écart entre les coûts. Finalement, en plus d'une simple comparaison des taux de non-réponse, les données dans les dossiers permettront d'effectuer certaines comparaisons des caractéristiques des non-répondants selon les méthodes utilisées, ce qui pourra fournir une indication des écarts dans le biais dû à la non-réponse entre les deux méthodes.

4.4 Recherche sur la mise en application

Si les renseignements obtenus à la suite de l'étude d'évaluation montrent que les nouvelles procédures permettent d'apporter des améliorations considérables à la qualité, avec des coûts et un taux de non-réponse raisonnables, des recherches additionnelles seront réalisées afin d'étudier les nombreuses questions opérationnelles qui

---

[2] Dans les deux essais préliminaires, la période de référence du 2e cycle a été réduite afin de permettre au programme de recherche de respecter les dates limites du calendrier de remaniement de l'enquête. Cet aspect de la conception des essais préliminaires peut avoir eu une incidence sur des résultats principaux, particulièrement ceux relatifs à la réduction apparente du biais dû à la lisière (voir la Section 5.2).

resteront inévitablement à régler (par exemple, la possibilité d'étendre la méthode à d'autres lieux, la collaboration des répondants pendant plusieurs cycles, l'utilisation de l'interview sur place ou de l'interview téléphonique assistée par ordinateur, les effets différentiels sur les sous-groupes, les coûts et l'incidence sur la qualité des réponses des composantes particulières des nouvelles procédures, etc.). La conception et les buts exacts de cette recherche sur la mise en application (ou opérationnelle) restent encore à préciser.

5. RÉSULTATS DES ESSAIS PRÉLIMINAIRES

L'objet principal du premier essai préliminaire et aussi un but important du second, était de faire l'essai sur le terrain des nouvelles procédures et des nouveaux instruments et de déterminer et corriger les problèmes les plus évidents. Bien qu'aucune des caractéristiques de base des nouvelles procédures se soit révélée irréalisable sur le terrain (et que plusieurs aient eu un succès surprenant), pendant tous les essais préliminaires, de nombreuses améliorations ont été apportées aux procédures et aux instruments à la suite de situations rencontrées sur le terrain et des observations des intervieweurs. Le second essai préliminaire a été très instructif à propos de l'échantillonnage à partir des divers systèmes de dossiers, comme essai de ces procédures pour l'étude d'évaluation. La fréquence avec laquelle la liste des ménages pour l'adresse fournie par l'organisme ou l'employeur n'incluait pas la personne cible dans l'échantillon est un résultat important de l'essai[3]. Pour expliquer cette attrition, on a choisi un plus grand nombre de cas dans l'échantillon pour le 1er cycle de l'étude d'évaluation. Un autre but du second essai préliminaire était de faire l'essai des procédures de saisie des données. Cet exercice, aussi, a été très instructif, faisant ressortir le besoin d'apporter des modifications importantes pour la prochaine phase de la recherche.

Le reste de la présente section résume les résultats des essais préliminaires dans trois domaines: la mise en application réussie des nouvelles procédures sur le terrain qui visent la qualité, les indicateurs de la qualité améliorée des mesures avec les nouvelles procédures et les domaines qui nécessitent manifestement des améliorations - les taux de non-réponse et les coûts.

5.1 Mise en application des procédures visant la qualité

Enregistrement sur bande magnétique. Les intervieweurs ont eu un succès raisonnable pour ce qui est d'enregistrer sur bande magnétique les interviews des essais préliminaires, bien que des améliorations puissent certainement être apportées. Dans chaque essai, environ 75% de toutes les interviews réalisées ont été enregistrées sur bande magnétique. Il vaut la peine de remarquer que, selon les rapports des intervieweurs, seulement un ou deux des cas où un enregistrement n'a pu être réalisé étaient attribuables aux répondants[4]. Dans presque tous les cas, le fait que l'enregistrement n'a pas été réalisé était dû à une défaillance mécanique, à une erreur de l'opérateur, ou à ce que l'intervieweur n'avait pas demandé au répondant s'il pouvait enregistrer l'interview (ce qui s'est produit souvent dans les cas de conversion de refus). Ces résultats prouvent de façon assez convaincante que l'enregistrement sur bande magnétique ne constitue pas une question importante pour les répondants.

Par contre, le système d'évaluation et d'amélioration du rendement des intervieweurs, dans son ensemble, la raison pour laquelle les enregistrements étaient effectués, n'a pas donné de très bons résultats lors des essais préliminaires. Les données objectives sont peu abondantes, mais il semble que des problèmes importants se soient produits lors de la conversion des résultats de l'examen des bandes en rétroaction utile pour ce qui est du rendement des intervieweurs. Le délai d'exécution qui était souvent beaucoup trop long, constituait un problème. Une difficulté plus fondamentale et pour laquelle aucune solution opérationnelle n'est immédiatement

---

[3] Les taux d'appariement observés – le taux auquel la personne cible dans l'échantillon était trouvée dans la liste des membres du ménage pour les ménages interviewés lors du 1er cycle – allaient d'un minimum de 68% pour la AFDC à 96% pour les échantillons de l'employeur et de l'assurance-chômage (Unemployment Compensation).

[4] Le fait que la non-réponse au 2e cycle était plus élevée dans les essais préliminaires de la SIPP-CR que ce n'est généralement le cas pour la SIPP courante (voir la Section 5.3) pourrait laisser supposer une réaction négative de la part des répondants au 1er cycle aux nouvelles procédures, y compris, peut-être, le besoin d'enregistrer les interviews. Toutefois, il n'y a aucune mention explicite d'un problème relatif aux enregistrements sur bande magnétique dans aucun des rapports de non-interview pour le 2e cycle.

apparente, est celle des réactions négatives des intervieweurs à l'examen des bandes. Nous visions à fournir un système continu de formation en cours d'emploi qui aiderait les intervieweurs à améliorer leur rendement; toutefois, les intervieweurs ont eu tendance à ne voir aucun aspect positif dans cette façon de procéder. Certains estimaient que le système ne tenait pas compte de toutes les interactions au cours des interviews qui ne se prêtent pas à un enregistrement audio et considéraient cet enregistrement surtout comme une façon de compter et d'étayer leurs erreurs. Les intervieweurs ont toutefois admis que les formules de contrôle communiquaient très clairement les buts de l'interview qui étaient prioritaires ainsi que les comportements qui nous intéressaient le plus.

Interviews de groupe et autodéclaration. Il semble que la mise en application des procédures d'interview de groupe et d'autodéclaration ait été couronnée de succès lors du 1er essai préliminaire (les données pour le 2e essai préliminaire ne sont pas encore disponibles). Les trois quarts de tous les adultes interviewés qui faisaient partie de ménages comptant plusieurs adultes ont participé à une interview de groupe et 92% de tous les adultes interviewés ont fourni leurs réponses par autodéclaration. Les procédures pour l'enquête SIPP courante donnent généralement un taux d'autodéclaration d'environ 65%. (L'enquête SIPP courante ne permet que d'interviewer des particuliers, on ne dispose donc pas de chiffres pour les interviews de groupe qui permettraient d'effectuer des comparaisons.)

5.2 Indicateurs de la qualité améliorée

Les preuves les plus concluantes de la qualité des données découlent, bien entendu, de l'appariement des données d'enquête avec les dossiers administratifs et des employeurs. Un ensemble limité de tels résultats appariés enquête/dossiers est actuellement disponible à partir des données du 2e essai préliminaire. Ces résultats, ainsi que deux autres ensembles d'analyses tirés des deux essais préliminaires - l'utilisation que font les répondants des dossiers ainsi qu'un biais dû à la lisière réduit - laissent supposer que les procédures révisées permettent effectivement d'améliorer la qualité des données.

Utilisation de dossiers. L'utilisation qu'ont fait les répondants de dossiers au cours des essais préliminaires a de loin dépassé nos attentes. Au niveau des ménages, 87% de tous les ménages (pour les deux essais préliminaires combinés) ont présenté au moins un document pour aider à déclarer le revenu, avec très peu de différence entre le 1er cycle et le 2e. Le taux d'utilisation de dossiers au niveau des sources de revenu était de 72% - c'est-à-dire que pour 72% des sources de revenu déclarées par les répondants, au moins un document a été utilisé pour justifier la date et le montant d'un paiement. De même, au niveau des paiements, les répondants ont utilisé des dossiers pour déclarer 63% des paiements qu'ils ont reçus. Le taux d'utilisation des dossiers au niveau des paiements pour le 2e cycle était de 74%, comparativement à 57% pour le 1er cycle, ce qui laisse supposer, à nouveau, que, bien qu'on puisse apporter encore beaucoup d'améliorations, les intervieweurs ont réussi à former les répondants à la tenue de dossiers entre les interviews[5].

Les procédures de l'enquête SIPP courante encouragent aussi les intervieweurs à demander aux répondants d'utiliser des dossiers. Selon les résultats résumés par Singh (1991, 1992), le taux d'utilisation des dossiers au niveau de la source de revenu était d'environ 20% dans les premiers cycles du panel de 1991 de l'enquête SIPP. Le succès plutôt limité de l'enquête SIPP courante à cet égard peut être attribuable en partie au fait que les intervieweurs craignent que s'ils demandent la production de dossiers, cela irritera les répondants, entraînant des ruptures et par la suite une non-réponse et que cela augmentera aussi la durée des interviews, diminuant ainsi leur efficience.

Biais dû à la lisière. Une analyse du biais dû à la lisière fait ressortir des preuves plus directes de la qualité améliorée avec les procédures révisées de l'enquête SIPP. Le tableau 1 montre un «indice global du biais dû à la lisière» - le rapport entre le nombre moyen de changements d'un mois à l'autre à la lisière et le nombre moyen de changements ailleurs qu'à la lisière - pour chaque essai préliminaire, groupés pour tous les genres de

---

[5] Un test t simple fait avec l'hypothèse d'indépendance des échantillons est significatif. Le fait de tenir compte de la corrélation entre les observations du 1er cycle et celles du 2e cycle ne change pas la conclusion qui découle du test original. Puisque certaines personnes ne figurent que dans un des cycles, nous avons ré-estimé les proportions d'utilisation des dossiers en n'incluant que les personnes qui faisaient partie des deux cycles. Les résultats sont fort semblables, nous concluons donc que l'utilisation de tous les cas disponibles ne cause pas de distorsion importante dans la conclusion qu'il y a différence.

revenus. Un indice dont la valeur est 1.0 montre qu'il n'y a pas de biais dû à la lisière; c'est-à-dire que l'indice est 1.0 si le nombre de transitions mesurées à la lisière est identique au nombre de transitions dans une paire de mois moyenne ailleurs qu'à la lisière. Pour le 1er essai préliminaire, l'indice global du biais dû à la lisière est .95; pour le 2e essai préliminaire, cet indice est légèrement plus élevé (1.55), ce qui est encore beaucoup plus faible que les résultats déclarés par Burkhead et Coder (1985) pour l'enquête SIPP courante[6].

Tableau 1: Résultats relatifs au biais dû à la lisière pour les essais préliminaires de la SIPP-CR et la SIPP courante.

INDICE DU BIAIS DÛ À LA LISIÈRE:	
SIPP-CR 1er essai préliminaire:	0.95
SIPP-CR 2e essai préliminaire:	1.55
Indices représentatifs du biais dû à la lisière pour l'enquête SIPP (Burkhead et Coder 1985):	
Assurance-chômage:	1.9
Gains:	2.2
Coupons alimentaires:	3.5
Sécurité sociale:	3.9
AFDC:	4.9
Pensions de sources privées:	6.3

Nous pouvons nous interroger sur les raisons pour lesquelles les nouvelles procédures de l'enquête SIPP semblent mieux répartir le changement déclaré. Marquis et Moore (1989) ont montré que le biais dû à la lisière est le résultat net à la fois d'une sous-déclaration des changements lors d'une interview (ailleurs qu'à la lisière) et d'une surdéclaration des changements entre les interviews (à la lisière). Le fait que, dans les nouvelles procédures, on mette l'accent sur chaque paiement peut encourager les répondants à déclarer le revenu touché dans tous (ou du moins dans la majorité de) ses détails compliqués. Les procédures courantes de l'enquête SIPP, parce qu'elles se concentrent sur des agrégats mensuels, éloignent les répondants des détails et les incitent à conter une brève histoire plausible. Lors de l'interview suivante, il se peut que les répondants content une histoire plausible légèrement différente; ce processus peut donc permettre de minimiser le changement au cours d'une interview et le forcer à paraître à la lisière. D'autres modifications aux procédures qui peuvent aussi avoir contribué à la réduction du biais dû à la lisière sont l'utilisation de périodes de référence qui se chevauchent, la solution des problèmes liés aux écarts entre les sources de revenu d'une interview à l'autre et la solution des problèmes liés aux écarts dans le revenu qu'on déclare avoir touché pendant la période de chevauchement. Il faut aussi remarquer que la réduction du biais dû à la lisière est, dans une mesure inconnue, un artefact de la conception des essais préliminaires, pour lesquels on a utilisé une période de référence réduite de 2 mois lors du 2e cycle plutôt que la période de référence de 4 mois utilisée pour l'enquête SIPP courante.

Erreurs de sous-déclaration. Comme nous l'avons mentionné plus haut, les données tirées des dossiers administratifs et des employeurs disponibles lors du 2e essai préliminaire renferment une preuve directe des effets des erreurs de mesure des procédures de la SIPP-CR. Jusqu'ici, les données déclarées dans le cadre de l'enquête par les participants connus aux programmes ont été évaluées en les comparant à une «vérité» basée sur des dossiers pour deux programmes, les coupons alimentaires (Food Stamps) et le revenu de sécurité sociale (Supplemental Security Income) (SSI).

Le tableau 2 présente les taux d'erreur liés à la sous-déclaration de la participation mensuelle aux programmes - c'est-à-dire, la proportion de mois au cours desquels les répondants ont «véritablement» participé aux

[6] Les résultats relatifs au biais dû à la lisière pour la SIPP-CR sont basés sur des données provenant de tous les ménages pour lesquels des interviews complètes ont été réalisées au cours des deux cycles d'interview; 74 lors du 1er essai préliminaire et 79 lors du 2e. Les données mentionnées par Burkhead et Coder sont tirées des trois premiers cycles d'interview du panel de 1984 de la SIPP, qui est composé d'environ 20 000 ménages.

programmes alors qu'ils n'ont rien déclaré à ce sujet dans le cadre de l'enquête[7]. Comme la taille de l'échantillon utilisé pour le 2e essai préliminaire est petite et à cause des différences considérables au niveau conceptuel entre les essais préliminaires et l'enquête SIPP courante, nous n'avons pas tenté d'effectuer de tests statistiques et ne pouvons donc faire aucune affirmation concernant la signification statistique des différences observées. Néanmoins, les données limitées dont nous disposons laissent supposer, à nouveau, que les procédures révisées vont dans la bonne direction pour ce qui est d'apporter des améliorations importantes à la mesure de statistiques clés de l'enquête SIPP.

Tableau 2: Sous-déclaration de la participation aux programmes pour les essais préliminaires de la SIPP-CR et la SIPP courante.

SOUS-DÉCLARATION DE LA PARTICIPATION MENSUELLE AUX PROGRAMMES:

% des mois de participation "véritable" déclarés comme "sans participation":

	SIPP-CR	SIPP courante
Coupons alimentaires	9.7%	23.7%
SSI	11.1%	23.2%

(Les résultats pour la SIPP courante sont tirés de Marquis et Moore 1990)

(Résultats de l'enquête SIPP courante (Marquis et Moore 1990))

5.3 Domaines qui nécessitent des améliorations - Non-réponse des ménages et coûts

Les essais préliminaires n'étaient pas conçus pour fournir des comparaisons opérationnelles définitives avec les procédures de l'enquête SIPP courante. Toutefois, les données relatives aux essais préliminaires laissent supposer qu'il se peut que, telles qu'elles sont conçues et mises en application actuellement, les nouvelles procédures donnent des résultats bien en-deçà du rendement de l'enquête SIPP courante dans deux domaines clés - la non-réponse et les coûts.

Non-réponse. Pour les deux essais préliminaires combinés, le taux de réponse des ménages du 1er cycle (le nombre de ménages interviewés divisé par le nombre de ménages admissibles) était de 73%; pour le 2e cycle, ce taux (basé seulement sur les ménages interviewés lors du 1er cycle) était de 87%, ce qui donne un taux de réponse longitudinal de 63%. Ce taux montre la proportion des ménages admissibles du 1er cycle qui ont été interviewés lors des deux cycles. Pour l'enquête SIPP courante, on obtient un taux de réponse au 1er cycle d'environ 92% et un taux longitudinal lors du 2e cycle d'environ 88%. Bien que les taux de l'enquête SIPP ne soient pas exactement comparables à ceux des essais préliminaires, à cause de différences dans les procédures (par exemple, pour l'enquête SIPP courante ont suit les personnes qui déménagent, alors que cela n'était pas prévu dans les procédures appliquées pour les essais préliminaires de la SIPP-CR), il est assez évident que les taux de réponse lors des essais préliminaires étaient beaucoup plus faibles dès le début du 1er cycle et que l'attrition a aussi vraisemblablement été plus élevée lors du 2e cycle.

Nous avons examiné les descriptions qu'ont fait les intervieweurs des circonstances entourant chaque non-interview à laquelle ils avaient fait face pour essayer de trouver des indications que les nouvelles procédures causaient le taux plus élevé de non-réponse. À une ou deux exceptions près, il n'y a pas suffisamment de preuves dans ces rapports qu'une non-interview quelconque résultait directement des nouvelles procédures. Les non-interviews causées par l'absence de répondants au domicile, qui représentaient entre 20 et 25% des cas de non-interviews, ne sont vraisemblablement pas dûs à une procédure spéciale quelconque utilisée pour l'enquête, certainement pas lors du 1er cycle d'interview. La majorité des non-interviews étaient des refus. Dans presque tous les cas, les refus lors du 1er cycle se sont produits avant que l'intervieweur ait même pu commencer à

[7] Les personnes faisant partie de l'échantillon du 2e essai préliminaire de la SIPP-CR ont connu 165 mois de participation véritable au programme des coupons alimentaires (Food Stamps), selon les dossiers administratifs pour lesquels elles ont déclaré 149 mois de participation lors de l'interview de la SIPP-CR; dans le cas du SSI, les chiffres comparables sont de 135 mois de participation véritable dont 120 ont été déclarés. Les données présentées par Marquis et Moore sont tirées de l'étude de vérification de dossiers de la SIPP (SIPP Record Check Study), qui faisait appel à un sous-ensemble de trois États pour les deux premiers cycles d'interview du panel de 1984 de la SIPP. Pendant cette période, les personnes admissibles de l'échantillon de la SIPP ont connu 1 451 mois de participation véritable au programme des coupons alimentaires (Food Stamps), d'après les dossiers administratifs, dont 1 107 ont été déclarés lors de l'interview de la SIPP courante; pour le SSI, les chiffres comparables sont de 919 mois de participation véritable dont 706 ont été déclarés.

expliquer l'objet de l'enquête et ce que la collaboration à l'enquête entraînait. Bien que, par définition, les personnes qui ont refusé de répondre à l'interview lors du 2e cycle connaissaient ce que l'interview comportait, même ces personnes, d'après les rapports des intervieweurs, n'ont pas mis en cause l'une quelconque des procédures cognitives pour expliquer leur refus de collaborer. Les personnes n'ont pas refusé de participer parce qu'on leur a demandé d'obtenir des documents ou parce qu'on était pour enregistrer l'interview. Il semble que les problèmes de non-réponse lors des essais préliminaires aient été beaucoup plus de nature administrative; il est arrivé souvent que les personnes qui pourraient éventuellement refuser de répondre et les répondants difficiles à retracer n'étaient pas repérés assez tôt pour que les mesures correctives appropriées puissent être prises, ou, s'ils étaient repérés tôt, il est arrivé souvent que les mesures de suivi n'étaient pas prises immédiatement.

Coûts. Bien qu'il soit difficile de comparer directement les coûts des essais préliminaires de la SIPP-CR aux coûts de l'enquête SIPP courante (à cause des tâches beaucoup plus petites dans l'enquête SIPP-CR, par exemple, et d'un plan d'échantillonnage comportant beaucoup de grappes pour l'enquête SIPP courante), il est évident que lors des essais préliminaires de l'enquête SIPP-CR, les coûts sur le terrain ont été considérablement plus élevés que ceux associés à la réalisation courante de l'enquête, peut-être jusqu'à 50% plus élevés. Une hypothèse évidente pour expliquer cette situation est le fait qu'on peut imputer les augmentations de coûts à certaines des caractéristiques des nouvelles procédures - le plus possible d'autodéclaration, des interviews de groupe, le fait d'insister pour disposer d'un cadre approprié dans lequel réaliser les interviews, l'utilisation de dossiers, etc., puisque ces procédures exigeaient de nombreuses visites additionnelles aux ménages, visites qui auraient été évitées selon les procédures utilisées pour l'enquête SIPP courante.

Nous avons examiné les rapports produits par les intervieweurs suite à leurs visites aux ménages pour le 1er cycle du 1er essai préliminaire, visites qui étaient toutes supposées avoir été enregistrées et rendu un jugement subjectif pour décider si chacune de ces visites aurait été nécessaire selon les procédures de l'enquête SIPP courante ou s'il s'agissait d'une entrée en communication «additionnelle», requise seulement pour appliquer les nouvelles procédures. Aucune des premières visites, par exemple, n'était considérée une visite «additionnelle»; toutes les visites faites pour obtenir des documents manquants relatifs au revenu étaient des visites «additionnelles». Les données figurant dans le registre des visites pour le 1er essai préliminaire ne montrent pas un nombre déraisonnable de visites «additionnelles» aux ménages (les données pour le 2e essai préliminaire n'ont pas encore été analysées). Bien qu'il soit impossible de trouver le nombre exact de visites «additionnelles», on peut en évaluer une limite supérieure; nous estimons qu'au plus 14% de toutes les visites sur place lors du 1er cycle afin d'interviewer des ménages étaient des visites «additionnelles». Même si ces visites «additionnelles» (et les nombreux appels téléphoniques additionnels qui n'auraient pas été nécessaires selon les procédures de l'enquête SIPP courante) ont contribué, sans aucun doute, aux coûts plus élevés des travaux sur le terrain, il ne semble pas qu'il y en ait suffisamment pour justifier tout l'écart dans les coûts.

Un autre élément qui contribue aux coûts plus élevés des nouvelles procédures est la durée effective de l'interview réalisée sur place. Pour le 2e essai préliminaire de la SIPP-CR, une interview du 1er cycle prenait, en moyenne, 71 minutes par ménage; pour l'enquête SIPP courante, la moyenne est d'environ 52 minutes par ménage. Il se peut que cette différence soit attribuable au manque d'expérience de l'intervieweur (tous les intervieweurs travaillant à la SIPP-CR effectuaient ce travail pour la première fois), ou elle peut être due aux procédures; dans l'un ou l'autre cas, il est peu probable que cette situation ait eu un effet considérable sur les différences de coûts observées.

Le fait que les intervieweurs ont effectué de nombreuses visites improductives (Krasko 1992) est une cause majeure beaucoup plus évidente des coûts plus élevés du travail sur le terrain lors des essais préliminaires. On a manifestement évité de réaliser des interviews en soirée, les intervieweurs ont donc effectué, pendant la journée, des visites répétées qui n'ont pas permis d'entrer en communication avec des répondants éventuels. Puisque les coûts de déplacement sont une composante importante des coûts du travail sur le terrain, ces visites improductives ont certainement contribué aux coûts directs d'interview plus élevés. Il se peut que le manque d'expérience des intervieweurs à titre d'intervieweurs travaillant à des enquêtes, que le fait qu'ils n'habitaient pas dans les secteurs qui leur avaient été confiés et qu'on n'avait pas mis l'accent sur les coûts et sur l'efficience (lors de la formation, dans la supervision et la rétro-information), expliquent tous pourquoi les intervieweurs ont effectué un nombre si élevé de visites improductives.

6. CONCLUSIONS ET PROCHAINES ÉTAPES

Bien que le travail sur les procédures révisées et «cognitives» de l'enquête SIPP soit encore loin d'être terminé, les renseignements recueillis dans le cadre d'essais préliminaires à petite échelle nous amènent à penser que les nouvelles procédures offrent la possibilité de réduire considérablement certains des problèmes de mesure importants liés à l'enquête. En même temps, les difficultés opérationnelles rencontrées lors des essais préliminaires - taux de non-réponse et coûts élevés - menacent manifestement la notion que ces procédures constituent une option viable pour une mise en application effective à l'échelle nationale.

L'étude d'évaluation actuellement en cours - une comparaison expérimentale en parallèle des nouvelles procédures et des procédures de l'enquête SIPP courante, dans laquelle on utilise des données tirées des dossiers administratifs comme critères de mesure - permettra d'obtenir des preuves sérieuses à propos des avantages qu'offrent les procédures révisées relativement aux erreurs de mesure. Si les résultats pour les erreurs de mesure se révèlent suffisamment positifs, le Census Bureau effectuera des recherches additionnelles afin d'étudier les nombreuses questions opérationnelles qui resteront à régler, y compris, bien entendu, comment ramener les taux de non-réponse et les coûts à des valeurs raisonnables, mais aussi la possibilité d'appliquer les résultats à d'autres lieux, la collaboration des répondants pendant plusieurs cycles, la meilleure façon d'utiliser l'interview sur place ou l'interview téléphonique assistée par ordinateur avec les nouvelles procédures, les effets différentiels de la nouvelle méthode pour recueillir des données sur le revenu dans des sous-groupes (plus particulièrement les sous-groupes à revenu élevé) et de nombreuses autres questions.

Deux de ces «autres» questions méritent une mention spéciale. Une a trait au comportement des intervieweurs, que les nouvelles procédures mènent manifestement dans de nouvelles directions. Une interprétation raisonnable des résultats des essais préliminaires en ce qui a trait à la non-réponse, est que, bien que les répondants soient peu enclins à ne pas collaborer quand les nouvelles procédures sont employées, il est fort possible que les intervieweurs manifestent un certain manque d'enthousiasme pour ce qui est de leur application. Notre examen des résultats de l'étude d'évaluation doit tenir compte des perceptions des intervieweurs: que trouvent-ils particulièrement difficile, s'il y a lieu, à propos des nouvelles procédures et pourquoi? Nous devons aussi être préparés à accepter la possibilité que d'autres révisions et améliorations doivent être apportées aux nouvelles procédures pour les rendre vraiment réalisables, il se peut aussi que de nouvelles méthodes de formation en classe et de nouvelles approches pour encadrer et surveiller les intervieweurs sur le terrain soient nécessaires.

La deuxième question porte sur l'ensemble de procédures de l'enquête SIPP-CR elles-mêmes. L'ensemble actuel de procédures a été créé en tenant compte des échéances du calendrier de remaniement de l'enquête SIPP. On ne nous a pas accordé de temps pour élaborer et améliorer les composantes individuellement, mais nous avons plutôt dû utiliser une méthode indiscutablement «globale». Bien que nous puissions faire des hypothèses, nous ne savons pas quelle partie de l'ensemble de procédures permet d'obtenir les gains véritables de qualité et où les gains ne sont pas suffisants pour justifier la dépense additionnelle. Si des coûts accrus et des taux de non-réponse plus élevés continuent d'accompagner toutes les améliorations de la qualité des réponses, il sera essentiel que la prochaine étape des recherches s'intéresse au coût discret et aux effets sur la qualité de chacune des composantes de ce qui constitue actuellement l'ensemble servant à effectuer des mesures dans le cadre de l'enquête SIPP-CR.

BIBLIOGRAPHIE

Burkhead, D., et Coder, J. (1985). Gross changes in income recipiency from the survey of income and program participation. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington, DC, 351-356.

Krasko, N. (1992). SIPP-CR pretest I type A analysis. Note de service non publiée du U.S. Bureau of the Census à Stephen Willette, février 1992.

Marquis, K. (1990). Report of the SIPP cognitive interviewing project. Rapport non publié, U.S. Bureau of the Census, août 1990.

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

Marquis, K., et Moore, J. (1990). Measurement errors in SIPP program reports. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., 721-745. Aussi disponible en tant que Report No. 9008 de la Census Bureau's SIPP Working Paper Series (juin 1990).

Marquis, K., Moore, J., et Bogen, K. (1991). A cognitive approach to redesigning measurement in the survey of income and program participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 413-418.

Moore, J., Bogen, K., et Marquis, K. (1992). Use of unstandardized interviewing techniques in a proposed redesign of the survey of income and program participation. Document non publié pour les Joint Meetings of the Census Advisory Committees of the American Marketing Association and the American Statistical Association, U.S. Bureau of the Census, Washington, DC, 22 et 23 octobre, 1992.

Moore, J., et Marquis, K. (1989). Utilisation des données de dossiers administratifs pour l'évaluation de la qualité des estimations d'enquêtes, *Techniques d'enquête*, 15, 1, 133-149.

Murray, T. S., Michaud, S., Egan, M., et Lemaître, G. (1991). Invisible seams? The experience with the Canadian labour market activity survey. *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 715-758.

Singh, R. (1991). SIPP 91: Wave 1 results of the record check study. Note de service non publiée du U.S. Bureau of the Census au SIPP Research and Evaluation Steering Committee, 19 décembre, 1991.

Singh, R. (1992). SIPP 91: Wave 2 results of the record check study. Note de service non publiée du U.S. Bureau of the Census au SIPP Research and Evaluation Steering Committee, 15 juin, 1992.

Recueil du Symposium 1992 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

REPÉRAGE DES RÉPONDANTS À L'ENQUÊTE NLSY ET CONSERVATION D'UN TAUX D'ACHÈVEMENT DE 90% POUR 13 VAGUES ANNUELLES

A. Schoua-Glusberg et E. Hunt[1]

RÉSUMÉ

L'enquête «National Longitudinal Survey of Labor Market Experience - Youth Cohort (NLSY)» existe depuis 1979. Parrainée actuellement par le U.S. Bureau of Labor Statistics, cette enquête consiste à réaliser annuellement des interviews directes avec chacune des personnes qui avaient été interviewées en 1979 (environ 9,800), sauf les membres de deux des échantillons sur-représentés. À la vague numéro 13, 98% des répondants avaient pu être repérés et plus de 91% avaient pu être interviewés. Dans cette communication, on analyse les méthodes qui ont servi à retracer les participants de cette enquête longitudinale annuelle. Le repérage efficace des répondants dans une enquête par panel est une façon de réduire au maximum l'érosion de l'échantillon.

MOTS-CLÉS: Repérage; NLSY; enquêtes longitudinales.

1. INTRODUCTION

1.1 L'enquête «National Longitudinal Survey of Labor Market Experience»

Je vous entretiendrai aujourd'hui des stratégies et des méthodes que nous employons dans l'enquête «National Longitudinal Survey of Labor Market Experience/Youth Cohort (NLSY)» pour repérer les répondants. L'étude est parrainée par le Bureau of Labor Statistics du Department of Labor des États-Unis. Le Center for Human Resource Research (CHRR) de l'Ohio State University est le fournisseur attitré, tandis que le National Opinion Research Center (NORC) est le sous-traitant chargé de la collecte des données.

Le NLSY est une enquête annuelle menée auprès de jeunes adultes sur leur expérience du marché du travail. Pour l'année de référence de l'enquête (1979), un total de 12 686 personnes dont l'âge variait entre 14 et 21 ans ont été sélectionnées afin de participer à une enquête parrainée par le Department of Labor et le Department of Defense des États-Unis. L'enquête devait, à l'origine, durer cinq ans, mais on a décidé de la poursuivre. Elle est réalisée chaque année et compte maintenant 13 ans d'existence. D'ici les deux prochaines semaines, nous allons terminer la collecte des données pour la quatorzième vague, et nous sommes à préparer la quinzième.

1.2 Taux d'achèvement du NLSY

Au fil des ans, les taux d'achèvement de l'enquête NLSY ont varié entre 90% et 96% du nombre réel de répondants de l'année de référence. Il est devenu de plus en plus difficile avec les années d'atteindre des taux d'achèvement aussi élevés, bien que nous n'ayons pas nécessairement observé de diminution constante. Certaines années, le taux d'achèvement a baissé, puis a remonté par la suite. Il y a de bonnes raisons qui expliquent ce phénomène. Ainsi, après avoir participé aux cinq premières vagues, bon nombre de répondants estimaient avoir rempli leur engagement initial de cinq ans et ne voulaient pas continuer à être interviewés. Il en est résulté une chute marquée du taux d'achèvement après la cinquième année. Autre exemple: dans la neuvième vague, la majorité des interviews ont été menées par téléphone. Or, c'est la seule fois où nous avons procédé de la sorte, et le taux d'achèvement a alors diminué. Mais il est remonté de nouveau dans la dixième vague dès que nous

[1] A. Schoua-Glusberg et E. Hunt, National Opinion Research Center, 1155 E., 60e Rue, Chicago (IL), É.-U. 60637.

avons repris les interviews directes. Quoiqu'il en soit, nous avons toujours réussi à interviewer plus de 90% des répondants sélectionnés à l'origine.

Nous explorons constamment de nouvelles façons de convaincre les personnes réfractaires de participer à l'enquête; bien entendu, nous appliquons celles qui nous semblent les plus pratiques et les plus efficaces. Il n'y a pas de doute que l'aptitude à repérer les répondants au fil des ans constitue l'un de grands succès obtenus par l'enquête NLSY dans ses efforts visant à empêcher l'érosion de l'échantillon.

2. REPÉRER LES RÉPONDANTS DE L'ENQUÊTE NLSY

2.1 Catégories de répondants suivis

Quelles catégories de répondants suivons-nous? Si nous reprenons les «règles de suivi» dont Graham Kalton nous a entretenus ce matin, nous cherchons en somme à retracer tous les répondants, sauf, bien entendu, les personnes décédées et celles appartenant à la catégorie dite des «refus agressifs» (c'est-à-dire les répondants qui menacent d'appeler la police ou leur avocat, ceux qui accueillent nos intervieweurs à la pointe du fusil et les gens du même acabit). Il y en a très peu. Nous suivons les personnes qui vivent en établissement, du moment qu'elles sont capables de répondre à nos questions. Par exemple, nous interviewons chaque année un certain nombre de détenus. Nous suivons aussi les répondants qui vont s'établir à l'étranger; nous en interviewons certains au téléphone, d'autres en personne. Nous suivons, ou en tous cas nous essayons de suivre, les personnes dont le domicile est inconnu et nous tentons de joindre également celles qui n'ont pas répondu dans les vagues précédentes.

NORC a une longue expérience du repérage des répondants difficiles à trouver. En particulier, nous travaillons depuis longtemps à repérer des jeunes en milieu défavorisé, une population qu'on retrouve d'ailleurs souvent dans les enquêtes. Notre expérience en ce domaine laisse généralement supposer qu'une approche souple et globale, qui tient compte des caractéristiques des diverses composantes de l'échantillon, est essentielle lorsqu'on recherche des personnes difficiles à trouver.

Penchons-nous maintenant sur les personnes introuvables. Qui sont-elles? À quoi ressemblent-elles? Nous entendons par «introuvables» les répondants qui n'ont pas été retracés à la fin de la période d'enquête sur le terrain. (Les chiffres que je vais citer correspondent aux 12 premières années de l'enquête, étant donné qu'il s'agit là des données officielles publiées jusqu'à présent.) Précisons d'abord que parmi les 12 686 répondants constituant l'échantillon initial, seulement 1 190 ont été au moins une fois classés introuvables. De ceux-là, le quart ou 25% sont restés introuvables plusieurs fois et les trois quarts (75%) ne l'ont été qu'une seule fois, tandis que 81% l'ont été une fois ou deux.

Au fil des ans, nous avons cherché à interviewer des échantillons de tailles différentes. Pendant les premières années, nous avons cherché à interviewer la totalité des 12 686 répondants initiaux. Par la suite, après la sixième vague, nous avons éliminé l'échantillon militaire sur-représenté. L'échantillon global a ainsi été ramené à environ 11 600 répondants. Il y a deux ans, nous avons laissé tomber l'échantillon sur-représenté composé de personnes de race blanche économiquement défavorisées, c'est-à-dire l'échantillon des «Blancs défavorisés» (population d'ailleurs particulièrement facile à repérer et à interviewer). Nous nous intéressons ainsi à quelque 9 800 personnes. De ce nombre, jamais plus de 2,5% ont été classées comme introuvables au terme d'une vague donnée. En chiffres absolus, le nombre le plus élevé d'introuvables au terme d'une vague a été de 293.

2.2 Caractéristiques démographiques des introuvables

Examinons les caractéristiques démographiques des introuvables. Sur le plan du sexe, 60% sont des hommes et 40%, des femmes, alors que les personnes des deux sexes étaient représentées également dans l'échantillon. S'agissant de la race et du groupe ethnique, on constate que les hispanophones sont les plus difficiles à trouver - ce qui n'est pas étonnant. En effet, ils forment 30% des introuvables, en regard de 15% de l'échantillon. Dans notre échantillon sur-représenté composé d'hispanophones, nous observons beaucoup de mouvements de part et d'autre de la frontière entre le Mexique et les États-Unis, avec un va-et-vient fréquent de nombreux répondants. Il nous arrive de retrouver ces personnes au Mexique et de les interviewer là-bas. Les Noirs

constituent 28% des introuvables et 25% de l'échantillon, tandis que les Blancs constituent 41% des introuvables et 59% de l'échantillon.

2.3 Outils de repérage

De quelles sources de renseignements disposons-nous pour le repérage? Chaque année, nous terminons les interviews en indiquant dans la dernière partie du questionnaire tous les renseignements utiles et récents permettant de localiser les répondants. Nous leur demandons chaque fois d'inscrire leur nom au complet, car il arrive que le nom change. Nous leur demandons leur adresse complète et leur numéro de téléphone, leur lieu de travail et s'ils nous permettent de communiquer avec eux au travail. Nous demandons aux femmes leur nom de jeune fille. Nous demandons aux répondants leur numéro de permis de conduire, s'ils comptent déménager, les noms des parents et amis susceptibles de toujours savoir où ils se trouvent. Que faisons-nous de tous ces renseignements? Chaque année, au moment où nous nous apprêtons à enquêter sur le terrain, nous incorporons tout nouveau renseignement dans une Fiche de localisation remise à l'intervieweur pour chaque cas qu'on lui confie. Au fil des ans, nous avons dû diminuer la taille du caractère d'imprimerie de cette Fiche de localisation, étant donné que nous essayons de grouper sur une seule feuille tous les renseignements recueillis d'une année à l'autre. Nous ne nous limitons pas aux renseignements les plus récents ou les plus utiles. Nous gardons aussi après chaque vague les coordonnées des personnes de référence, habituellement des parents et amis, dont les noms nous ont été fournis par les répondants. Il y a quelques années, nous avions décidé de supprimer de vieilles références contenues dans la Fiche de localisation pour faire place à de nouveaux éléments et produire ainsi un document plus lisible. Cependant, les intervieweurs ont jugé que l'initiative n'avait pas été heureuse, parce qu'il arrive qu'une vieille référence s'avère utile quand toutes les autres pistes ne mènent nulle part.

2.4 Repérage avant l'enquête sur le terrain

Nous ne nous contentons pas de recueillir des renseignements que nous mettons à jour chaque année avec les répondants; nous suivons aussi d'autres pistes. D'une année à l'autre, nous perfectionnons et révisons constamment la marche à suivre en fonction des renseignements accumulés, en vue surtout de la consultation des bases de données. Permettez-moi de vous décrire notre démarche habituelle. Chaque année, avant même d'essayer de prendre contact avec un répondant, le personnel du bureau central vérifie les bases de données électroniques, par exemple les bases de données «National Change of Address». Même si on nous donne une nouvelle adresse postale, l'ancienne demeure sur la Fiche de localisation. L'expérience nous a enseigné que, parfois, les modifications ou les mises à jour dans les bases de données étaient incorrectes. Dans le passé, nous avons commis l'erreur de supprimer l'ancienne adresse pour la remplacer par la nouvelle, supposément meilleure. Maintenant, nous gardons toujours l'ancienne adresse. Cette vérification préliminaire est à peu près tout ce que nous faisons avant de prendre contact avec les répondants.

En mars, nous envoyons par la poste un premier avis auquel nous joignons un coupon sur lequel les répondants peuvent corriger ou modifier leur adresse existante. Les coupons retournés constituent le point de départ des actions ultérieures entreprises par le bureau central. Dans les cas où le coupon n'est pas retourné, nous supposons -- une conjecture certes hasardeuse -- que l'avis a été envoyé à la bonne adresse. Aucune autre action n'est donc entreprise en pareil cas avant l'enquête sur le terrain. Les coupons retournés peuvent être divisés en différentes catégories. Certains établissent que l'adresse postale était effectivement correcte. (Nous ne demandons pas aux répondants de nous retourner le coupon si l'adresse imprimée est correcte, mais il y en a qui prennent la peine de nous le retourner en précisant que les renseignements que nous avons sur eux sont toujours corrects.) D'autres nous font part d'un changement d'adresse ou de numéro de téléphone. Nous recevons aussi quelques coupons du bureau de poste avec une mention de nouvelle adresse: la lettre est livrée au répondant par le bureau de poste, lequel nous informe de la nouvelle adresse. Et enfin d'autres avis n'ont pu être livrés à cause d'une adresse incomplète ou erronée.

Examinons, à titre d'exemple, ce qui s'est produit durant la douzième vague, année typique sur le plan du repérage. Si nous excluons les «refus agressifs», nous avions mis à la poste plus de 10 500 avis. Dans environ la moitié des cas, nous n'avions rien reçu en retour. Si on additionne tous les coupons que le bureau de poste ou les répondants nous ont retournés avec des corrections et les avis qui n'ont pu être livrés, le tiers des avis mis à la poste comportaient un ou plusieurs changements. On voit donc que, d'une année à l'autre, beaucoup de répondants changent d'adresse. Bien entendu, il est arrivé que des répondants nous signalent une correction due

à des erreurs produites à l'étape de la saisie. À la fin du premier envoi à la douzième vague, nous nous sommes donc retrouvés avec 783 répondants dont nous ne connaissions pas l'adresse et que nous devions retracer d'une façon ou d'une autre. Ce qui veut dire que nous avons pu retracer environ les deux tiers des répondants dont l'avis n'a pu être livré. On notera que cette proportion reste à peu près la même d'une vague à l'autre.

2.5 Repérage durant la collecte des données

Une autre façon de considérer le problème des répondants introuvables est de surveiller, d'une semaine à l'autre, l'évolution des cas durant la période d'enquête sur le terrain, laquelle commence à la fin de mai et se termine à la fin d'octobre. Au pire moment, entre 5% et 7% des répondants de l'échantillon sont introuvables. À la fin de la collecte des données, nous sommes en général capables d'en retracer les deux tiers. Ainsi, au terme de la collecte des données de la douzième vague, seulement 246 répondants sont restés introuvables.

Une fois que les résultats du premier envoi postal sont totalisés et que les adresses sont révisées, on procède à l'enquête sur le terrain. Les intervieweurs passent par plusieurs étapes pour retracer les cas qui leur sont confiés. Voulant entrer en contact par téléphone avec certains répondants, ils se heurtent à toutes sortes de difficultés, notamment des lignes débranchées. Ils se rendent à la dernière adresse connue du répondant sans arriver à le retracer, ou encore ils tombent sur un cas de répondant introuvable que les avis n'avaient pas décelé.

Les intervieweurs ont recours à un certain nombre d'outils connus: assistance-annuaire, répertoires téléphoniques, registres des statistiques de l'état civil, annuaires croisés et annuaires d'adresses permettant de prendre contact avec les voisins, changements d'adresse au bureau de poste local ou vérifications des anciennes adresses, vérifications au Department of Motor Vehicle des renseignements sur le permis de conduire que nous recueillons chaque année, recensements des électeurs, registres des répartiteurs (impôts). Voilà donc autant de méthodes courantes qu'adoptent les intervieweurs. Leurs superviseurs, de leur côté, peuvent consulter des bases de données informatisées à cette étape, par exemple celles des agences d'évaluation du crédit et des compagnies de téléphone, comme quelqu'un le mentionnait ce matin.

Quelques répondants ne peuvent être retracés par les moyens que je viens de décrire; il faut alors jouer au détective dans l'entourage du dernier domicile ou suivre d'autres pistes que nous aurions pu obtenir dans les années antérieures. L'intervieweur commence son travail de repérage en utilisant certains outils propres au cas qui l'occupe et recueille le plus de renseignements possibles dans le secteur géographique de l'ancienne adresse. Il parle aux voisins que le répondant aurait pu donner comme références ou s'entretient avec les commerçants du secteur. Nous n'insistons pas sur l'emploi de ces méthodes, mais elles sont parfois tout ce qui nous reste et donnent de bons résultats.

À quoi attribuons-nous le succès de nos efforts de repérage? Nous le devons en grande partie à la persistance des intervieweurs et à leur connaissance de l'échantillon. Les parents des répondants font preuve d'une bonne collaboration. Comme l'âge des jeunes variait entre 14 et 21 ans au moment où nous avons commencé à les interviewer, nous avons eu de fréquents contacts avec les parents au fil des ans. En général, les intervieweurs ont une attitude très positive à l'égard de l'échantillon, et c'est une question d'honneur pour eux de retrouver les jeunes visés par l'enquête.

De nombreuses anecdotes nous ont été racontées. Ainsi, un intervieweur a parcouru les champs de laitue de la Californie à la recherche de répondants qui manquaient à l'appel et qu'il a fini par retrouver. Dans les villes, il y a cette intervieweuse qui s'est rappelé que tel répondant aimait le jazz. Elle a visité tous les magasins de musique du quartier jusqu'à ce qu'elle tombe sur quelqu'un qui l'a mise sur la bonne piste. Une autre intervieweuse a découvert que le répondant qu'elle cherchait était devenu un sans-abri et habitait dans un boisé. Elle a parcouru le boisé à pied en appelant à haute voix le répondant jusqu'à ce qu'elle le trouve. Voilà autant d'exemples de persistance et d'ingéniosité.

3. CONCLUSION

Le coût de nos efforts de repérage doit être pris en compte. Bien que nous ne tenions pas un compte distinct de nos activités de repérage, contrairement à d'autres activités propres aux interviews, nous pouvons affirmer que le rapport entre la durée de l'interview et le nombre total d'heures consacrées à un cas dans le cadre de l'enquête NLSY n'est pas beaucoup plus élevé que pour bien d'autres de nos enquêtes à petit budget. Par conséquent, nous estimons que nous ne passons pas beaucoup de temps à faire du repérage. Très peu de cas nous obligent à sortir tout l'arsenal. Si on répartit ces activités de repérage sur un échantillon aussi vaste, on constate que le coût par cas n'est pas élevé.

Un dernier point: il ne faut pas abandonner la recherche des personnes introuvables. Nous avons noté qu'à la fin de la douzième vague, sur 246 personnes classées introuvables, seulement 60 n'avaient pu être retracées dans les deux vagues précédentes. Voyant comment les répondants deviennent des introuvables et comment nous les retraçons, nous constatons qu'ils essaient plutôt de nous éviter. D'autre part, lorsque nous les retrouvons, la majorité de ces personnes (environ 75%) deviennent par la suite de véritables répondants et cessent de jouer à cache-cache. La solution consiste donc à toujours poursuivre les recherches, car, dès que nous retraçons les introuvables, ils sont habituellement fidèles.

SESSION 2

Sélection des échantillons et pondération

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

TIRAGE COORDONNÉ D'ÉCHANTILLONS STRATIFIÉS

F. Cotton et C. Hesse[1]

RÉSUMÉ

Ce papier décrit un système de tirage coordonné d'échantillons stratifiés utilisé à l'INSEE depuis quelques années pour les enquêtes annuelles d'entreprises. Ce système repose sur l'affectation aux unités de la base de sondage de numéros aléatoires variables, recalculés après chaque tirage selon une formule simple qui dépend du type de coordination désiré. On peut avec cette méthode coordonner simplement des tirages stratifiés suivant des critères différents, ou même portant sur des unités différentes mais liées: un exemple de ce dernier type d'application, portant sur des entreprises et leurs établissements, est décrit en détail, et quelques simulations sont présentées pour donner un aperçu des effets de coordination obtenus.

MOTS CLÉS: Coordination; échantillons stratifiés; numéros aléatoires; enquêtes d'entreprises.

1. INTRODUCTION

L'enquête annuelle d'entreprises réalisée par le système statistique d'entreprises français est en fait une juxtaposition de plusieurs enquêtes, menées conjointement par divers organismes:

- des services statistiques de ministères (enquêtes sur l'industrie, les industries agricoles et alimentaires, le bâtiment, les transports);
- deux départements de l'INSEE (enquêtes dans le commerce et les services, enquête sur les très petites entreprises industrielles).

Les entreprises enquêtées sont sélectionnées par tirage aléatoire simple stratifié (TASST). Au-dessus d'un seuil de taille qui varie selon le secteur d'activité, les enquêtes sont exhaustives. La gestion de la base de sondage et le tirage des échantillons sont effectués à l'INSEE (Institut National de la Statistique et des Études Économiques), qui dispose du répertoire central des entreprises et des établissements (Sirène) à partir duquel est obtenue la base de sondage.

Ce système d'échantillonnage a fait l'objet d'une refonte informatique et méthodologique en 1989. À cette occasion, de nouvelles techniques statistiques ont été adoptées, en particulier une technique de tirage et de coordination d'échantillons par attribution de numéros aléatoires, technique qui fait l'objet de ce papier.

2. TIRAGES AVEC ATTRIBUTION DE NUMÉROS ALÉATOIRES

On considère une population U composée d'individus i, $i = 1,...,N$. Cette population est partitionnée en strates $\{U_h\}_{h=1,...,H}$ en fonction d'un critère connu pour tous les individus. L'objectif est d'effectuer dans cette population des tirages aléatoires simples stratifiés (TASST), définis ici comme des juxtapositions de tirages aléatoires simples indépendants dans chaque strate.

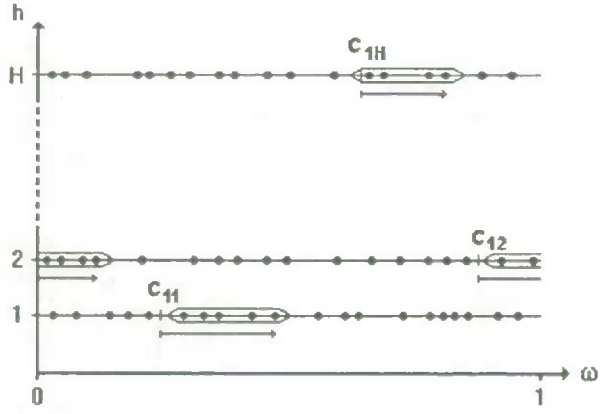
Pour commencer, on ne considérera que le tirage de deux échantillons s_1 et s_2, sans changement intermédiaire de la population ni de la stratification. Notons n_{1h} (respectivement n_{2h}) le nombre d'unités que l'on souhaite sélectionner dans la strate h pour s_1 (respectivement s_2).

---

[1] C. Hesse est responsable de la cellule méthodologie à la Direction des Statistiques Economiques de l'INSEE (France). F. Cotton travaile au Département des Projets de l'INSEE, 15, boulevard Gabriel Péri, F-92245 Malakoff Cedex.

On attribue indépendamment à chaque individu i un numéro ω_i tiré selon une loi uniforme sur $[0,1[$. $(\omega_1,...,\omega_N)$ suit donc une loi uniforme sur $\Omega = [0,1[^N$. Une manière simple de procéder au tirage de s_1 (figure 1) consiste à ordonner dans chaque strate les individus selon les ω_i croissants, à se fixer pour chaque strate, indépendamment des ω_i, une origine $c_{1h} \in [0,1[$, puis à sélectionner dans la strate h les n_{1h} premiers individus dont le numéro aléatoire vérifie $\omega_i \geq c_{1h}$. On raisonnera modulo 1 partout dans ce papier, c'est-à-dire en l'occurrence que si l'on arrive à la fin de la strate avant d'avoir sélectionné n_{1h} unités, on retourne vers l'origine des ω_i, et donc on sélectionne en complément les premiers individus tels que $\omega_i + 1 \geq c_{1h}$. Une telle séquence d'unités rangées selon les ω_i sera appelée une fenêtre d'interrogation.

Figure 1: Tirage stratifié par attribution de numéros aléatoires.



Chaque ligne horizontale représente une strate, et chaque point une unité, les unités étant rangées selon leurs numéros aléatoires. Les unités entourées sont sélectionnées lorsque l'on applique la technique décrite ci-dessus.

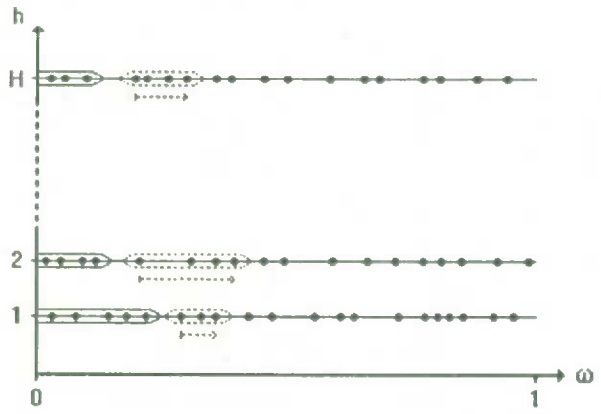
Il est évidemment possible de sélectionner «vers la gauche» de c_{1h}, c'est-à-dire dans le sens des ω_i décroissants: on prend alors les n_{1h} derniers individus tels que $\omega_i < c_{1h}$ (toujours modulo 1). Sauf mention explicite du contraire, on considérera ici que les échantillons sont tirés «vers la droite».

On procédera pour s_2 comme pour s_1, en sélectionnant les unités à partir d'un ensemble d'origines $\{c_{2h}\}_{h=1,...,H}$. On peut obtenir des effets de coordination entre s_1 et s_2 en jouant sur les c_{1h} et les c_{2h}, ou sur le sens de la sélection (vers la gauche ou vers la droite). Supposons, pour illustrer ce point, que les c_{1h} sont tous nuls et que s_1 est tiré vers la droite: cela n'entraîne aucune perte de généralité.

On peut coordonner s_1 et s_2 de façon positive, c'est-à-dire en reprenant dans s_2 le maximum d'unités déjà tirées par s_1, en choisissant $c_{2h} = 0$ pour tout h. Il n'est d'ailleurs dans ce cas pas nécessaire que la stratification de s_1 et celle de s_2 soient identiques.

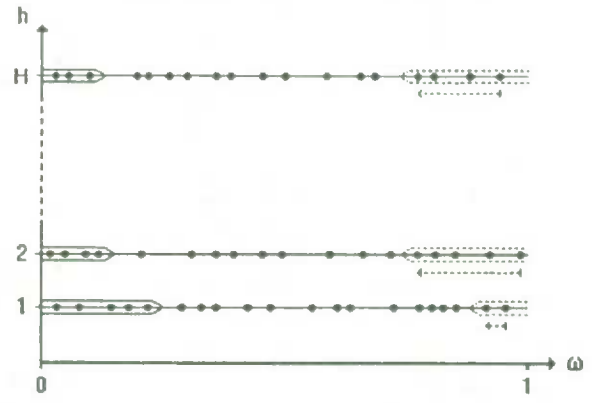
Pour obtenir une coordination négative, il faut décaler les fenêtres d'interrogation entre le tirage de s_1 et celui de s_2. Les figures 2 et 3 décrivent deux méthodes possibles. On voit que celle de la figure 2 résiste mal à un changement de stratification entre s_1 et s_2, puisque l'on «désynchronise» les origines des fenêtres d'interrogation. Dans l'exemple de la figure, si la dernière unité sélectionnée pour s_1 dans la strate 1 passe dans la strate 2 avant le tirage de s_2, elle sera de nouveau sélectionnée pour s_2: la coordination négative ne sera pas maximale. Au contraire, le choix, dans chaque strate, de la fenêtre constituée des n_{2h} dernières unités de la strate conduit à une coordination négative maximale entre les deux échantillons. Cette situation est illustrée dans la figure 3, où le second échantillon est tiré avec des fenêtres de mêmes origines que le premier (nulles, en l'occurrence), mais dans le sens des ω_i décroissants.

Figure 2: Coordination négative de deux échantillons.



Le premier échantillon étant tiré avec des origines nulles (traits pleins), on peut coordonner négativement un second échantillon en utilisant des fenêtres d'interrogation dont l'origine se situe après la dernière unité sélectionnée lors du premier tirage (pointillés).

Figure 3: Coordination négative de deux échantillons.



Le premier échantillon étant tiré avec des origines nulles (traits pleins), on peut coordonner négativement un second échantillon en utilisant des fenêtres d'interrogation situées en fin de strates (pointillés).

3. RENUMÉROTATION DES UNITÉS

Considérons maintenant deux TASST pouvant adopter des stratifications différentes. Lorsque l'on recherche une coordination négative, on peut, plutôt que de changer la fenêtre d'interrogation entre les deux tirages, garder la fenêtre fixe mais «bouger» les unités en changeant leurs numéros aléatoires ω_i.

Notons $\alpha_{h1}(j)$, $j \in \{1,...,N_h\}$ le j-ième numéro aléatoire, par ordre croissant, dans la strate h du premier tirage: c'est le numéro aléatoire ω_i d'un individu i, où i peut être spécifié en fonction de j et h. On a, les numéros aléatoires étant presque sûrement tous différents:

$$\alpha_{h1}(1) < ... < \alpha_{h1}(N_h).$$

On a sélectionné pour s_1 les individus tels que $j = 1,...,n_{1h}$, et on cherche un procédé de renumérotation:

$$\nu_{hl}:[0,1[^h \rightarrow [0,1[^h$$
$$\alpha_{hl} \rightarrow \beta_{hl}$$

associant à l'individu i de numéro $\alpha_{hl}(j)$ un nouveau numéro $\beta_{hl}(j)$, de telle façon que le classement selon ce nouveau numéro rejette les individus tirés par s_1 en fin de strate. Le choix de l'INSEE s'est porté sur la transformation ν_{hl}^1:

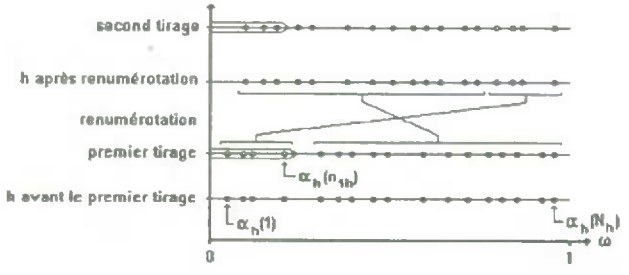
$$\begin{cases} \beta_{hl}(j) = \alpha_{hl}(j) + \alpha_{hl}(N_h) - \alpha_{hl}(n_{hl}) & j \le n_{hl} \\ \beta_{hl}(j) = \alpha_{hl}(j) - \alpha_{hl}(n_{hl}) & j > n_{hl}. \end{cases} \quad (1)$$

D'autres spécifications sont possibles, notamment la transformation ν_{hl}^p par permutation des numéros:

$$\begin{cases} \beta_{hl}(j) = \alpha_{hl}(j + N_h - n_{hl}) & j \le n_{hl} \\ \beta_{hl}(j) = \alpha_{hl}(j - n_{hl}) & j > n_{hl}. \end{cases} \quad (2)$$

L'échantillon s_2 s'obtient en sélectionnant les unités dont les nouveaux numéros sont au début des strates de s_2. La figure 4 fournit une illustration de cette méthode de coordination avec renumérotation.

Figure 4: Renumérotation des unités.



Ce graphique illustre l'évolution dans le temps de la répartition des unités dans une strate h. Les unités sélectionnées dans le premier échantillon (en bas) sont rejetées en fin de strate par la renumérotation, tandis que celles qui n'ont pas été tirées sont décalées vers le début de la strate, et sont donc sélectionnées prioritairement lors du second tirage (en haut).

Notons qu'il est possible, par renumérotation, de ne rejeter en fin de strate qu'une partie des individus sélectionnés pour le premier échantillon: il suffit de remplacer, dans la formule (1), n_{1h} par un entier plus petit. Cela permet d'obtenir des effets de coordination partielle entre les échantillons.

La transformation ν_{hl}^1 possède de bonnes propriétés[2]. En particulier, elle ne modifie pas la loi de probabilité sur Ω. De plus, le tirage joint $p(s_1, s_2)$ selon cette méthode est identique à celui qui consiste, sans changer les numéros, à utiliser pour s_1 les fenêtres situées au début des strates (de s_1), et pour s_2 les fenêtres situées à la fin des strates (de s_2). On obtient donc la coordination négative maximale des deux TASST.

La méthode de tirage avec renumérotation se généralise aisément au cas où le nombre d'échantillons à tirer dépasse 2. Elle permet en particulier la rotation maximale des unités dans des TASST successifs, avec changements de la stratification. Il suffit d'enchaîner les séquences:

---

[2] Voir Tirages coordonnés d'échantillons, F. Cotton, C. Hesse, document de travail INSEE n° E9206.

- tirage (la première fois) ou renumérotation des ω_i;
- restratification de la population;
- échantillonnage des unités au début des strates.

Dans le cas où le nombre d'échantillons est supérieur à 2, la renumérotation prend l'avantage sur la technique de sélection de s_2 en fin de strate de la figure 3: avec cette dernière, on ne sait en effet pas trop quelles fenêtres utiliser pour s_3 afin d'assurer une coordination négative avec à la fois s_1 et s_2.

4. APPLICATION: TIRAGE D'UNITÉS DE NIVEAUX DIFFÉRENTS

On peut réaliser de multiples effets de coordination entre échantillons par simple manipulation des numéros aléatoires associés aux unités, à condition de ne pas modifier la distribution uniforme de ces numéros.

Illustrons ce point dans une autre situation: considérons une base de sondage contenant deux types liés d'unités, par exemple des entreprises et leurs établissements. Il peut parfois être intéressant de coordonner des tirages portant sur ces deux types d'unités, notamment dans une optique de répartition de la charge statistique sur les enquêtés.

Supposons qu'un tirage p_1 a été effectué sur les entreprises en utilisant la méthode de renumérotation linéaire. On souhaite maintenant procéder à un tirage p_2 sur les établissements, mais en évitant si possible de sélectionner des établissements appartenant à des entreprises tirées par p_1. De telles entreprises ont, après renumérotation des ω_i plutôt élevés. Il faut que leurs établissements aient également des numéros aléatoires élevés pour ne pas être sélectionnés au début de leur strate par p_2. L'idée est donc de relier le numéro de l'entreprise et le plus petit des numéros de ses établissements.

Si au contraire p_1 portait sur les établissements, on peut chercher, dans un tirage p_2 d'entreprises, à éviter celles dont des établissement ont été sélectionnés par p_1. On souhaite en conséquence, lorsque l'un des numéros d'établissements est élevé, que le numéro de l'entreprise le soit aussi, ce qui fait donc penser à un lien entre ce dernier et le plus grand des numéros des établissements de l'entreprise.

Ce lien doit préserver l'uniformité des distributions des numéros ω_i d'entreprises et des numéros ω_{ij} d'établissements. À cette fin, on utilise la propriété que si $F(x)$ est la fonction de répartition d'une variable aléatoire X, $F(X)$ suit la loi uniforme sur $[0,1[$. Soit n le nombre d'établissements de l'entreprise i (pour être tout à fait correct, il faudrait indicer n par i, le nombre d'établissements pouvant varier selon les entreprises).

Dans le sens établissements → entreprise, pour calculer initialement les numéros d'entreprises à partir de ceux des établissements, ou pour répercuter sur les entreprises une renumérotation des établissements, on peut poser:

$$\omega_i = [\max(\omega_{i1},...,\omega_{in})]^n, \text{ ou :} \tag{3}$$

$$\omega_i = 1 - (1 - \min(\omega_{i1},...,\omega_{in}))^n. \tag{4}$$

On parlera respectivement de lien par le max et de lien par le min.

On utilisera le sens entreprise → établissements pour répercuter par exemple une renumérotation des entreprises sur les établissements. On peut ainsi maintenir le lien par le min en attribuant à l'établissement ayant initialement le plus petit numéro la nouvelle valeur $1-(1-\omega_i)^{1/n}$, dont la loi est celle du minimum de n réels tirés selon une loi uniforme dans $[0,1[$.

Pour attribuer les nouveaux numéros des autres établissements de l'entreprise, on utilise la propriété que, conditionnellement au minimum des ω_{ij}, les ω_{ij} supérieurs doivent suivre également une loi uniforme entre ce minimum et 1, pour que leur loi marginale soit uniforme. On peut donc, soit tirer ces numéros selon une loi

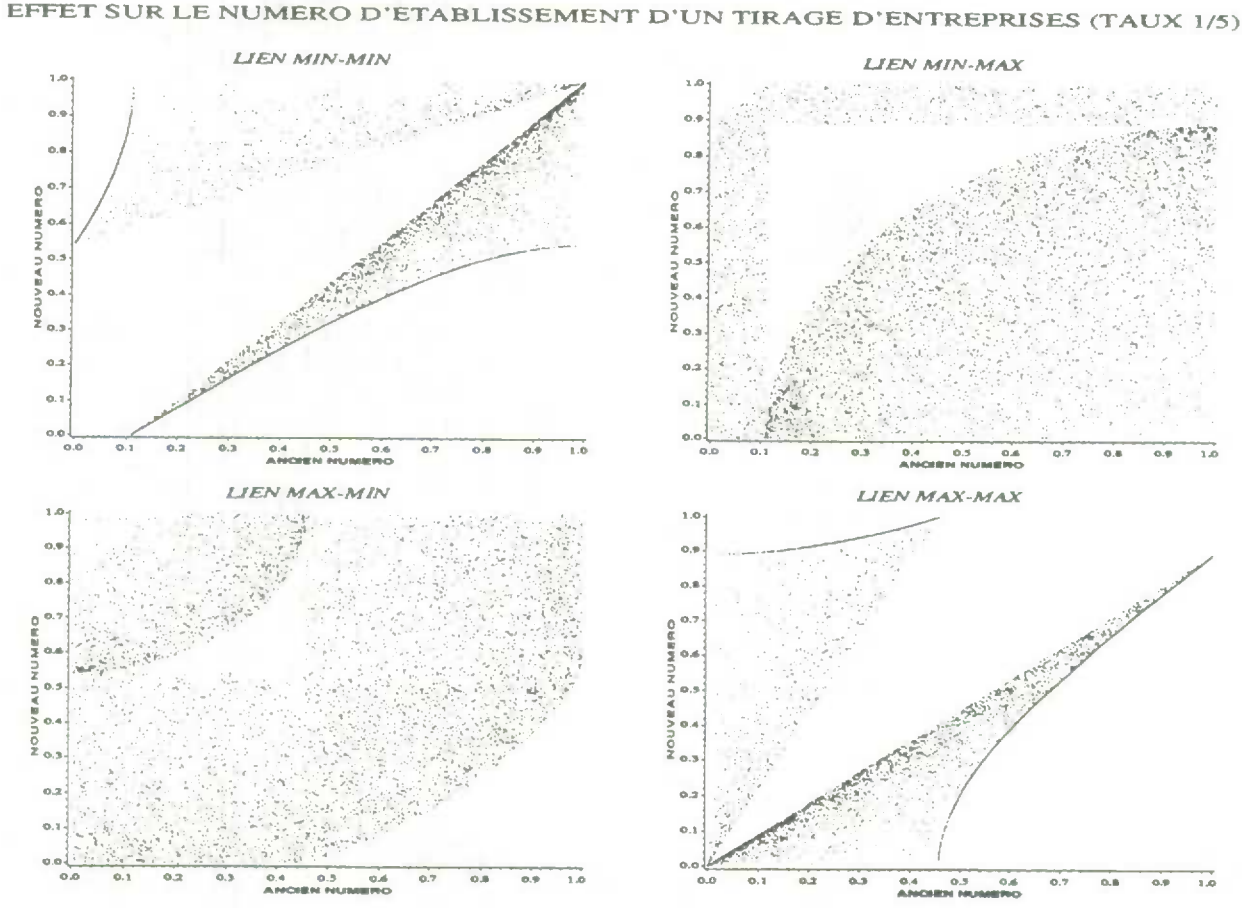
uniforme entre ce minimum et 1, soit les répartir dans cet intervalle avec des écarts proportionnels à ceux de leur distribution précédente (supposée être déjà uniforme).

Le lien entreprise → établissements par le max se fait en attribuant à l'établissement ayant initialement le plus grand numéro le nouveau numéro $\omega_i^{1/n}$, et en procédant de manière symétrique à ce qui précède pour le calcul des nouveaux numéros des autres établissements.

Les figures 5 et 6 présentent des simulations dans des cas simples, qui permettent de visualiser les effets de coordination obtenus.

Dans la figure 5, on considère une population non stratifiée de 10 000 entreprises de deux établissements. Les établissements se voient affecter un numéro tiré au hasard entre 0 et 1. Le numéro aléatoire de l'entreprise est calculé en fonction des numéros d'établissements par une des formules ci-dessus. On procède ensuite à un sondage aléatoire simple sur les entreprises puis à une renumérotation sur ces entreprises. Le numéro des établissements est ensuite recalculé en utilisant là encore un lien par le min ou par le max.

Figure 5: Simulations sur 5 000 entreprises de deux établissements.



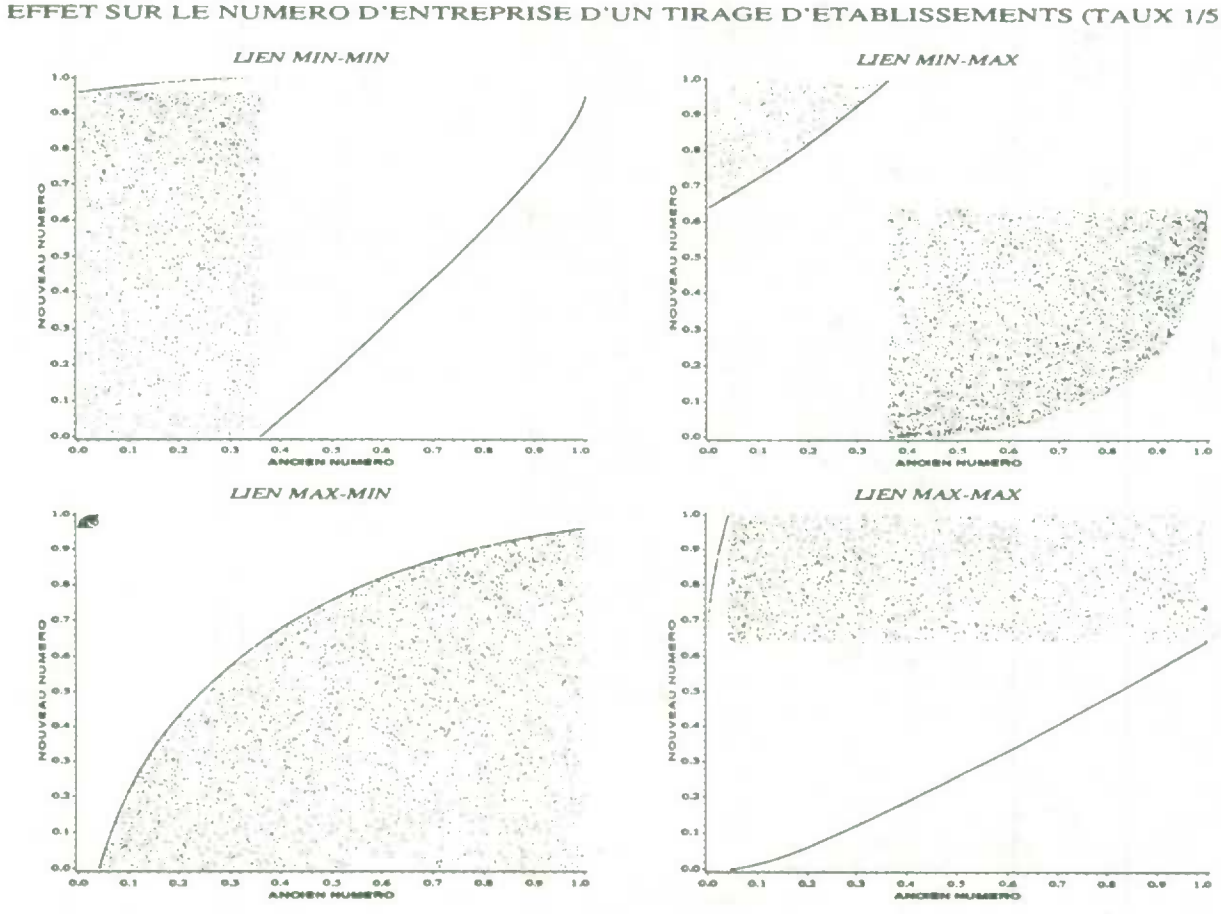
On a tracé sur la figure la correspondance entre l'ancien numéro des établissements et le nouveau pour chacun des liens possibles (min-min, min-max, max-min et max-max). Chaque établissement est représenté par un point d'abscisse son ancien numéro et d'ordonnée son nouveau. Les quatre figures sont composées de deux ensembles de points: le premier, situé plutôt en haut et à gauche, correspond aux établissements dont l'entreprise à été tirée; le second, plutôt en bas et à droite, aux établissements dont l'entreprise n'a pas été tirée. La distinction est particulièrement nette dans le cas des liens min-min et max-max. On constate que les établissements dont l'entreprise a été tirée voient leur numéro augmenter, c'est-à-dire leur probabilité de sélection dans le prochain tirage d'établissement diminuer. L'effet est inversé pour les établissements dont l'entreprise n'a pas été tirée.

Le choix du lien à utiliser dépend de l'effet de coordination que l'on désire. Dans le cas du lien min-min, tous les établissements dont l'entreprise a été tirée ont un nouveau numéro élevé (il est aisé de calculer qu'il est supérieur à $1 - \sqrt{\tau}$, où τ est le taux de sondage, 1/5 dans notre exemple), ce qui assure une bonne coordination négative avec le prochain tirage. En revanche, la coordination avec un éventuel tirage précédent sur les établissement est moins bonne, puisque certains établissements d'abscisse élevée voient leur entreprise sélectionnée. Inversement, avec le lien max-max, la coordination négative est bonne avec le tirage précédent, mais moins bonne avec le tirage suivant, puisque certains établissements dont l'entreprise est tirée ont un nouveau numéro proche de l'origine. Le lien max-min assure une bonne coordination avec à la fois le tirage suivant et le tirage précédent, mais présente d'autres désavantages, en particulier celui de modifier de façon plus importante les numéros des établissements dont l'entreprise n'est pas tirée.

Dans la figure 6, sur la même population que précédemment, on simule cette fois un tirage d'établissements, qui modifie donc par renumérotation les numéros de ces établissements. Chaque point représente une entreprise: son abscisse est le numéro initial de l'entreprise, calculé à partir des numéros de ses établissement avant le tirage, et son ordonnée est le numéro de l'entreprise recalculé sur les numéros d'établissements après tirage. Chaque image est constituée de trois zones, qui sont alternativement des courbes et des nuages de points, correspondant aux entreprises dont les deux établissements ont été tirés (en haut à gauche), à celles dont un seul établissement a été tiré, et enfin à celles dont aucun établissement n'a été tiré (en bas à droite).

On note sur la figure que les entreprises dont les deux établissements ont été tirés subissent une forte augmentation de leur numéro, et donc leur probabilité de sélection ultérieure est largement diminuée. Le même effet, moins fort, est observé sur les entreprises dont un seul établissement est tiré. Les autres entreprises voient leur numéro diminuer. Là encore, on peut comparer les avantages et les inconvénients des différents liens possibles, le meilleur compromis semblant cette fois être le lien min-max.

Figure 6: Simulations sur 5 000 entreprises de deux établissements.



Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

SÉLECTION ET MISE À JOUR D'UN ÉCHANTILLON PERMANENT HAUTEMENT STRATIFIÉ

J.L. Czajka et A.L. Schirm[1]

RÉSUMÉ

L'Internal Revenue Service des É.-U. a prélevé 90 000 déclarations d'impôt sur le revenu des particuliers parmi son échantillon de 1987 pour constituer un panel annuel. Cet échantillon est fortement stratifié selon le revenu, les taux de sondage variant de 0,02% à 100%. Les données des trois premières années permettent de constater que les mouvements interstrates ont modifié sensiblement la composition de l'échantillon. Dans cette communication, on étudie les questions que soulève le comportement dynamique d'un échantillon permanent hautement stratifié, notamment les conséquences pour l'échantillonnage, l'élaboration de poids pour l'estimation transversale et longitudinale, les méthodes d'élargissement de l'échantillon et les effets à long terme de ce comportement dynamique sur la précision des estimations calculées à partir de l'échantillon.

MOTS CLÉS: Revenu; stratification a posteriori; pondération.

1. INTRODUCTION

Cette communication examine un problème dont on a parlé très peu jusqu'à maintenant dans la littérature statistique et qui a trait à la conception et à l'analyse de panels ou d'échantillons permanents. Ce problème vient de l'emploi de la stratification dans le prélèvement d'un panel, plus précisément de la combinaison de deux facteurs:

- stratification selon des caractéristiques (par ex.: le revenu) qui varient dans le temps;
- probabilités d'échantillonnage très variables.

Dans ces circonstances, la composition d'un panel n'est pas fixe par rapport aux variables de stratification. Le comportement dynamique du panel influe sur la qualité des estimations calculées à partir des données de panel et attribue à la conception de panel une complexité que n'ont pas les échantillons transversaux.

Nous avons éprouvé ce problème en utilisant un échantillon permanent de déclarations de revenus des particuliers des É.-U. (par opposition aux déclarations de revenus des sociétés) conçu par la division des statistiques du revenu (Statistics of Income Division - SOI) de l'Internal Revenue Service (IRS) des É.-U. Cet échantillon, formé de 90 000 unités de déclaration, est très fortement stratifié, l'intervalle des taux de sondage couvrant quatre ordres de grandeur. En nous servant de cet échantillon comme exemple, nous allons examiner quelques-unes des conséquences d'une modification de la composition de l'échantillon permanent et voir comment cela peut influer sur l'élaboration du plan d'échantillonnage.

Le plan de cette communication est le suivant. Dans la section 2, nous énumérons divers usages des données fiscales provenant de panels et nous décrivons le panel de déclarations de revenus de particuliers de la SOI. La section 3 porte sur les changements qui peuvent survenir au fil des ans dans la composition et les caractéristiques du panel. Dans la section 4, nous expliquons comment ces changements peuvent influer sur la précision des estimations calculées à partir de l'échantillon. Ensuite, nous examinons deux méthodes générales - repondération et augmentation de l'échantillon - pour compenser les effets d'une modification de la composition de

[1] J.L. Czajka et A.L. Schirm, Mathematica Policy Research Inc., 600 ave. Maryland, S.-O., pièce 550, Washington (DC), É.-U. 20024-2512.

l'échantillon. Dans la section 6, nous étudions des méthodes de conception de panel qui situent l'échantillonnage dans une perspective longitudinale. Enfin, la section 7 renferme nos conclusions.

2. PANEL DE LA STATISTICS OF INCOME DIVISION (SOI)

Pour évaluer toutes les conséquences que peut avoir une modification de la composition du panel pour la conception et l'analyse d'enquêtes longitudinales, il est important de savoir non seulement comment sont recueillies les données de panel, mais aussi à quels usages elles sont destinées. Pour les besoins de la communication, nous énumérons ici des usages possibles des données fiscales provenant de panels, puis nous décrivons comment est conçu le panel de la SOI.

2.1 Usages des données fiscales provenant de panels

Les données de panel tirées de déclarations de revenus peuvent servir à de nombreux usages auxquels les données transversales ne conviennent pas. Voici des exemples de questions auxquelles on peut tenter de répondre avec des données fiscales de panel.

- De quelle manière l'éclatement de la famille influe-t-il sur le revenu et le statut de déclarant des anciens membres?
- Quel est l'aspect de la courbe de revenu à différentes étapes du cycle de vie? Et comment cet aspect varie-t-il selon les grandes catégories de revenu?
- Comment le revenu relatif de différents segments de population a-t-il évolué pendant une certaine période?
- Quelle est l'évolution à long terme de la réalisation de gains en capital?
- Dans quel délai les contribuables peuvent-ils utiliser les pertes reportées à des exercices ultérieurs?
- Comment les dons de charité varient-ils en fonction du revenu ou du taux d'imposition marginal?
- Comment les contribuables se prévalent-ils du crédit d'impôt accordé pour une perte nette d'exploitation: l'utilisent-ils pour des années d'imposition antérieures ou ultérieures?
- Dans quelles conditions deux personnes formant un couple choisissent-elles de produire chacune une déclaration de revenus?

Certaines applications consistent dans l'analyse de phénomènes qui sont longitudinaux par essence (ex.: variation du revenu durant le cycle de vie). D'autres applications supposent des études de comportement qui peuvent être faites à l'aide de données transversales mais pour lesquelles des données longitudinales sont beaucoup plus appropriées (ex.: réponse à une modification des taux d'imposition). D'autres applications encore consistent dans l'étude d'événements transversaux (ex.: l'utilisation d'un crédit d'impôt particulier) pour lesquels on peut trouver des variables explicatives clés dans les années antérieures.

Évidemment, les données de panel peuvent servir tout aussi bien à des applications de nature transversale, surtout lorsque d'importants éléments d'information sont recueillis uniquement auprès d'un échantillon permanent. Bien que ce ne soit pas le cas actuellement du panel de la SOI, car il existe un échantillon transversal annuel de taille comparable, d'autres panels ont été enrichis de données qui ne proviennent pas de l'échantillon transversal, et on projette de faire la même chose pour le panel de la SOI.

2.2 Caractéristiques de l'échantillon permanent de la SOI

À chaque année, la division des statistiques du revenu (SOI) prélève un échantillon de déclarations de revenus des particuliers dans la population des déclarations qui sont traitées cette année-là. L'échantillon aléatoire

stratifié est très gros, son effectif dépassant souvent 100 000 déclarations. Avec l'année d'imposition 1987, la SOI a constitué un panel de près de 90 000 unités de déclaration. L'échantillon de l'année de référence a été prélevé dans l'échantillon transversal de 1987 et il est représentatif des déclarations de revenus de personnes non à charge, traitées par l'IRS en 1988. Les déclarations produites par les membres du panel ont été prélevées à chaque année depuis la formation du panel.

À l'image de l'échantillon transversal duquel il est issu, l'échantillon permanent de l'année de référence est caractérisé par des probabilités d'échantillonnage très différentes selon la strate. L'échantillon est stratifié selon le type de déclaration et le niveau de revenu. Il existe sept types de déclaration définis de façon hiérarchique, c'est-à-dire qu'une déclaration est classée dans la première catégorie dont elle répond aux critères. Les types de déclaration et les numéros de strate correspondants sont listés ci-dessous:

28	revenu élevé, non imposable;
38	bénéfice net ou perte nette élevés (entreprises);
80-84	revenu gagné de source étrangère (formule 2555);
90-94	crédit pour impôt étranger (formule 1116);
60-68	entreprise individuelle, commerciale (Annexe C);
50-58	entreprise individuelle, agricole (Annexe F);
40-48	revenu ne provenant pas d'une entreprise, commerciale ou agricole (toutes les autres déclarations).

Les déclarations des deux premières catégories sont échantillonnées à 100%. Dans le cas des autres catégories, les taux de sondage sont fonction du revenu ou (pour les déclarations d'entreprises commerciales ou agricoles) des recettes totales. Il existe cinq classes de revenu pour les déclarations de l'«étranger» et neuf classes pour les autres types de déclaration. Si l'on prend, par exemple, les déclarations qui ne viennent ni d'une entreprise commerciale ni d'une entreprise agricole, la strate 40 comprend les déclarations qui indiquent des revenus inférieurs à $25 000; le taux de sondage appliqué à ces déclarations était de 0.04%. La strate 48 comprend les déclarations qui indiquent des revenus de $5 millions ou plus; ces déclarations ont été échantillonnées avec une probabilité égale à 1. Pour une description complète des strates de panel, se référer à Schirm et Czajka (1992).

3. ÉVOLUTION DE LA COMPOSITION DE L'ÉCHANTILLON

3.1 Variation nette

Nous nous intéressons maintenant à la variation nette de la composition de l'échantillon par rapport aux variables qui ont servi à définir les strates d'échantillonnage. Autrement dit, si nous devions prélever un nouvel échantillon une ou deux années après l'année de référence de l'échantillon permanent et que nous devions utiliser les mêmes définitions de strate en tenant compte toutefois du revenu de l'année courante, dans quelles strates se trouveraient désormais les membres du panel?

Le tableau 1 donne, pour l'année de référence et les deux années suivantes (1988 et 1989), la distribution des déclarations de revenus des membres du panel selon la strate d'échantillon (membres choisis dans l'année de référence comme personnes non à charge). Les deux dernières colonnes du tableau indiquent pour 1988 et 1989 la variation en pourcentage par rapport à l'année de référence.

Nous observons dans la partie supérieure du tableau de fortes variations nettes du nombre de déclarations pour certains types de déclaration. Le nombre de déclarations indiquant un revenu élevé non imposable (strate 28) diminue de 70% au bout d'un an et de 75% au bout de deux ans. Le nombre de déclarations indiquant un bénéfice net ou une perte nette élevés (strate 38) chute de 41% au bout d'un an et de 56% au bout de deux ans. En revanche, le nombre de déclarations qui indiquent un crédit pour impôt étranger (strates 90 à 94) augmente très sensiblement en deux ans, le taux d'augmentation variant de 54 à 442%, selon la catégorie de revenu.

Dans la partie inférieure du tableau, nous observons de fortes variations pour les niveaux supérieurs de revenu. Le nombre de déclarations indiquant des revenus de $1 million ou plus (numéros de strate se terminant par un 6, un 7 ou un 8) chute de 40% en règle générale au bout de deux ans, tandis que le nombre de déclarations

indiquant des revenus de $500 000 à $1 million (numéros de strate se terminant par un 5) augmente de 50 à 60%.

Tableau 1: Déclarations de revenus de l'échantillon permanent selon la strate d'échantillonnage, 1987-1989: membres du panel choisis comme personnes non à charge.

Strate	Nombre de déclarations			Variation par rapport à 1987	
	1987	1988	1989	1988	1989
Total	89 755	86 353	87 318	-3.8%	-2.7%
28	873	259	220	-70.3	-74.8
38	9 590	5 635	4 202	-41.2	-56.2
80	29	37	48	27.6	65.5
81	7	29	42	314.3	500.0
82	120	158	158	31.7	31.7
83	167	101	79	-39.5	-52.7
84	39	29	16	-25.6	-59.0
90	50	96	141	92.0	182.0
91	55	171	298	210.9	441.8
92	531	992	1 345	86.8	153.3
93	957	1 222	1 656	27.7	73.0
94	747	1 083	1 149	45.0	53.8
60	3 089	2 901	3 651	-6.1	18.2
61	3 527	3 557	3 959	.9	12.2
62	3 763	3 779	3 707	.4	-1.5
63	2 291	2 517	2 595	9.9	13.3
64	1 896	3 208	3 529	69.2	86.1
65	1 078	1 678	1 734	55.7	60.9
66	1 732	1 514	1 483	-12.6	-14.4
67	1 684	1 350	981	-19.8	-41.7
68	985	812	609	-17.6	-38.2
50	259	301	495	16.2	91.1
51	493	517	619	4.9	25.6
52	374	375	396	.3	5.9
53	177	227	315	28.2	78.0
54	337	363	373	7.7	10.7
55	198	288	320	45.5	61.6
56	554	331	331	-40.3	-40.3
57	626	370	267	-40.9	-57.3
58	176	121	104	-31.3	-40.9
40	19 548	17 523	19 321	-10.4	-1.2
41	10 757	11 635	12 395	8.2	15.2
42	7 909	8 728	7 669	10.4	-3.0
43	2 559	3 176	3 059	24.1	19.5
44	3 176	3 417	3 299	7.6	3.9
45	1 415	2 141	2 187	51.3	54.7
46	3 343	2 211	2 087	-33.9	-37.6
47	2 939	2 135	1 487	-27.4	-49.4
48	1 705	1 366	990	-19.9	-41.9

Tableau 2: Pourcentage de membres du panel ayant changé de strate depuis l'année de référence.

Strate de l'année de référence	1988	1989
Total	37.7%	48.9%
28	79.5	87.3
38	43.4	59.7
80	29.6	60.0
81	57.1	57.1
82	32.7	57.4
83	48.4	66.0
84	48.6	78.4
90	45.5	63.3
91	69.1	66.7
92	52.4	61.6
93	55.4	62.8
94	50.0	53.8
60	35.0	40.3
61	41.6	54.6
62	40.8	57.1
63	48.6	63.2
64	49.2	64.0
65	64.5	76.6
66	69.4	78.6
67	63.5	79.6
68	57.2	70.8
50	29.2	28.1
51	34.5	41.4
52	44.5	62.8
53	52.4	69.1
54	50.5	69.9
55	60.2	77.2
56	66.7	79.0
57	62.5	75.9
58	71.4	81.4
40	13.5	14.4
41	26.3	35.8
42	26.7	45.8
43	46.1	63.1
44	46.8	60.9
45	59.9	71.9
46	65.7	74.5
47	62.2	76.7
48	53.4	65.5

3.2 Variation brute

Les variations nettes que l'on observe dans le tableau 1 peuvent être rattachées à des variations brutes beaucoup plus grandes. Le tableau 2 indique pour chaque strate de l'année de référence le pourcentage de membres du

panel qui se trouvaient dans une autre strate en 1988 et en 1989. Globalement, 38% et 49% des déclarations n'étaient plus dans leur strate originale en 1988 et en 1989 respectivement. Parmi les déclarations échantillonnées en 1987 qui indiquaient un revenu élevé non imposable, 80% n'étaient plus dans cette strate en 1988 et 87% ne l'étaient plus en 1989. De même, 60% des membres du panel qui avaient déclaré en 1987 un bénéfice ou une perte nets élevés ne faisaient plus partie de cette strate en 1989. Si l'on regarde maintenant les trois derniers groupes de strates, on constate que la probabilité de changer de strate dépend fortement du niveau de revenu en 1987. Par exemple, plus de 60% des déclarations qui étaient à l'origine dans l'une ou l'autre des strates n° 43 à 48 avaient changé de strate en 1989, alors que seulement 14% des déclarations qui appartenaient à la strate 40 en 1987 n'étaient plus dans cette strate deux ans plus tard.

Le tableau 3 sert à ventiler les effectifs des strates de 1987 en fonction des strates de 1989 pour les membres du panel échantillonnés dans les strates 40 à 48 (c.-à-d. déclarations indiquant un revenu qui ne provient ni d'une entreprise commerciale ni d'une entreprise agricole) et qui sont demeurés dans cette série de strates. À cause des très grosses différences de taux de sondage entre les catégories de revenu, on observe beaucoup plus de cas de recul que de cas de progression. Ainsi donc, tandis que des membres du panel échantillonnés à l'origine dans la strate 48 (revenu de $5 millions et plus) se trouvent répartis dans les sept strates inférieures en 1989, très peu de membres des strates inférieures - si ce n'est des trois qui précèdent immédiatement la strate 48 - sont passés à la strate 48 entre 1987 et 1989. En règle générale, les hausses de revenu qui permettraient aux membres de l'échantillon permanent de faire un bond de plus de deux catégories de revenu sont rares.

Tableau 3: Mobilité interstrates entre 1987 et 1989: strates 40 à 48.

Strate de 1987	Strates de 1989								
	40	41	42	43	44	45	46	47	48
40	14 833	1 381	135	11	*	*	*	*	*
41	2 361	6 940	622	18	5	*	*	*	*
42	396	2 292	4 362	210	14	*	*	*	*
43	79	159	881	951	153	10	*	*	*
44	54	93	280	803	1 242	175	30	5	*
45	26	26	69	101	387	396	120	17	4
46	43	39	116	192	399	674	848	247	31
47	44	41	85	101	235	293	562	692	198
48	23	12	23	36	113	89	119	250	594

* L'effectif par cellule est de trois ou moins.

4. PERTE DE PRÉCISION

Si l'on fait abstraction des questions de couverture de la population, le panel de la SOI produira des estimations non biaisées de moyennes et de totaux si les enregistrements du panel sont pondérés par l'inverse de leur probabilité de sélection pour l'année de référence. Or, un accroissement de la variance intrastrate causé par une modification des caractéristiques des membres du panel entraîne une perte de précision par rapport à l'année de référence (ou à un échantillon transversal de taille équivalente pour l'année courante). Cette perte de précision sera observée pour toute caractéristique qui a rapport à la stratification originale.

Le tableau 4 donne, pour l'échantillon permanent et l'échantillon transversal de la SOI, les coefficients de variation (CV) des estimations de totaux pour certaines variables de la déclaration de revenus pour les années 1988 et 1989. Les CV relatifs aux échantillons transversaux de 1988 et de 1989 ont été rajustés en fonction de la taille et de la composition par strate de l'échantillon permanent de l'année de référence.

Tableau 4: Coefficients de variation (en %) des estimations de totaux pour certaines variables de la déclaration de revenus, 1988 et 1989.

Item	1988 Échantillon transversal de 1988	Panel de 1988 (stratifié)		Échantillon transversal de 1989	Panel de 1989 (stratifié)	
		strates de 1987 seulement	stratifié a posteriori selon 1988		strates de 1987 seulement	stratifié a posteriori selon 1989
RBR ou déficit						
Revenu	.15	.31	.15	.15	.44	.21
Déficit	2.28	2.95	2.60	2.47	3.17	2.75
Salaires et traitements	.27	.29	.24	.28	.37	.25
Intérêt imposable	1.19	1.00	.97	.99	.99	.95
Dividendes	1.72	2.81	1.76	1.55	1.72	1.52
Revenus de pension ou de rente (dans le RBR)	1.73	1.47	1.41	1.39	3.20	1.44
Bénéfice net ou perte nette (entreprises)						
Bénéfice	1.36	1.46	1.24	1.29	1.58	1.28
Perte	3.24	3.47	3.17	3.16	3.77	3.46
Gain ou perte en capital nets						
Gain	1.08	3.18	1.98	1.20	4.93	3.61
Perte	2.79	2.16	2.24	2.40	2.15	2.28
Gain ou perte supplémentaires						
Gain	4.93	6.67	5.98	5.29	6.01	5.67
Perte	7.56	9.24	9.11	7.75	8.99	8.35
Revenu ou perte nets selon l'annexe E						
Revenu	1.42	2.72	1.83	1.53	3.62	1.79
Perte	1.63	1.78	1.65	1.74	2.67	2.02
Déductions détaillées - total	.60	.53	.51	.55	.87	.55
Impôt total à payer	.24	.56	.26	.27	.73	.37

Si nous comparons les CV de l'échantillon transversal à ceux de l'échantillon permanent pondéré en fonction de la stratification de 1987 seulement (poids du plan de sondage), nous constatons que pour les deux années étudiées, le CV pour l'échantillon permanent est presque deux fois plus élevé et même trois ou quatre fois plus élevé que le CV pour l'échantillon transversal pour cinq variables de la déclaration de revenus. Notons plus particulièrement qu'en 1988 le CV d'échantillon permanent pour le RBR (revenu brut ajusté), notion qui se rapproche sensiblement de la notion de revenu utilisée dans le plan de sondage de la SOI, est plus de deux fois supérieur au CV d'échantillon transversal pour la même variable. En 1989, le CV d'échantillon permanent est trois fois plus élevé que le CV d'échantillon transversal pour la même variable. On observe des rapports semblables pour la variable «Impôt total à payer».

5. MÉTHODES POUR COMPENSER L'EFFET D'UNE MODIFICATION DE LA COMPOSITION DU PANEL

Deux méthodes générales peuvent être utilisées pour compenser après coup les effets d'une modification de la taille et de la composition d'un échantillon permanent. La première méthode consiste à repondérer les observations du panel et la seconde, à augmenter l'échantillon en y ajoutant de nouvelles observations.

5.1 Repondération

La stratification a posteriori est un moyen de corriger les poids de l'échantillon permanent dans le but de compenser les effets d'une modification de la composition de l'échantillon. En ce qui concerne le panel de la SOI, nous observons d'importants transferts d'unités entre les catégories de revenu d'où ont été tirées au départ les déclarations de l'échantillon permanent et les catégories de revenu dans lesquelles ces déclarations seraient classées pour le tirage d'un échantillon transversal dans les années subséquentes. On peut accroître la précision des estimations transversales en stratifiant *a posteriori* sur l'appartenance aux strates selon l'année courante, à l'aide de chiffres de population que l'on peut aisément obtenir.

Les CV qui figurent dans les troisième et sixième colonnes du tableau 4 sont le résultat d'une stratification a posteriori de l'échantillon permanent en fonction des chiffres de population de l'année courante (ceux-là même qui sont utilisés pour la pondération des estimations de l'échantillon transversal). La stratification a posteriori a pour effet de réduire sensiblement le CV estimé pour chaque variable pour laquelle nous avons observé auparavant un large écart entre le CV pour l'échantillon transversal et le CV pour l'échantillon permanent. Dans le cas du RBR, des dividendes et de l'impôt total à payer, en 1988, et dans le cas des revenus de pension ou de rente, en 1989, la stratification a posteriori ramène le CV pour l'échantillon permanent au même niveau que le CV pour l'échantillon transversal. Les résultats sont presque aussi probants dans le cas du revenu net selon l'annexe E pour les deux années.

Bref, même la forme très élémentaire de stratification a posteriori qui a été exécutée ici compense largement l'effet négatif d'une modification de la composition de l'échantillon sur la précision des estimations transversales. Lorsque de fortes différences subsistent, on peut améliorer davantage la précision en effectuant une stratification a posteriori en fonction de nouveaux chiffres de population. Si cela ne suffit pas, il faudrait envisager des méthodes d'estimation davantage basées sur des modèles.

5.2 Augmentation de l'échantillon

Certains plans d'échantillonnage prévoient l'ajout d'observations dans le but de tenir compte de la croissance inexprimée de la population cible. Une méthode plus courante est de remplacer purement et simplement un panel par un autre. Dans la Survey of Income and Program Participation par exemple, la durée d'un panel est de 2 ans et demi et de nouveaux panels sont introduits à chaque année. La question du fardeau de réponse et celle de l'attrition sont probablement plus importantes que la question de la couverture. En revanche, dans le cas d'un panel de dossiers administratifs, le fardeau de réponse n'est généralement pas une question pertinente et l'attrition se limite principalement aux unités qui quittent la population cible. Compte tenu de la durée virtuellement plus longue du panel de dossiers administratifs, la question de la couverture prend de l'importance et nous sommes d'avis qu'il en va de même des méthodes qui servent à compenser les effets d'une modification de la composition de l'échantillon.

Une méthode générale destinée à cette fin pourrait consister à élargir le panel à des intervalles précis afin de compenser la perte d'observations qui présentent des caractéristiques particulières. En ce qui concerne le panel de la SOI par exemple, on pourrait compenser la perte de déclarations qui indiquent des revenus élevés en introduisant dans le panel à chaque année ou à tous les deux ans de nouvelles unités d'observation à revenu élevé. Dans un ouvrage antérieur, nous avons mis en doute l'efficacité de cette méthode pour ce qui a trait à l'amélioration du taux de couverture offert par un échantillon permanent, en faisant valoir notamment qu'il est très difficile de sonder efficacement les segments qui manquent (Czajka et Schirm 1992). En revanche, si l'augmentation de l'échantillon a pour but d'ajouter des observations dans des segments bien définis comme des strates d'échantillon, le remplacement *peut* se faire efficacement (en reprenant simplement les méthodes d'échantillonnage originales avec les caractéristiques de l'année courante).

Le fait de recueillir des données des années précédentes sur ces nouvelles unités d'observation aura deux avantages pour le panel. Premièrement, il sera possible d'utiliser ces unités d'observation pour accroître le degré de représentativité de l'échantillon par rapport aux transitions vers le haut. Deuxièmement, l'analyse d'un échantillon permanent à durées de panel différentes posera moins de difficultés.

6. CONSIDÉRATIONS RELATIVES À LA CONCEPTION DE PANELS

Les méthodes exposées ci-dessus servent à résoudre les problèmes que crée une modification de la composition de l'échantillon permanent. Comment concevoir un échantillon permanent qui soit stratifié selon les exigences et soit en même temps plus à l'abri des effets d'une modification de la composition?

6.1 Échantillonnage en fonction de la mi-durée

Une solution qui ressort de notre analyse est de former un panel de manière à ce que celui-ci ait telle composition à la *mi-durée* et non à l'année de référence. Il y a diverses manières d'y arriver. L'une d'elles consiste à tirer l'échantillon permanent dans l'année médiane de sa durée de vie prévue. Une fois cette opération accomplie, on pourrait recueillir des données rétrospectivement jusqu'à un point déterminé dans le temps puis prospectivement. Cette méthode permettrait l'étude des transitions d'une strate à l'autre. Le Département du Trésor des États-Unis a constitué un panel de cette manière à partir de données de déclarations de revenus (Hubbard, Nunns et Randolph, 1992). Une autre méthode consiste à déterminer la composition que devrait avoir le panel à la mi-durée et à définir à partir de cet objectif, en se servant des probabilités de transition estimées, l'échantillon qui devrait être formé à l'année de référence et dont la composition initiale évoluera vers la composition souhaitée à la mi-durée. Le succès de cette méthode dépend de deux choses: 1) la connaissance des probabilités de transition pertinentes et 2) l'élimination des faibles probabilités par l'introduction de conditions. Dans le cas du panel de la SOI par exemple, si l'on voulait que ce panel ait la composition souhaitée à la mi-durée pour, disons, la strate 28, il faudrait des probabilités de transition très faibles, donc des échantillons énormes dans l'année de référence. Pour obtenir l'effectif voulu pour la strate 28, il nous faudrait définir des sous-strates avec des probabilités de transition beaucoup plus élevées, de sorte que nous puissions procéder à un suréchantillonnage. Il arrive trop souvent que nous ignorions les conditions qui s'appliquent ou que nous ne soyons pas capables de les reproduire à l'aide des données dont nous disposons pour la stratification.

6.2 Conception de panel dans une perspective longitudinale

L'approche dont nous venons de parler est encore de nature essentiellement transversale. Il existe d'autres approches qui permettent d'envisager le problème de la conception de panel dans une perspective longitudinale; cela signifie essentiellement que l'on ajoute une dimension temporelle à la population dans laquelle nous souhaitons prélever un échantillon. Si, au moment de la conception d'un panel, nous disposions de données longitudinales sur les unités d'échantillonnage éventuelles, comment concevrions-nous l'échantillon?

Si l'on considère quelques-uns des usages des données fiscales longitudinales dont nous avons parlé plus haut, on peut envisager diverses variables de stratification pour obtenir un échantillon longitudinal *idéal* de déclarations de revenus. Retenons-en trois: le revenu cumulatif, les transitions d'une catégorie de revenu à l'autre (ou, de façon plus générale, d'un «état» à l'autre) et les «événements» observés (par exemple, l'utilisation de crédits d'impôt ou la décision par les membres d'un couple de remplir chacun de leur côté une déclaration de revenus). Une enquête rétrospective permettrait de tirer un échantillon en fonction de ces caractères, mais les coûts de sélection préliminaire seraient exorbitants, et on connaît bien, en outre, les limites des données d'enquêtes rétrospectives. Et si on optait pour un échantillon de dossiers administratifs? Les données couplées nécessaires pourraient ne pas exister ou il pourrait être impossible de créer des couplages. Quelles autres solutions s'offrent à nous? Considérons quelques avenues pour chacun des trois types de variable longitudinale.

6.2.1 Échantillonnage du revenu cumulatif

On peut prélever un échantillon selon le revenu cumulatif en tentant de prévoir la valeur de cette variable à partir de données transversales sur le revenu. Cela peut se faire en calculant le revenu cumulatif prévu comme

la somme pondérée de composantes de revenu multiples, les composantes plus «stables» étant plus fortement pondérées que les composantes moins stables. On retrouve quelque chose de semblable dans la notion de «revenu permanent» qu'utilisent les économistes; leur approche prend souvent la forme d'une prédiction par régression du revenu à l'aide de plusieurs variables ayant trait au capital humain et autres variables.

L'utilisation du revenu cumulatif comme variable de stratification présente des avantages et des inconvénients. Un des avantages est la sous-pondération des composantes de revenu instables. Il se peut que des composantes instables comme les gains en capital ou le revenu d'une société de personnes aient été à l'origine de l'inclusion d'unités de déclaration à revenu élevé dans l'échantillon de l'année de référence, mais ces unités seraient passées par la suite à des niveaux de revenu beaucoup plus bas. Un autre avantage de l'emploi du revenu cumulatif comme critère de stratification est le fait d'utiliser des variables de stratification comme des prédicteurs du revenu cumulatif qui prévoient aussi le changement ou le déterminent.

Parmi les inconvénients, notons la relative incapacité de prévoir les cas de personnes qui jouissent de nouvelles sources de revenu ou qui subissent de fortes variations de revenu (surtout positives). Un deuxième inconvénient est la possibilité que de bons prédicteurs du revenu cumulatif ne puissent servir à la stratification. En ce qui concerne le panel de la SOI, le fait de devoir s'en tenir aux prédicteurs qui figurent dans la déclaration de revenus est évidemment très restreignant; il n'existe pas de mesure directe du capital humain.

6.2.2 Échantillonnage des transitions

Nous avons remarqué que la stratification du panel de la SOI entraîne une forte surreprésentation des cas de transition vers les catégories de revenu inférieures par rapport aux cas de transition vers les catégories de revenu supérieures et ce déséquilibre est d'autant plus marqué que la transition est grande. La stratification d'un panel devrait refléter l'importance relative de divers types de transition pour les chercheurs qui se serviront des données. Plus l'intérêt pour les transitions sera grand, moins il sera indiqué d'effectuer une stratification globale.

Pour suréchantillonner efficacement des types particuliers de transition, il faut généralement connaître les facteurs qui déterminent ces transitions. Par exemple, quels contribuables à faible revenu sont susceptibles de faire partie du groupe de ceux dont le revenu connaîtra une forte croissance? Si l'on veut suréchantillonner des cas de transition d'un niveau de revenu faible à un niveau de revenu élevé, il faut pouvoir distinguer les unités de déclaration en fonction de la probabilité qu'elles connaissent une forte hausse de revenus. Si nous avions cette capacité, nous deviendrions probablement nous-mêmes millionnaires!

6.2.3 Échantillonnage des événements

Nous avons montré que la stratification du panel de la SOI produit beaucoup plus d'unités de déclaration à revenu élevé non imposable et d'unités de déclaration avec bénéfice net ou perte nette élevés dans l'année de référence que dans les années suivantes. Nous pouvons nous servir de ces deux types de déclaration pour illustrer les difficultés que pose la représentation d'«événements» dans le temps avec un panel. Les besoins de la recherche pourront exiger une représentation relativement uniforme de ces événements dans le temps. Pour atteindre cet objectif, il faut pouvoir déterminer les segments de population d'où proviendront dans les années à venir les rares groupes de contribuables à revenu élevé non imposable ou ceux avec bénéfice net ou perte nette élevés, de sorte que ces segments puissent faire l'objet d'un suréchantillonnage. Cela nous ramène à la connaissance que nous avons (ou que nous n'avons pas) des variables qui prédisent ou déterminent le changement et à la possibilité (ou l'impossibilité) de les utiliser comme critères de stratification.

7. CONCLUSION

Si la stratification d'un échantillon permanent comporte des variables qui ont des valeurs non fixes, alors une variation des caractéristiques de l'échantillon peut avoir pour effet de réduire la précision des estimations à long terme et de diminuer sensiblement le nombre d'observations qui pourraient servir à l'étude de comportements susceptibles d'intéresser l'analyste. Nous avons montré qu'en effectuant une stratification a posteriori en fonction des valeurs «courantes» des variables de stratification originales, on pouvait accroître sensiblement, dans les années qui suivent l'année de référence, la précision des estimations de caractéristiques transversales calculées

à partir d'un échantillon permanent. Nous avons avancé l'idée qu'il pouvait être nécessaire d'élargir l'échantillon initial en vue d'accroître le degré de représentation de certaines caractéristiques courantes si l'on voulait obtenir des échantillons d'une certaine taille dans les années qui suivent l'année de référence. Nous avons aussi souligné l'importance d'envisager la conception de panels dans une perspective longitudinale et nous avons fourni des exemples qui montrent comment aborder la question sous cet angle. Pour élaborer un échantillon permanent dans une perspective longitudinale, il faut définir un ordre de priorité pour les objectifs des analystes qui utiliseront les données. Les plans de sondage varieront selon l'ordre de priorité.

En prolongement de notre analyse, nous sommes d'avis que l'exposé que nous venons de faire suppose certaines conclusions quant à la durée d'un panel. Même si l'érosion d'un panel n'est pas une question importante ici, l'incapacité de compenser certains des effets négatifs d'une modification de la composition de l'échantillon (ou l'incapacité de réaliser l'échantillon idéal) limitera à certains égards la durée de vie du panel. Or, cette remarque peut ne pas s'appliquer à tous les segments du panel. C'est pourquoi nous proposons d'envisager des panels à durées multiples: certains segments du panel initial servent pendant de longues périodes tandis que d'autres ne durent que le temps de trois périodes d'observation. Les premiers répondraient à certains objectifs de recherche et les seconds, à d'autres objectifs. Par exemple, l'étude de phénomènes du cycle de vie nécessiterait l'observation d'un panel pendant une période relativement longue tandis que l'étude des transitions d'un «statut de déclarant» à un autre s'accommoderait bien d'un panel d'une durée relativement courte (ou serait probablement plus efficace avec un panel de ce genre).

REMERCIEMENTS

Cette étude a été rendue possible grâce au soutien financier de la division des statistiques du revenu de l'IRS, à qui les auteurs tiennent à exprimer leur reconnaissance. Les auteurs adressent aussi des remerciements à William Randolph et à Robert Gillette, de l'Office of Tax Analysis, pour avoir accepté de discuter des possibilités d'utilisation future des données fiscales longitudinales. Ils remercient également Roderick J.A. Little, de la UCLA et de Datametrics, pour ses commentaires et suggestions utiles sur les questions de la précision et de la stratification a posteriori, ainsi que Daisy Ewell, de Mathematica Policy Research, pour avoir exécuté avec adresse de nombreuses tâches de programmation complexes.

BIBLIOGRAPHIE

Czajka, J.L., et Schirm, A.L. (1992). Enhancing the representativeness of a longitudinal sample of individual tax returns: Weighting and sample supplementation. Dans *Proceedings of the Eighth Annual Census Bureau Research Conference.*

Hubbard, R.G., Nunns, J.R., et Randolph, W.C. (1992). Household income mobility during the 1980s: A statistical assessment based on tax return data. U.S. Department of the Treasury.

Schirm, A.L., et Czajka, J.L. (1992). Weighting a panel of individual tax returns for cross-sectional estimation. Dans *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

MÉTHODES DE PONDÉRATION POUR L'ENQUÊTE SUR LA DYNAMIQUE DU TRAVAIL ET DU REVENU

P. Lavallée et L. Hunter[1]

RÉSUMÉ

En 1994, Statistique Canada réalisera pour la première fois une enquête par panel à grande échelle auprès des ménages. L'enquête sur la dynamique du travail et du revenu (EDTR) portera sur les particuliers et les ménages et servira à observer dans le temps leur activité sur le marché du travail ainsi que les changements touchant leur revenu et leur situation familiale. En plus de fournir des données longitudinales, l'enquête permettra de produire des estimations transversales. Nous commencerons d'abord par décrire le plan de sondage de l'EDTR. Deuxièmement, nous discuterons de la détermination des poids de base qui correspondent, pour la plupart des particuliers, à l'inverse des probabilités de sélection. Troisièmement, nous examinerons les rajustements effectués pour la non-réponse et la post-stratification qui sera utilisée pour l'EDTR. Nous terminerons notre communication par une brève conclusion et un aperçu de projets futurs.

MOTS CLÉS: Enquête longitudinale, estimation transversale, probabilités de sélection, post-stratification.

1. INTRODUCTION

En 1994, Statistique Canada lancera une enquête par panel à grande échelle auprès des ménages. L'enquête sur la dynamique du revenu et du travail (EDTR) portera sur les particuliers et les ménages et servira à observer dans le temps leur activité sur le marché du travail ainsi que les changements touchant leur revenu et leur situation familiale. L'enquête vise d'abord à fournir des données longitudinales. Des estimations annuelles (souvent appelées estimations transversales) seront aussi produites.

Dans la présente communication, on aborde le problème de la représentativité de l'échantillon longitudinal de l'EDTR pour l'établissement des estimations transversales. Pour atteindre ce but, il faut élaborer un plan de pondération approprié. Il s'agit de concevoir la meilleure méthode de pondération possible, c'est-à-dire une méthode à la fois efficace et réalisable sur le plan opérationnel et possédant en plus la propriété de ne pas introduire de biais. En dépit du fait que l'on s'intéresse surtout ici à l'estimation transversale, il faut garder à l'esprit que les études longitudinales donnent généralement une représentation transversale de l'année de sélection de leurs échantillons longitudinaux. Par conséquent, la présente communication examinera, d'une certaine manière, la pondération longitudinale et la pondération transversale.

Puisque de nombreuses enquêtes sont de nature transversale, il est souvent question de la pondération transversale dans les ouvrages sur l'échantillonnage. Toutefois, dans le cas présent, l'aspect longitudinal de l'enquête complique légèrement la situation. L'échantillon n'est pas resélectionné de façon indépendante à chaque vague d'interviews; il ne fait l'objet d'aucune mise à jour importante et se compose d'unités qui, pour la plupart, ont été choisies lors d'occasions précédentes. Il faut donc tenir compte de certains aspects particuliers en raison de la nature longitudinale de l'échantillon.

Nous commencerons d'abord par décrire le plan de sondage de l'EDTR. Deuxièmement, nous discuterons de la détermination des probabilités de sélection. Troisièmement, nous examinerons les rajustements faits pour la non-réponse et la post-stratification. Enfin, l'exposé se terminera par une brève conclusion et un aperçu de projets futurs.

---

[1] P. Lavallée et L. Hunter, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

2. PLAN DE SONDAGE

La population cible de l'EDTR comprend toutes les personnes, quel que soit leur âge, qui résident dans les provinces du Canada, à l'exception de celles qui vivent dans les territoires, les établissements institutionnels, les réserves indiennes et les camps militaires.

L'échantillon de l'EDTR de 1993 sera un sous-échantillon de l'enquête canadienne sur la population active (EPA). L'EPA permet de produire des estimations mensuelles de l'emploi total, du travail autonome et du chômage total. Elle fait appel à un plan de sondage stratifié à plusieurs degrés qui est fondé sur une base aréolaire où les logements constituent les unités finales d'échantillonnage. Toutes les personnes appartenant aux ménages qui occupent les logements choisis font partie de l'échantillon de l'EPA. Pour constituer l'échantillon, on a recours à un plan avec renouvellement selon lequel chaque mois, un de six groupes de renouvellement est remplacé après avoir fait partie de l'échantillon pendant six mois. Chaque groupe de renouvellement contient environ 10 000 ménages, ce qui représente à peu près 20 000 personnes. Pour plus de détails au sujet du plan de sondage de l'EPA, voir Singh et coll. (1990).

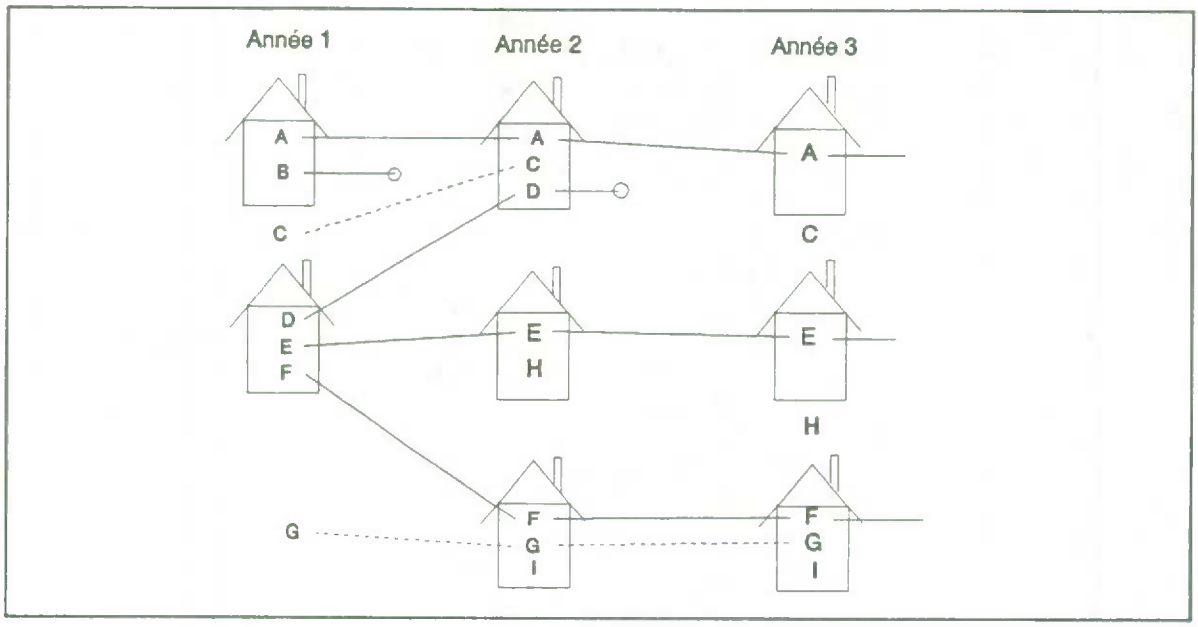
Pour 1993, l'échantillon de l'EDTR proviendra de deux groupes de renouvellement qui sortent de l'EPA. Par conséquent, il constituera un sous-échantillon de l'EPA. Cet échantillon, ou panel, contiendra environ 20 000 ménages. En 1996, un deuxième panel de 20 000 ménages sera choisi pour former un échantillon total de 40 000 ménages.

L'EDTR suivra les particuliers dans le temps, mais s'intéressera aussi aux caractéristiques des ménages. Toutes les personnes appartenant aux ménages qui occupent les logements sélectionnés seront choisies au début de l'enquête pour faire partie de l'échantillon de l'EDTR. Lors d'interviews ultérieures, les personnes vivant avec les particuliers compris dans l'échantillon initial seront aussi interviewées afin que des données puissent être obtenues pour le ménage au complet. Ces personnes représenteront soit des nouveaux entrants, soit des cohabitants. Les *nouveaux entrants* sont les personnes qui se joignent à la population cible et sont choisies pour faire partie de l'échantillon de l'EDTR. Ils sont représentés par les lettres H et I dans la figure 1. Par exemple, un nouvel entrant pourrait être un enfant né en 1994 ou une personne venue d'un pays étranger. Les *cohabitants* sont les personnes qui sont choisies pour faire partie de l'échantillon de l'EDTR, mais ne sont pas de nouveaux entrants dans la population cible. En fait, un cohabitant est une personne qui faisait partie de la population cible dès la première année, mais n'avait pas été choisie à ce moment-là. Par exemple, un cohabitant pourrait être une personne qui s'est mariée à un membre de l'échantillon de l'EDTR après la sélection de l'échantillon initial. Dans la figure 1, les lettres C et G représentent des cohabitants. Les personnes qui quittent la population cible, correspondant aux lettres B et D dans la figure 1, sont des *sortants*.

Il faudra tenir compte des nouveaux entrants et des sortants pour conserver la représentativité transversale. Il convient de signaler que la composante longitudinale de l'EDTR ne sera pas mise à jour après sa sélection. Tous les nouveaux membres (ajoutés, par exemple, parce qu'ils vivent avec des personnes de la composante longitudinale) ne seront pris en considération qu'à des fins transversales.

Un des problèmes que posent les nouveaux entrants a trait à la façon dont ils seront inclus dans l'échantillon. Certains nouveaux entrants seront échantillonnés parce qu'ils se joindront à un ménage longitudinal. On considère qu'un ménage est longitudinal s'il contient au moins une personne qui fait partie de l'échantillon longitudinal. Toutefois, cette méthode n'inclut pas les ménages ne comptant que de nouveaux entrants. On prévoit sélectionner les ménages contenant des nouveaux entrants au moyen d'un échantillon de logements obtenus de l'EPA. Notons que l'échantillon de l'EDTR, qui est un sous-échantillon de l'EPA, a été tiré initialement d'un échantillon de logements. Les logements utilisés pour la sélection des nouveaux entrants pourraient être les logements ayant servi à l'origine à la sélection du premier échantillon ou un nouvel ensemble de logements choisis indépendamment chaque année à partir de groupes de renouvellement sortant de l'EPA. Le fait de choisir de nouveau ou non des logements n'influe pas sur les caractéristiques statistiques de l'échantillon, mais a une incidence certaine sur les coûts d'exploitation. L'utilisation de l'échantillon de logements initiaux est coûteuse parce qu'il faut alors visiter chacun des logements chaque année pour trouver de nouveaux entrants. Si les logements échantillonnés proviennent de ceux qui sortent de l'EPA par renouvellement, le dépistage des nouveaux entrants peut être fait directement au cours des interviews de l'EPA. Toutefois, pour simplifier la discussion, nous supposerons que les *logements choisis initialement* serviront à la sélection des

Figure 1: Interview des particuliers dans les ménages.



nouveaux entrants.

3. PONDÉRATION DE BASE

La détermination des *poids de base* est un des points liés à la pondération de l'échantillon de l'EDTR. Pour la plupart des personnes, ces poids correspondent à l'inverse des probabilités de sélection. Les *poids de base* sont en fait les poids à utiliser dans le processus d'estimation avant tout rajustement ou toute post-stratification. La détermination des poids de base est compliquée par le fait que les cohabitants et les nouveaux entrants peuvent faire partie de l'échantillon à n'importe quelle vague d'interview en se joignant à un ménage longitudinal.

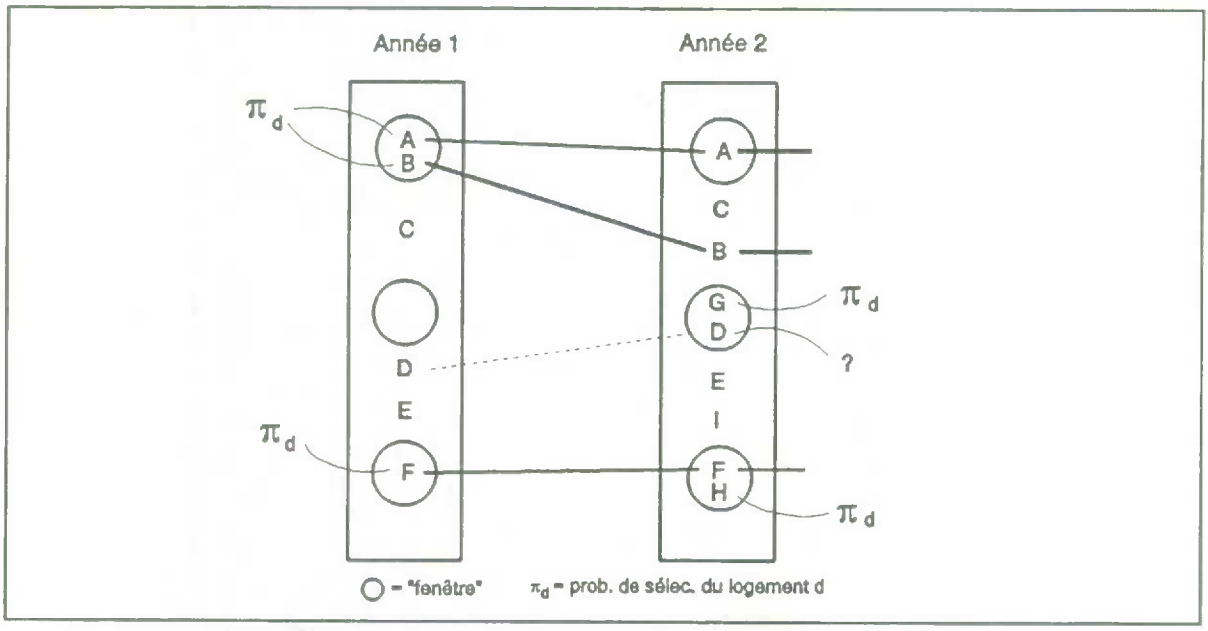
La première année, les personnes sont choisies au moyen d'un échantillon de logements tirés de la base aréolaire de l'EPA. On attribue alors à chaque personne j du logement choisi d la probabilité de sélection π_d. Au cours des années suivantes, les personnes choisies initialement conservent les probabilités de sélection attribuées la première année. Pour chaque personne j, nous avons donc la probabilité de sélection $\pi_j^{(o)}$ égale à π_d.

Pour les besoins de l'estimation transversale, l'échantillon sera mis à jour par suite de la présence des nouveaux entrants et des cohabitants. Les nouveaux entrants qui se joignent à l'échantillon par l'entremise d'un logement initial d auront la probabilité de sélection π_d du logement. C'est exactement comme si les nouveaux entrants avaient été choisis au moment de la sélection initiale. Il faut signaler que si un nouvel entrant est échantillonné du fait qu'il s'est joint à un ménage longitudinal qui occupe un logement initial d, on lui attribuera aussi la probabilité de sélection π_d du logement. En fait, les logements initiaux peuvent être considérés comme des "fenêtres" permettant d'entrer dans l'échantillon, comme le montre la figure 2. Les personnes échantillonnées par l'intermédiaire de ces "fenêtres" reçoivent la probabilité de sélection π_d du logement d et peuvent donc être considérées comme ayant été échantillonnées de façon légitime. Elles seront désignées par l'expression *personnes légitimes* (personnes initialement choisies et nouveaux entrants vivant dans un logement initial).

Au cours d'une année après la sélection initiale, prenons un logement initial qui est maintenant occupé par un ménage longitudinal. Il est possible que ce ménage comprenne des cohabitants et (ou) des nouveaux entrants en plus des membres longitudinaux. Du fait qu'ils ont été choisis la première année, ces derniers ont assurément des probabilités de sélection (voir les lettres A et F de la figure 2). On attribue aux nouveaux entrants la probabilité de sélection du logement étant donné que le ménage occupe un logement initial (voir les lettres G et H de la figure 2). Par ailleurs, les cohabitants n'ont pas la probabilité de sélection du logement parce qu'ils ne proviennent pas de l'échantillon initial, même s'ils occupent maintenant un logement initial (voir la lettre D

de la figure 2). Quiconque appartient à la population cible peut entrer dans l'échantillon par une "fenêtre" seulement au moment où l'échantillon initial est prélevé; seuls les nouveaux entrants dans la population visée peuvent entrer dans l'échantillon par une "fenêtre" lors de vagues subséquentes. Par conséquent, dans le cas d'un ménage échantillonné occupant un logement initial, les personnes longitudinales et les nouveaux entrants sont les seuls à être considérés des personnes légitimes.

Figure 2: Sélection des personnes à travers des "fenêtres".



Prenons maintenant un ménage longitudinal qui n'occupe pas un logement initial. Là encore, ce ménage peut comprendre des cohabitants ou des nouveaux entrants, en plus des membres longitudinaux. Comme le ménage ne réside pas dans un logement initial, on n'attribue pas aux nouveaux entrants la probabilité de sélection du logement. Quant aux cohabitants, ils n'ont pas la probabilité de sélection du logement parce qu'ils ne faisaient pas partie de l'échantillon initial. Lorsqu'un ménage longitudinal n'occupe pas un logement initial, seuls ses membres longitudinaux sont considérés des personnes légitimes.

Les probabilités de sélection des personnes illégitimes ne sont pas faciles à déterminer. Lorsque de nouveaux entrants s'ajoutent à un ménage longitudinal qui n'occupe pas un logement initial, on ne sait pas au juste quelles probabilités de sélection leur attribuer. Dans le cas des cohabitants, l'attribution de probabilités de sélection pose aussi un problème, étant donné que la seule raison pour laquelle ils sont ajoutés à l'échantillon est parce qu'ils se sont joints à un ménage longitudinal, même si ce ménage réside dans un logement initial.

En raison de son appartenance à la population cible dès la première année, chaque cohabitant a, en théorie, une probabilité de sélection. Toutefois, cette probabilité est habituellement inconnue. L'exemple suivant nous permet d'illustrer cette situation: dans un plan de sondage à plusieurs degrés, les cohabitants auraient pu faire partie d'une unité primaire d'échantillonnage (UPÉ) non incluse dans l'échantillon. Comme on ne se rend généralement pas dans les UPÉ non choisies, les probabilités de sélection des personnes dans ces UPÉ demeurent inconnues. Il faut mentionner que même si les probabilités de sélection des cohabitants pouvaient être établies avec précision, il serait nécessaire de mettre à jour les probabilités de sélection de toutes les personnes échantillonnées afin de rendre compte du fait que l'échantillon compte maintenant de nouveaux membres. Ce processus s'avérerait complexe et coûteux. Par conséquent, les cohabitants sont toujours considérés comme des personnes illégitimes.

Deux méthodes ont été proposées pour régler le problème de l'attribution des probabilités de sélection (ou poids de base) aux nouveaux entrants ne résidant pas dans un logement initial de même qu'aux cohabitants. La première méthode, que nous appellerons la méthode du partage des poids, a été décrite en partie par Ernst

(1989). La pondération pour l'Enquête «Survey of Income and Program Participation» (SIPP) s'inspire de cette méthode. La deuxième méthode, celle qui est proposée par J.N.K. Rao, est fondée sur l'estimation composite. Ces deux méthodes sont présentées ci-après.

3.1 Méthode du partage des poids

En gros, cette méthode s'inspirant de Ernst (1989) attribue à chaque personne illégitime choisie un poids de base établi à partir de la moyenne des poids calculée pour chaque ménage. Un *poids initial* qui correspond à l'inverse de la probabilité de sélection est d'abord obtenu pour chaque personne légitime. Deuxièmement, un *poids initial* de zéro est attribué à chaque personne illégitime. Le poids de base est ensuite obtenu en calculant la moyenne des poids initiaux au niveau du ménage. Enfin, le poids final est assigné à tous les membres du ménage. Il faut signaler que le fait d'attribuer le même poids de base à toutes les personnes présente l'avantage non négligeable d'assurer la cohérence des estimations des particuliers et des ménages.

Pour présenter de façon stricte la méthode du partage des poids, il faut considérer deux cas distincts.

Cas 1: Ménages occupant un logement initialement choisi d.

Supposons que le ménage choisi i de l'année 2 occupe un logement initial d. Posons que le ménage i contient $M_i^{(O)}$ personnes choisies durant l'année 1 (c.-à-d. les personnes longitudinales initiales), $M_i^{(C)}$ cohabitants et $M_i^{(N)}$ nouveaux entrants. Comme nous l'avons indiqué précédemment, on attribue aux nouveaux entrants la probabilité de sélection du logement d. On attribue à chaque membre j du ménage i du logement initial d le poids initial

$$w_{ij}' = \begin{cases} 1/\pi_j^{(O)} & \text{pour } j=1,\ldots,M_i^{(O)} \\ 0 & \text{pour } j=(M_i^{(O)}+1),\ldots,(M_i^{(O)}+M_i^{(C)}) \\ 1/\pi_d & \text{pour } j=(M_i^{(O)}+M_i^{(C)}+1),\ldots,(M_i^{(O)}+M_i^{(C)}+M_i^{(N)}) . \end{cases} \tag{1}$$

Puis, le poids de base w_i du ménage i est obtenu par

$$w_i = \frac{1}{M_i}\sum_{j=1}^{M_i} w_{ij}' , \tag{2}$$

où $M_i = M_i^{(O)}+M_i^{(C)}+M_i^{(N)}$. Le poids de base w_i est finalement attribué à chaque membre du ménage.

Cas 2: Ménages n'occupant pas un logement initial.

Prenons le ménage choisi i qui n'occupe pas un logement initial. Nous attribuons à chaque membre j du ménage i le poids initial

$$w_{ij}' = \begin{cases} 1/\pi_j^{(O)} & \text{pour } j=1,\ldots,M_i^{(O)} \\ 0 & \text{pour } j=(M_i^{(O)}+1),\ldots,(M_i^{(O)}+M_i^{(C)}) \\ 0 & \text{pour } j=(M_i^{(O)}+M_i^{(C)}+1),\ldots,(M_i^{(O)}+M_i^{(C)}+M_i^{(N)}) . \end{cases} \tag{3}$$

Le poids de base w_i du ménage i est obtenu au moyen de l'équation (2). Là encore, le poids de base w_i est assigné à chaque membre du ménage.

L'estimation $\hat{Y}$ de la population totale Y est finalement obtenue par

$$\hat{Y} = \sum_{i=1}^{n}\sum_{j=1}^{M_i} w_{ij} y_{ij} , \tag{4}$$

où n est le nombre total de ménages choisis et $w_{ij} = w_i$ pour tous les membres j du $i^{\text{ième}}$ ménage choisi. En procédant comme Ernst (1989), on peut démontrer que cette estimation est sans biais.

3.2 Méthode de l'estimation composite

La méthode de l'estimation composite proposée par J.N.K. Rao est axée davantage sur l'établissement d'un estimateur que sur la pondération de chaque personne choisie. En fait, elle fait appel à la modélisation pour l'établissement d'un estimateur relatif aux personnes illégitimes. Brièvement, les personnes choisies sont réparties en deux ensembles, D_L et D_{NL}, selon qu'elles sont légitimes ou non. Un estimateur de la population totale Y est ensuite produit à partir de chacun des deux ensembles. On obtient l'estimation à partir de l'ensemble D_{NL} de personnes illégitimes en supposant que ces dernières sont choisies selon un plan d'échantillonnage aléatoire simple stratifié. Un estimateur composite est finalement créé à l'aide des deux estimations pondérées par le nombre relatif de personnes dans chaque ensemble.

L'établissement de l'estimateur composite repose sur le raisonnement suivant: en temps normal, on tiendrait uniquement compte des personnes légitimes pour l'estimation de quantités transversales telles que la population totale Y. Toutefois, étant donné que tous les membres des ménages sont interviewés pour obtenir les estimations des ménages, le fait de ne pas compter les personnes illégitimes équivaudrait à ne pas considérer une partie de l'information. Il s'agit donc de tenir compte des personnes illégitimes, mais seulement en proportion de leur importance relative.

Ensemble D_L: personnes légitimes

Chaque personne j de l'ensemble D_L se fait attribuer un poids qui correspond à l'inverse de sa probabilité de sélection π_j. Une estimation $\hat{Y}_L$ de la population totale Y est ensuite calculée de la façon suivante:

$$\hat{Y}_L=\sum_{j\in D_L}\frac{y_j}{\pi_j}. \tag{5}$$

Ce résultat correspond simplement à l'estimateur de Horvitz-Thompson.

Ensemble D_{NL}: personnes illégitimes

On suppose que l'échantillon de l'ensemble D_{NL} correspond à un échantillon aléatoire simple stratifié tiré d'une superpopulation qui pourrait être décrite au moyen du modèle suivant:

$$Y_j=\mu_h+\epsilon_j \; si \; j\in h, \tag{6}$$

où $E(\epsilon_j)=0$, $\mathrm{Var}(\epsilon_j)=\sigma_h^2$ et $\mathrm{Cov}(\epsilon_j,\epsilon_{j'})=0$ pour $j\neq j'$. On suppose que les strates h correspondent à un niveau agrégé comme par exemple celui des provinces.

Une deuxième estimation $\hat{Y}_{NL}$ de la population totale Y est alors calculée de la façon suivante:

$$\hat{Y}_{NL}=\sum_{h=1}^{H}\frac{M_h}{m_{NL,h}}\sum_{j=1}^{m_{NL,h}}y_{hj}, \tag{7}$$

où M_h est le nombre de personnes dans la strate h, $m_{NL,h}$ est le nombre de personnes illégitimes échantillonnées dans la strate h et H est le nombre total de strates.

L'estimation finale de la population totale Y est finalement établie à partir d'un estimateur composite de la forme suivante:

$$\hat{Y}_c=\frac{1}{(m_L+m_{NL})}(m_L\hat{Y}_L+m_{NL}\hat{Y}_{NL}), \tag{8}$$

où m_L et m_{NL} représentent le nombre de personnes dans D_L et D_{NL} respectivement.

L'estimateur composite $\hat{Y}_c$ attribue un poids initial à chaque personne j du ménage i de la façon suivante:

$$w'_{cij}=\begin{cases}\dfrac{m_L}{(m_L+m_{NL})}\dfrac{1}{\pi_j} & \text{pour } j\in D_L \\ \dfrac{m_{NL}}{(m_L+m_{NL})}\dfrac{M_h}{m_{NL,h}} & \text{pour } j\in D_{NL} \text{ et } j\in h\ .\end{cases} \tag{9}$$

Ces poids pourraient servir directement à l'estimation de différentes caractéristiques y. Toutefois, pour assurer la cohérence des estimations des personnes et des ménages, il est préférable d'avoir un seul poids par ménage. Par conséquent, on propose donc de faire la moyenne des poids initiaux w'_{cij} pour chaque ménage afin d'obtenir les poids de base.

$$w_{ci}=\frac{1}{M_i}\sum_{j=1}^{M_i} w'_{cij}\ . \tag{10}$$

Puis, ce poids est assigné à chaque personne, légitime ou non, qui appartient au ménage i.

En fin de compte, l'estimateur composite $\hat{Y}_c$ est biaisé en fonction du plan. Ce biais provient de $\hat{Y}_{NL}$ en grande partie parce que l'échantillon tiré de D_{NL} n'est pas un échantillon aléatoire simple stratifié. On peut aussi alléguer la possibilité que la deuxième estimation $\hat{Y}_{NL}$ n'est pas vraiment représentative de la population actuelle Y parce que les personnes illégitimes forment une population de personnes "sociables" du fait qu'elles font partie de l'ensemble D_{NL} seulement parce qu'elles se sont jointes à un ménage non vide. Autrement dit, les personnes "non sociables" qui ne vivront jamais avec d'autres personnes n'ont aucune chance d'être comptées parmi les personnes illégitimes. Si elles agissent différemment des personnes légitimes, un biais peut alors être introduit dans l'estimation. Il faut mentionner que le biais de $\hat{Y}_{NL}$ sera vraisemblablement faible à cause de la taille relativement petite de l'échantillon m_{NL}, mais prendra de l'importance avec le temps à mesure que m_{NL} augmentera.

3.3 Résultats de simulations

Il a été nécessaire d'effectuer des simulations afin de déterminer laquelle des deux méthodes était la meilleure, celle du partage des poids ou de l'estimation composite. Heureusement, le plan d'échantillonnage et les questions sur le travail de l'EDTR sont semblables à ceux d'une enquête déjà existante, l'Enquête sur l'activité (EA). Par conséquent, on a utilisé les données de l'EA pour faire les simulations.

L'EA a porté sur deux panels et pour chacun d'entre eux, l'échantillon était composé de cinq groupes de renouvellement de l'EPA. Le premier panel a été interviewé pendant une période de deux ans (1986-1987) et le deuxième, pendant une période de trois ans (1988-1989-1990). Pour le panel de trois ans, des renseignements ont aussi été recueillis auprès des nouveaux entrants et des cohabitants, tout comme on projette de le faire pour l'EDTR. Toutefois, à la différence de l'EDTR, le plan de sondage de l'EA prévoit des échantillons distincts, mais chevauchants, pour sa composante transversale et sa composante longitudinale. Pour l'échantillon longitudinal, les personnes échantillonnées la première année ont été retracées et réinterviewées la deuxième (et troisième) année. Dans le cas de l'échantillon transversal, les logements choisis initialement ont été visités de nouveau la deuxième (et troisième) année, et toute personne dans le champ de l'enquête qui y vivait était interviewée. Par conséquent, les personnes choisies initialement qui n'ont pas déménagé ont fourni des données tant longitudinales que transversales, et celles qui ont déménagé n'ont fourni que des données longitudinales; les personnes qui n'ont pas été choisies la première année, mais qui se sont installées dans un logement initial n'ont fourni que des données transversales. Selon le raisonnement à la base des rajustements pour pondération faits dans le cadre de l'EDTR, plutôt que d'avoir un échantillon transversal distinct, il est possible d'atteindre le même degré de représentativité à l'aide des personnes longitudinales et de leurs cohabitants. On s'est fondé sur les données du deuxième panel de l'EA pour tester cette hypothèse.

Pour commencer, les données des deux premières années (1988-1989) ont été utilisées. Chaque personne appartenait à une de trois catégories: personne longitudinale initiale, nouvel entrant dans la population cible ou cohabitant. Chaque ménage était réparti selon qu'il occupait un logement initial ou un logement non

échantillonné, ou non initial, par suite du déménagement d'une personne longitudinale. D'après le fichier de l'EA de 1989, il y avait 49 874 personnes à la fois dans l'échantillon longitudinal et dans l'échantillon transversal; 8 016 d'entre elles étaient des répondants longitudinaux seulement, 14 874 étaient uniquement des répondants transversaux et 2 355 étaient des cohabitants qui vivaient avec des répondants longitudinaux seulement. Des 60 245 personnes dans l'échantillon longitudinal, 82.8% faisaient aussi partie de l'échantillon transversal.

Tableau 1: Comparaison des estimations nationales, répondants de 25-64 ans, EA 88-89 (post-stratification selon la province, l'âge et le sexe).

	Fichier transversal (1989)		Partage des poids Estimation du fichier longitudinal	Estimation composite Estimation du fichier longitudinal	
	estimation	c.v. (%)		Poids dont on n'a pas fait la moyenne	Poids dont on a fait la moyenne
Nombre moyen sem. travail	38.0	0.5	38.5	38.4	38.3
Nombre moyen sem. chômage1	2.0	2.7	1.9	2.0	2.0
Nombre moyen sem. inact.	12.7	1.3	12.2	12.3	12.4
Sem. travail = 0 (%)	19.5	1.4	18.5	18.5	18.4
Sem. travail = 1-26	8.2	2.7	8.1	8.2	8.2
Sem. travail = 27-48	10.8	2.2	10.8	10.9	10.9
Sem. travail = 49+	61.5	0.7	62.6	62.3	62.2
Sem. chômage1 = 0 (%)	88.0	0.3	88.5	88.3	88.3
Sem. chômage1 = 1-26	9.1	1.9	8.8	9.0	8.9
Sem. chômage1 = 27-48	2.5	4.4	2.4	2.4	2.4
Sem. chômage1 = 49+	0.4	15.7	0.3	0.3	0.3
Sem. inact. = 0 (%)	64.4	0.1	65.2	65.0	64.9
Sem. inact. = 1-26	12.5	6.7	12.6	12.7	12.6
Sem. inact. = 27-48	5.0	17.1	4.8	4.9	4.9
Sem. inact. = 49+	18.1	10.7	17.4	17.4	17.6
Nombre d'emplois = 0 (%)	19.5	1.4	18.5	18.5	18.7
Nombre d'emplois = 1	66.5	0.5	69.7	69.5	69.4
Nombre d'emplois = 2	11.4	2.1	9.8	9.9	9.9
Nombre d'emplois = 3+	2.5	4.4	2.0	2.0	2.0
ÉM = marié (%)	74.6	0.5	76.8	76.6	77.5
ÉM = célibataire	15.7	2.3	13.6	13.7	12.9
ÉM = veuf/veuve	2.0	4.9	2.1	2.1	2.1
ÉM = séparé(e)/divorcé(e)	7.7	3.0	7.5	7.6	7.6

PA = population active, ÉM = état matrimonial

Un poids de départ a été attribué à chaque personne longitudinale d'après la probabilité de sélection établie en fonction du logement occupé par cette personne la première année. On a ensuite procédé à un rajustement par quotient dans les composantes des strates afin de compenser la non-réponse aux interviews de la deuxième année. Ces poids corrigés en fonction de la non-réponse ont été employés comme poids initiaux pour les différentes méthodes de pondération décrites en 3.1 et en 3.2. Pour la méthode de l'estimation composite, on a supposé que la stratification pour les personnes illégitimes se faisait au niveau provincial. Une post-

stratification finale a été effectuée selon chaque méthode en fonction des effectifs du recoupement des provinces, des groupes d'âges et des sexes. Des poids finaux ont été calculés d'après les méthodes proposées pour les personnes longitudinales et leurs cohabitants qui ont participé à l'enquête en 1989. Les estimations ont ensuite été établies à l'aide de ces poids et ont été comparées aux estimations de l'EA fondées sur les données transversales de 1989.

Des estimations de proportions de la population ont été calculées selon l'état matrimonial, le nombre d'emplois tenus en 1989, le nombre de semaines de travail, de chômage et d'inactivité en 1989 ainsi que le nombre moyen de semaines de travail, de chômage et d'inactivité. Les résultats sont présentés au tableau 1. De façon générale, la méthode de l'estimation composite a donné d'aussi bons résultats que la méthode du partage des poids en ce sens que les estimations se rapprochaient des estimations transversales réelles. On a observé très peu d'écart entre les résultats obtenus au moyen des deux versions de la méthode d'estimation composite (utilisation de poids différents et identiques pour différents membres du ménage). En raison de la propriété qu'elle a de ne pas introduire de biais, la méthode du partage des poids pourrait alors être préférable à la méthode de l'estimation composite.

4. RAJUSTEMENTS EN FONCTION DE LA NON-RÉPONSE ET POST-STRATIFICATION

La présente section porte sur les corrections qui seront apportées aux poids de base afin d'améliorer les estimations. Ces corrections correspondent à un facteur multiplicatif que l'on appelle poids-g (voir Särndal et coll. (1992)). Autrement dit, pour chaque personne j du ménage i, le poids de base w_{ij} est multiplié par un certain facteur de correction afin d'obtenir le poids final $w_{ij}^F = g_{ij|s} w_{ij}$. Habituellement, le poids-g $g_{ij|s}$ est fonction de la méthode de pondération et des échantillons sélectionnés. Malheureusement, les résultats des simulations ne sont pas encore disponibles.

4.1 Rajustements pour la non-réponse

Comme toutes les enquêtes, l'EDTR aura à faire face au problème de la non-réponse. Nous pouvons nous attendre à un certain taux de non-réponse à une ou à plusieurs questions ou au questionnaire en entier, de même qu'à des non-répondants à une vague donnée ou à des non-répondants irréductibles. En outre, il peut y avoir des données manquantes sur des membres d'un ménage ou sur le ménage au complet. Des mesures correctives devront donc être prises pour résoudre ces cas complexes de non-réponse.

Dans la mesure du possible, on aura recours à l'imputation pour régler les cas de non-réponse. Pour effectuer l'imputation, la principale exigence est qu'au moins un membre faisant partie du ménage au moment de l'interview ait répondu au questionnaire. Par conséquent, il est peu probable que l'on fasse appel à cette méthode si tous les membres d'un ménage sont des non-répondants. À titre d'exemple, un nouveau ménage formé par une personne ayant quitté un ménage choisi ne fera pas l'objet d'une imputation si cette personne est un non-répondant parce que nous ne possédons aucun renseignement sur la composition actuelle de son ménage. Lorsque l'imputation ne peut pas être faite, une correction pour tenir compte de la non-réponse sera apportée aux poids de base.

Le rajustement d'un poids en fonction de la non-réponse peut être effectué en établissant un modèle de réponse (ou de non-réponse). Avec la modélisation de la réponse, on fait un ensemble de suppositions portant sur le véritable mécanisme de réponse de l'enquête; ce mécanisme nous étant généralement inconnu. La méthode de régression logistique s'avère alors particulièrement pratique (voir Little (1986) et Hunter, Michaud et Torrance (1992)). La fonction de réponse logistique multiple est la suivante:

$$E(R_i|\underline{x}_i) = [1+\exp(-\underline{\beta}'\underline{x}_i)]^{-1}, \qquad (11)$$

où R_i est la variable dépendante, $\underline{\beta}$ est le vecteur-colonne des paramètres de régression et $\underline{x}_i$ est un vecteur de variables indépendantes disponibles pour tous les ménages. La variable dépendante R_i égale 1 si le ménage i est un répondant; autrement, elle égale 0. Par conséquent, $E(R_i|\underline{x}_i)$ peut être considéré comme la probabilité de réponse $\theta_{i|s}$, qui peut dépendre de l'échantillon s.

Après avoir fait l'estimation de $\underline{\beta}$ selon la méthode du maximum de vraisemblance, nous obtenons une probabilité de réponse estimée pour un ménage ayant la valeur $\underline{x}_i$:

$$\hat{\theta}_{i|s}=[1+\exp(-\hat{\underline{\beta}}'\underline{x}_i)]^{-1}. \tag{12}$$

Après avoir corrigé le poids de base w_i du ménage i, le poids de base corrigé $w_i^{\wedge}$ est donné par:

$$w_i^{\wedge}= w_i \times \frac{1}{\hat{\theta}_{i|s}}. \tag{13}$$

Il faut mentionner que les probabilités de réponse estimées peuvent servir à former des groupes de réponse homogène (GRH) dans lesquels on suppose que tous les éléments échantillonnés ont la même probabilité de réponse. Celle-ci est alors simplement estimée au moyen du taux de réponse dans chaque GRH (voir Särndal et coll. (1992)).

L'utilisation d'un modèle de réponse pour l'EDTR pose un problème, soit celui de la disponibilité des variables auxiliaires à employer comme variables dépendantes $\underline{x}_i$ avec la fonction de réponse logistique (11). Comme les ménages qui feront l'objet d'un rajustement pour non-réponse seront ceux dont on n'a obtenu aucune réponse des membres actuels du ménage, il est probable qu'il n'y aura presque aucun renseignement sur le ménage actuel. Par exemple, il serait inutile d'utiliser les régions et (ou) la taille des ménages comme variables auxiliaires dans le cas des ménages que l'on a pas retrouvés. Par conséquent, les corrections faites pour tenir compte de la non-réponse des ménages à l'aide d'un modèle de réponse ne semble pas une solution pratique pour l'EDTR. On espère en fait que le problème de la non-réponse des ménages pourra être réglé en grande partie par la post-stratification.

4.2 Post-stratification

Dans les enquêtes par échantillonnage, on a recours à la post-stratification pour deux grandes raisons. Premièrement, cette méthode sert à corriger la sous-représentation de certaines sous-populations dans l'échantillon. Par exemple, cette sous-représentation peut être imputable à la non-réponse. Plus particulièrement, si les probabilités de réponse $\theta_{i|s}$ obtenues au moyen de l'équation (11) reposent sur les mêmes variables utilisées pour définir les post-strates, la post-stratification corrigera de façon implicite la non-réponse (voir Särndal et coll. (1992)). Deuxièmement, si l'on constate que les variables d'intérêt sont homogènes dans certaines catégories (ou post-strates), la post-stratification selon ces catégories permettra d'améliorer la précision des estimations.

Étant donné que l'échantillon de l'EDTR est un sous-échantillon de l'EPA, il convient, pour des raisons d'ordre pratique, de prendre en considération au moins les mêmes variables employées par l'EPA pour la post-stratification. Il est cependant possible d'accroître cet ensemble de variables utilisées pour la post-stratification en y ajoutant d'autres variables pertinentes. Les estimations de l'EPA sont présentement post-stratifiées en fonction des régions et des groupes âge-sexe (voir Singh et coll. (1990)). Cette post-stratification est effectuée au niveau des particuliers, mais selon une approche intégrée qui donne un poids égal à tous les membres d'un ménage (voir Lemaître et Dufour (1987)).

Les variables de post-stratification que l'on pense utiliser pour l'EDTR sont la région, l'âge, le sexe, le revenu et la mobilité interprovinciale. Ces variables ne seront pas toutes recoupées parce que les effectifs de population ne sont pas disponibles et qu'un tel recoupement produirait un grand nombre de cellules, y compris de nombreuses cellules sans unités échantillonnées. Deux ensembles d'effectifs de population sont mis à l'essai: pour un de ces ensembles, on recoupe les régions, les groupes âge-sexe et les catégories de revenu, et pour l'autre, on recoupe les provinces, les groupes âge-sexe et la mobilité interprovinciale.

La post-stratification en fonction du revenu vise principalement à corriger la sous-représentation des ménages à faible revenu et à revenu élevé. D'une part, il apparaît que les ménages à faible revenu ont tendance à avoir une plus grande mobilité que les ménages à revenu moyen ou élevé, ce qui les rend plus difficiles à retracer. Une fois retracés, ils sont davantage enclins à ne pas collaborer, particulièrement lorsque les questions portent sur des sujets de nature délicate comme le revenu. D'autre part, une bonne proportion des ménages à revenu élevé sont des non-répondants parce qu'ils refusent de fournir les renseignements requis ou bien parce qu'ils en

sont incapables (par ex., leurs affaires sont gérées par un comptable). Pour faire la post-stratification en fonction du revenu, il faudra établir des catégories de revenu qui seront déterminées à l'aide de simulations basées sur les données de l'EA, avec l'ajout d'une variable revenu correspondant à la variable "revenu sujet à imposition" de l'Enquête sur les finances des consommateurs. Les effectifs de population qui seront utilisés pour la post-stratification proviendront des données fiscales de Revenu Canada - Impôt. Les nombres tirés des données fiscales seront rajustés pour tenir compte des problèmes relatifs au champ d'observation qui sont causés par le fait que les particuliers ne remplissent pas tous une déclaration de revenus.

En effectuant la post-stratification selon la mobilité interprovinciale, nous supposons qu'il y a une différence entre le comportement des personnes qui déménagent et celles qui ne déménagent pas. Cette post-stratification vise surtout à améliorer la précision des estimations. Il faut noter qu'elle peut aussi servir à corriger la non-réponse des ménages qui est imputable à des problèmes reliés au dépistage des ménages. La post-stratification selon la mobilité peut être difficile à réaliser si les post-strates sont déterminées en recoupant l'ensemble complet des provinces d'origine et des provinces de destination. Nous pourrions alors penser à établir les post-strates uniquement selon la province de destination. Dans le cas de la post-stratification de la province du Manitoba, par exemple, la seule distinction qui serait faite au niveau des particuliers serait entre les personnes qui viennent de s'installer dans la province (personnes ayant déménagé) et les autres (personnes n'ayant pas déménagé). Cette post-stratification est beaucoup plus simple que le recoupement de l'ensemble complet des provinces d'origine et des provinces de destination, mais n'est peut-être pas aussi efficace si les personnes ont un comportement passablement différent selon leur province d'origine. Les effectifs de population utilisés pour la post-stratification selon la mobilité peuvent être obtenus auprès de la Division de la démographie de Statistique Canada.

Une fois que les variables utilisées pour la post-stratification auront été déterminées, on procédera à la mise en oeuvre en se fondant soit sur l'approche intégrée de Lemaître et Dufour (1987), dont on se sert actuellement pour l'EPA, soit sur les estimateurs par calage sur marges développés par Deville et Särndal (1992). Ces estimateurs tiennent compte de l'approche intégrée de Lemaître et de Dufour (1987), mais permettent d'exercer un certain contrôle sur les poids afin d'éviter les valeurs négatives. L'utilisation des estimateurs par calage sur marges dans le cadre de l'EPA a été étudiée par Stukel et Boyer (1992).

5. CONCLUSION ET PROJETS FUTURS

Dans la présente communication, nous avons examiné deux méthodes servant à pondérer l'échantillon de l'EDTR en vue d'obtenir les meilleures estimations transversales possible. D'après les résultats obtenus, il semble que la méthode du partage des poids constitue le meilleur choix en raison de la simplicité des suppositions qu'elle comporte et de la propriété qu'elle a d'être sans biais.

Nous avons aussi décrit l'utilisation des variables employées pour la post-stratification afin de corriger les problèmes de sous-représentation et d'améliorer la précision des estimations. Des études sont en cours pour déterminer dans quelle mesure le revenu et (ou) la mobilité interprovinciale constituent des variables utiles pour la post-stratification.

REMERCIEMENTS

La présente recherche a été rendue possible grâce à l'aide financière du Fonds global - Recherche et développement de Statistique Canada. Les auteurs profitent de l'occasion pour remercier Hussain G. Choudry et J.N.K. Rao pour leurs observations utiles.

BIBLIOGRAPHIE

Cochran, W.G. (1977). *Sampling Techniques, Third Edition*, John Wiley and Sons, New York.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 418, 376-382.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. *Panel Surveys*, John Wiley and Sons, New York.

Gouriéroux, C., et Roy, G. (1978). Enquête en deux vagues: renouvellement de l'échantillon. *Annales de l'INSEE*, 29, 115-135.

Hunter, L., Michaud, S., et Torrance, V. (1992). Modelling non-response in a longitudinal survey. Discussion présentée à 1992 Conference of the American Statistical Association.

Lavallée, P. (1992). Sample representativity for the survey of labour and income dynamics. Rapport interne de Statistique Canada.

Lemaître, G. (1989). Variance estimation for surveys using the LFS frame weighting system, user documentation.

Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 2, 211-220.

Little, R.J.A. (1986). Survey nonresponse adjustments. *International Statistical Review*, 54, 2, 139-157.

Särndal, C.-E., Swensson, B., et Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Singh, M.P., Drew, J.D., Gambino, J.G., et Mayda, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*, Statistique Canada, Catalogue 71-526.

Stukel, D.M., et Boyer, R. (1992). Calibration estimation: An application to the Canadian Labour Force survey. Document de travail de Statistique Canada SSMD-009E.

SESSION 3

Non-réponse et érosion

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

SUIVI DE L'ENQUÊTE SUR LA SANTÉ DES JEUNES ONTARIENS: ÉVALUATION DES EFFETS DE L'ÉROSION DE L'ÉCHANTILLON (PARTIE II)

M.H. Boyle, B. Wheaton, D.R. Offord, Y.A. Racine et G. Catlin[1]

RÉSUMÉ

Dans la présente communication, on utilise des renseignements provenant d'un suivi, effectué quatre années plus tard, des enfants de 4 à 12 ans ayant participé à l'Enquête sur la santé des jeunes Ontariens (ESJO) en 1983 afin de présenter une procédure à variables multiples pour évaluer les effets de l'érosion de l'échantillon. La procédure comporte quatre parties: (1) modélisation statistique de l'érosion de l'échantillon à l'aide de données du 1er cycle; (2) production de probabilités à partir d'un modèle de prédiction de l'érosion d'un échantillon pour tous les membres de l'échantillon; (3) évaluation du besoin de construire des poids unitaires pour représenter les effectifs perdus pour l'enquête; et (4) évaluation de l'incidence des poids au moyen d'une comparaison des estimations statistiques de la vraisemblance d'une issue défavorable (trouble mental) basées sur des analyses non pondérées et pondérées. Les résultats laissent supposer que les effets de l'érosion de l'échantillon dépendent de variables. L'utilisation de poids a eu une incidence sur certaines des huit variables incluses dans l'analyse, réduisant, pour deux de ces variables, le degré d'association (risque relatif) avec une issue défavorable de 6.87 à 4.35 et de 5.05 à 3.55.

MOTS CLÉS: Érosion de l'échantillon; enquêtes de suivi; troubles de l'enfance.

1. INTRODUCTION

1.1 Description du problème

Les études longitudinales sont vulnérables aux effets défavorables de l'érosion de l'échantillon ou de la perte d'effectif de l'échantillon. Les pertes d'effectif dans le temps peuvent se produire à cause d'un décès, d'une migration, d'une impossibilité de retracer ou d'un refus. La non-participation pour ces raisons augmente invariablement avec chaque période de collecte des données, elle diminue la précision des estimations statistiques et peut mener à des distorsions systématiques ou biais parce que l'échantillon devient non représentatif. On ne doit pas négliger la possibilité que la perte d'effectif invalide les conclusions d'études longitudinales portant sur des enfants. Pour les études longitudinales de troubles mentaux chez les enfants, il n'est pas rare que de 20 à 30% des sujets soient perdus pour l'étude à partir du premier suivi. Cela s'est produit dans le suivi pour l'Île de Wight (Schachar, Rutter et Smith 1981), pour l'étude de suivi dans les quartiers intermédiaires de Manhattan (Gersten, Langner, Eisenberg, Simcha-Fagan et McCarthy 1976) et pour l'Enquête sur la santé des jeunes Ontariens (Offord et coll. 1992). Dans des études de la consommation d'alcools et autres drogues chez les adolescents, des pertes d'effectifs de 20 à 55% lors du suivi ne sont pas rares (Johnston, O'Malley et Eveland 1978; Donovan, Jessor et Jessor 1983; Kandel et Logan 1984; O'Malley, Bachman et Johnston 1984; Brook, Whiteman, Gordon et Cohen 1986; Newcomb et Bentler 1988).

Peu d'études empiriques ont été effectuées pour évaluer les effets des pertes d'effectif sur l'exactitude des estimations statistiques et de la validité des inférences provenant d'études longitudinales menées auprès d'enfants. Généralement, on ne tient pas compte de la question (p. ex., Schachar et coll. 1981) ou l'on s'y attaque en comparant les caractéristiques des participants et des non-participants et en soutenant qu'il n'existe pas d'écarts

[1] M. Boyle, D. Offord et Y. Racine, Department of Psychiatry, McMaster University, Hamilton (Ontario), Canada. B. Wheaton, Department of Sociology, University of Toronto, Toronto (Ontario), Canada. G. Catlin, Division des enquêtes spéciales, Statistique Canada, Parc Tunney, 24-Q, Édifice R.H. Coats, Ottawa (Ontario), Canada.

importants pour les variables étudiées une à la fois (p. ex., Gersten et coll. 1976; Johnson et coll. 1978; Kandel et Logan 1984; Brook et coll. 1986) ou dans diverses combinaisons (p. ex., Newcomb et coll. 1986; Macera, Jackson, Farach et Pate 1988). On peut se demander si ces méthodes sont adéquates pour détecter un biais attribuable à l'érosion de l'échantillon. À l'aide de documents portant sur les essais cliniques, on a pu démontrer que, même avec un échantillonnage aléatoire, des différences qui ne sont pas statiquement significatives dans les caractéristiques de groupes peuvent mener à des estimations biaisées des issues du traitement (Altman 1985).

1.2 Objectifs de la communication

En 1987, une étude de suivi relative à l'Enquête (initiale) sur la santé des jeunes Ontariens (ESJO) a été réalisée afin d'évaluer l'issue de certains troubles de l'enfance, de déterminer les variables qui prédisent la persistance des troubles (pronostic) et d'évaluer le risque de troubles, au fil des ans, parmi les enfants classés, à l'origine, comme ne souffrant pas de troubles mentaux (Offord et coll. 1992). Les pertes d'effectif lors du suivi de l'ESJO ont fourni la motivation nécessaire pour décrire la nature de cette perte et pour en évaluer l'incidence sur diverses estimations statistiques afin de quantifier l'issue, le pronostic et le risque (Boyle, Offord, Racine et Catlin 1991).

La présente communication prolonge, de plusieurs façons importantes, les analyses antérieures de l'érosion de l'échantillon. Tout d'abord, une vaste étendue de données sur les enfants, les parents et la famille sont utilisées pour étudier les écarts entre les participants et les non-participants lors du suivi. Les analyses déjà réalisées se concentraient seulement sur les troubles de l'enfance, le revenu familial et le dysfonctionnement familial. Deuxièmement, des méthodes à variables multiples sont utilisées pour modéliser la participation et pour calculer des poids d'ajustements afin de compenser les écarts entre les participants et les non-participants. Les premières analyses se basaient sur des méthodes bidimensionnelles et sur la stratification pour le calcul des poids d'ajustements. La présente communication commence par une brève description de l'ESJO initiale et de son suivi pour ensuite se concentrer sur l'érosion de l'échantillon.

2. MÉTHODES DE RECHERCHE UTILISÉES DANS L'ESJO

2.1 L'ESJO initiale

La méthodologie utilisée dans l'ESJO initiale est décrite en détail ailleurs (Boyle et coll. 1987). Brièvement, la population cible comprenait tous les enfants nés entre le 1er janvier 1966 et le 1er janvier 1979, dont le lieu de résidence habituel était un logement occupé par un ménage en Ontario. L'unité d'échantillonnage était le logement occupé par un ménage; la base de sondage était le recensement de 1981 et le choix de l'échantillon s'est fait par échantillonnage aléatoire stratifié par grappes à partir du fichier des logements occupés par un ménage produit pour le recensement. La base de sondage excluait seulement trois groupes d'enfants représentant 3.3% de la population des enfants âgés de 4 à 16 ans: les enfants vivant dans des réserves indiennes; ceux qui vivaient dans des logements collectifs, comme les établissements institutionnels et ceux qui habitaient des logements construits après le 1er juin 1981 (jour du recensement). La collecte des données a été effectuée entre janvier et mars 1983. Le taux de participation obtenu était de 91.1% des ménages admissibles et seulement 3.9% de ces ménages ont refusé de participer. Les autres cas de non-participation étaient dûs à des raisons telles que la maladie et l'impossibilité d'entrer en communication avec le ménage.

2.2 Suivi de l'ESJO

Tous les enfants et toutes les familles qui ont participé à l'ESJO initiale étaient admissibles à l'étude de suivi; ces personnes ont été retracées en octobre et en novembre 1986. La Division des enquêtes spéciales de Statistique Canada a recueilli les données sur les participants au suivi en avril et en mai 1987. Les méthodes utilisées pour mesurer la gravité des troubles mentaux chez les enfants de 8 à 16 ans en 1987 étaient celles qui avaient été employées pour les enfants de 4 à 12 ans en 1983. Très brièvement, la mesure de la gravité de chaque trouble mental inclus dans les analyses présentées ici (trouble des conduites, hyperactivité et trouble caractériel) était basée sur les critères de diagnostic figurant dans le DSM - III. La liste de contrôle du comportement des enfants (Child Behaviour Checklist (CBCL) (Achenbach et Edelbrock 1981) a fourni l'ensemble de base des éléments utilisés pour mesurer les critères de diagnostic et des éléments additionnels ont

été créés quand on jugeait que les éléments de la CBCL ne décrivaient pas un critère particulier de façon adéquate. Pour les enfants de moins de 12 ans, des résultats ont été établis à l'aide de listes de contrôle remplies par les parents et les enseignants des enfants pour évaluer les troubles chez l'enfant. Pour les adolescents, enfants de 12 ans et plus, ce sont les parents et les enfants eux-mêmes qui ont complété les listes de contrôle pour évaluer les troubles. Les résultats attribués aux troubles de comportement ont été additionnés afin d'établir certains échelons de classification. Les limites établies pour classifier chaque trouble comme étant présent ou absent étaient les échelons qui établissaient le mieux une distinction entre les diagnostics posés indépendamment par des pédopsychiatres pour un échantillon aléatoire stratifié d'enfants (N = 194) qui avaient participé à l'ESJO initiale. Des détails additionnels sur la classification des troubles sont présentés dans un autre document (Boyle et coll. 1987). Un nouvel instrument a été élaboré afin d'évaluer les troubles mentaux chez les personnes de 17 à 20 ans en 1987. Ce groupe fera l'objet d'un rapport distinct.

2.3 Érosion de l'échantillon lors du suivi de l'ESJO

2.3.1 Portée de l'érosion de l'échantillon

Dans la présente communication, l'évaluation de l'érosion de l'échantillon lors du suivi est limitée aux enfants âgés de 4 à 12 ans dans l'ESJO initiales qui étaient âgés de 8 à 16 ans lors du suivi. Dans la communication, on ne tient pas compte des données sur les évaluations faites par les enseignants, ce qui donne un échantillon plus considérable pour les analyses que celui qui était disponible lors des études de 1983 (étude initiale) et de 1987 (étude de suivi). Quatre-vingt-onze pour cent des ménages avaient accepté de participer à l'ESJO initiale, mais des renseignements manquants sur les questionnaires de 1983 ont réduit l'échantillon disponible pour l'analyse de 2 279 à 1 843 enfants. Lors du suivi, il y avait 1 402 enfants avec données complètes tant pour 1983 que 1987 et 441 enfants avec données complètes en 1983, mais pour lesquels certaines données manquaient en 1987. Les analyses présentées dans cette communication sont limitées aux détériorations de l'échantillon qui se sont produites lors du suivi.

2.3.2 Caractéristiques des participants et des non-participants au suivi

Les participants et les non-participants en 1987 ont été comparés à l'aide de trois ensembles de variables évaluées en 1983:

(1) mesures des troubles mentaux chez l'enfant - le centre d'intérêt principal du suivi de l'ESJO;

(2) les caractéristiques socio-démographiques des enfants et des familles qui pourraient permettre de distinguer les participants des non-participants; et

(3) des variables, dont on sait qu'elles sont en corrélation avec des troubles mentaux chez l'enfant dans des analyses transversales, qui pourraient être examinées en perspective afin d'évaluer leurs risques potentiels.

Les trois ensembles de variables, accompagnés de brèves définitions, sont regroupés au tableau 1, sous les caractéristiques des enfants et des parents/de la famille.

Avant de tester l'importance statistique des écarts entre les participants et les non-participants par rapport aux variables définies au tableau 1, on a examiné les structures d'association entre les mesures ordinales et par intervalles (c.-à-d., nombre de frères et soeurs, dysfonctionnement familial, revenu familial, nombre de confidents et scores négatifs associés aux troubles affectifs) et le statut du participant a été examiné pour trouver des effets non linéaires. Selon la structure d'association observée, les variables ont été codées à nouveau afin de distinguer de façon maximale entre les participants et les non-participants.

Le tableau 2 montre les résultats de la comparaison des participants et des non-participants par rapport aux variables définies au tableau 1. La variable chi-carré (X^2) est utilisée pour tester les écarts statistiquement significatifs entre les groupes. Parmi les vingt comparaisons, il y a dix variables qui permettent de faire une distinction entre les groupes pour une valeur-p inférieure à .05. La grandeur des écarts dépasse 10.0 pour cent pour trois comparaisons: enfants âgés de 9 à 12 ans (46.6 par opposition à 60.3), résidence urbaine (59.3 par opposition à 69.6) et un score de 12 à 20 pour le dysfonctionnement familial (50.5 par opposition à 37.9). La

grandeur des écarts statistiquement significatifs entre les participants et les non-participants par rapport aux sept autres variables ne dépasse pas 7.5 pour cent (tableau 2). Si l'on utilise la signification statistique et la grandeur de l'écart comme critères, rien ne prouve que, pour les troubles mentaux chez l'enfant, il existe une distinction entre les participants et les non-participants (tableau 2).

3. MÉTHODES D'ÉVALUATION DE L'ÉROSION DE L'ÉCHANTILLON

3.1 Renseignements généraux

Les conditions pour qu'il y ait érosion sélective de l'échantillon ont été décrites ailleurs (Greenland 1977; Criqui 1979; Kleinbaum, Morgenstern et Kupper 1981; Boyle et coll. 1991). Brièvement, il y a érosion sélective accompagnée d'un biais lorsque les pertes d'effectif dans les catégories de risque ne sont pas réparties uniformément parmi les catégories d'issue possibles. Le sens et l'importance du biais dépendent de la répartition des pertes d'effectif dans les cellules d'un tableau de classement du risque en fonction des issues possibles. Pour quantifier exactement l'importance du biais attribuable à l'érosion sélective de l'échantillon, il faut disposer de renseignements sur l'issue pour les non-participants, obtenus dans une étude distincte. Comme les études de ce genre sont dispendieuses et difficiles à entreprendre, les chercheurs disposent rarement de données sur l'issue relatives aux non-répondants.

Tableau 1: Définition des variables.

Enfant

Trouble de comportement: Caractérisé par de la violence physique contre les personnes ou les biens et (ou) par une infraction grave aux normes sociales.

Hyperactivité: Caractérisée par l'inattention, l'impulsivité et l'activité motrice.

Trouble affectif: Caractérisé surtout par des sentiments d'anxiété et de dépression.

Au moins un trouble: Au moins un trouble parmi les suivants: trouble de comportement, trouble entraînant un déficit de la capacité d'attention et trouble affectif.

Difficultés dans les relations interpersonnelles: Un enfant a été classé comme ayant des difficultés dans les relations interpersonnelles si l'un des parents a indiqué, sur une échelle de cinq points, que l'enfant ne s'entendait «pas très bien, souvent des problèmes», option de réponse 4, ou que l'enfant ne s'entendait «pas bien du tout, toujours des problèmes», option de réponse 5, dans une ou plus des trois circonstances suivantes: avec d'autres enfants comme ses amis ou ses camarades de classes, avec les enseignants à l'école ou avec les membres de sa famille.

Requiert une aide professionnelle: Un enfant a été classé comme requérant une aide professionnelle si l'un des parents a répondu de façon positive à deux questions: «Au cours des six derniers mois, pensez-vous qu'il(elle) a eu des problèmes émotifs ou de comportement?» et «Croyez-vous qu'il(elle) a ou avait besoin de l'aide d'un professionnel pour régler ces problèmes?».

Échec scolaire antérieur: L'un des parents a déclaré que l'enfant a échoué ou redoublé une année à un moment quelconque pendant ses études.

Incapacités fonctionnelles: L'un des parents a déclaré que, depuis au moins six mois, l'enfant souffrait d'une incapacité sur le plan de l'activité physique, de la mobilité, pour assurer son entretien personnel ou qu'il a de la difficulté à jouer son rôle.

Parents/famille:

Famille monoparentale: Il n'y a que le père ou la mère qui habite le domicile.

Résidence urbaine: La résidence est située dans une région urbaine comptant au moins 1 000 habitants avec une densité de 400 habitants ou plus au kilomètre carré (Statistique Canada 1982).

Famille nombreuse: La famille compte au moins quatre frères ou soeurs âgés de 4 à 16 ans. Pour comparer les participants et les non-participants, le nombre de frères et de soeurs a été recodé à deux niveaux: (1) au moins trois frères ou soeurs et (2) 0,1 ou 2 frères ou soeurs.

Domicile surpeuplé: Un rapport entre le nombre de personnes et le nombre de pièces $\geq$ 1.0.

Famille mobile: La famille a déménagé au moins deux fois au cours des deux dernières années. Pour comparer les participants et les non-participants, la mobilité a été recodée à deux niveaux: (1) au moins trois déménagements et (2) 0,1 ou 2 déménagements.

Dysfonctionnement familial: D'après les renseignements fournis par l'un des parents, cette personne a obtenu un score compris entre 27 et 48 (parmi les valeurs possibles de 12 à 48) sur l'échelle générale de fonctionnement à 12 questions tirée de l'instrument d'évaluation du fonctionnement de la famille de McMaster (McMaster Family Fonctioning Assessment Device) (Byles, Byrne, Boyle et Offord 1988). Le fonctionnement de la famille est évalué par rapport à six dimensions: solution de problèmes, communication, rôles, capacité de répondre aux besoins affectifs des membres, participation à la vie affective de la famille et encadrement. Pour comparer les participants et les non-participants, on a recodé le dysfonctionnement familial à trois niveaux: (1) scores de 12 à 20; (2) scores de 21 à 25 et (3) scores supérieurs à 25.

Troubles nerveux chez les parents: L'un des parents déclare avoir été traité ou que son conjoint a été traité, à un moment quelconque, pour troubles nerveux.

Revenu familial < $10 000: Le revenu familial total au cours de l'année précédant l'enquête (1982) était < $10 000. Pour comparer les participants et les non-participants, nous avons recodé le revenu familial à trois niveaux: (1) moins de $10 000, (2) de $10 000 à $39 999 et (3) $40 000 ou plus.

Absence de confident: L'un des parents déclare n'avoir aucune personne à qui confier ses difficultés ou ses problèmes personnels. Pour comparer les participants et les non-participants, nous avons recodé le nombre de confidents à deux niveaux: (1) de 1 à 5 confidents et (2) 0,6 ou 7 confidents.

Score négatif associé aux troubles affectifs: L'un des parents a obtenu un score de 5 ou plus (gamme de valeurs allant de 0 à 10) pour l'échelle négative des troubles affectifs à 5 questions élaborée par Bradburn.

Bien que des données sur l'issue ne soient pas disponibles pour les non-participants lors des études de suivi, tous les renseignements recueillis lors des évaluations antérieures peuvent être utilisés pour distinguer les caractéristiques spéciales des non-participants qui pourraient rendre l'échantillon non représentatif. Cette évaluation, effectuée une variable à la fois, est présentée au tableau 2 et on l'a mentionnée dans la section précédente. Compte tenu du fait qu'un certain nombre de variables du tableau 2 qui permettent de faire une distinction entre les participants et les non-participants lors du suivi peuvent aussi être des facteurs de risque pour les troubles chez l'enfant, il existe des raisons de supposer qu'il y a eu érosion sélective de l'échantillon et introduction d'un biais.

Tableau 2: Répartition en pourcentage des caractéristiques de 1983 selon l'état pour le suivi en 1987.

	État pour le suivi en 1987			
	Participants	Non-participants		
Caractéristiques de 1983	N = 1 402	N = 441	X^2(dl)	valeur p
Enfant				
De sexe masculin	51.1	50.6	0.02(1)	ND
Âgé de 9 à 12 ans	46.6	60.3	24.79(1)	0.00
Au moins un trouble	6.9	5.7	0.66(1)	ND
Troubles de comportement	0.9	1.4	0.27(1)	ND
Hyperactivité	1.9	2.0	0.00(1)	ND
Troubles affectifs	5.5	4.8	0.23(1)	ND
Difficultés dans les relations interpersonnelles	2.4	5.0	6.64(1)	0.01
Requiert une aide professionnelle	2.4	4.1	2.78(1)	ND
Échec scolaire antérieur	7.3	10.7	4.49(1)	0.04
Incapacités fonctionnelles	4.0	5.9	2.42(1)	ND
Parents/famille				
Famille monoparentale	8.9	8.6	0.01(1)	ND
Résidence urbaine	59.3	69.6	14.76(1)	0.00
Au moins 3 frères et soeurs	29.3	35.8	6.37(1)	0.02
Domicile surpeuplé	13.7	17.7	3.96(1)	0.05
Au moins 3 déménagements	1.7	3.6	4.93(1)	0.03
Dysfonctionnement familial				
(1) score entre 12 et 20	50.5	37.9	26.65(2)	0.00
(2) score entre 21 et 25	37.2	42.4		
(3) score > 25	12.3	19.7		
L'un des parents (ou les deux) a (ont) été traité(s) pour troubles nerveux	21.5	19.3	0.91(1)	ND
Revenu familial				
(1) <$10 000	5.7	9.3	10.70(2)	0.01
(2) $10 000-$39 999	66.3	68.3		
(3) $40 000 ou plus	28.0	22.4		
De 1 à 5 confidents	11.8	18.8	13.40(1)	0.00
Score négatif associé aux troubles affectifs	13.1	17.9	6.11	0.02

3.2 Corrections par pondération

Une méthode recommandée pour faire un test afin de savoir s'il y a eu ou non érosion sélective de l'échantillon et introduction d'un biais consiste à élaborer des poids d'ajustement visant à compenser les érosions de l'échantillon beaucoup trop élevées parmi les répondants que l'on considère à risque d'avoir des issues défavorables. L'élaboration de poids d'ajustements basés sur des analyses bidimensionnelles a été présentée dans un article déjà publié (Boyle et coll. 1991). La méthode utilisée ici est basée sur des procédures multidimensionnelles et s'inspire du travail de Aneshensel, Becerra, Fielder et Schulrer (1989). Voici les étapes qui sont suivies:

(1) estimation d'une équation de régression logistique basée sur les données de l'évaluation de 1983 afin de prédire la perte d'effectif dans l'échantillon en 1987 sous forme de variable binaire;

(2) utilisation de l'équation pour produire des probabilités d'érosion pour toutes les personnes dans l'échantillon de suivi;

(3) évaluation du besoin de stratifier les répondants selon la probabilité de perte d'effectif lors du suivi et d'élaborer des poids qui reflètent l'importance de l'érosion de l'échantillon dans chaque strate; et

(4) élaboration de poids pour les répondants, définis comme étant le rapport entre le nombre de répondants en 1983 et le nombre de répondants en 1987 pour chaque strate définie par la probabilité de perte d'effectif lors du suivi en 1987.

3.2.1 Étapes 1 et 2

Nous avons utilisé des analyses de régression logistique multiple ascendante, établies à l'aide du logiciel SPSS, pour construire une équation afin de prédire la perte d'effectif dans l'échantillon en 1987 à partir des données de l'évaluation de 1983. Les variables étudiées étaient celles qui, dans le tableau 2, ont un critère X^2 supérieur à 1.50. En plus des variables choisies pour l'évaluation, trois interactions ont été précisées: dysfonctionnement familial et revenu, domicile surpeuplé et 3 déménagements ou plus et de 1 à 5 confidents et score négatif associés aux troubles affectifs. Les critères statistiques utilisés pour élaborer le modèle comprenaient une probabilité pour l'inclusion d'une variable fixée à 0.10 et une probabilité pour l'exclusion d'une variable fixée à 0.15. On a utilisé des critères statistiques larges afin de maximiser l'exactitude prédictive du modèle. Le modèle final comprenait les effets principaux suivants: description de l'enfant: âgé de 9 à 12 ans et qui éprouve des difficultés au niveau des relations interpersonnelles; description des parents/de la famille: résidence urbaine, au moins trois frères ou soeurs, au moins trois déménagements et dysfonctionnement familial et deux interactions: dysfonctionnement familial et revenu et de 1 à 5 confidents et score négatif au chapitre de l'affect. À l'aide des coefficients de régression et des valeurs observées pour les caractéristiques de 1983, on peut estimer la probabilité d'érosion pour tout répondant en 1983.

3.2.2 Étape 3

L'étape 3 comporte une évaluation du besoin de stratifier les répondants selon leur probabilité d'érosion lors du suivi et de construire des poids qui reflètent l'importance de la diminution de l'échantillon dans chaque strate. Les preuves pertinentes pour cette évaluation comprennent:

(1) la mesure dans laquelle la probabilité d'érosion lors du suivi permet d'effectuer une distinction entre les participants et les non-participants en 1987; et

(2) la mesure dans laquelle les estimations du risque d'issue défavorable en 1987 parmi les participants varient selon la probabilité d'érosion.

Ces deux conditions doivent exister, dans une certaine mesure, pour que tout système de pondération basé sur les probabilités de diminution de l'échantillon ait une incidence.

Pour examiner la mesure dans laquelle la probabilité d'érosion lors du suivi permet de faire une distinction entre les participants et les non-participants en 1987, nous avons effectué un test t qui portait sur les probabilités d'érosion estimées à partir de la régression logistique calculée à l'étape 2. Pour déterminer la mesure dans laquelle les estimations du risque d'issue défavorable en 1987 sont modifiées par la probabilité d'érosion, les répondants du suivi ont été divisés en deux strates selon leur probabilité d'érosion: probabilité élevée, définie comme une probabilité d'érosion >25% et faible probabilité, définie comme une probabilité d'érosion ≤25%. Nous avons ensuite calculé des estimations, propres à chaque strate, du degré d'association (risque relatif) entre les facteurs de risque potentiel évalués en 1983 et au moins un trouble mental évalué en 1987. Nous avons utilisé le test d'homogénéité de la variable X^2 avec un degré de liberté (voir Schesselman 1982), pour évaluer si les estimations propres à chaque strate étaient significativement différentes entre elles.

Tel que prévu, les non-participants en 1987 avaient une probabilité d'érosion plus élevée (M = 0.281) que les participants (M = 0.226), t(1 841) = 10.37, p<.000. De plus, il existe des données qui montrent que, dans certains cas, le risque d'issue défavorable est modifié par la probabilité d'érosion.

Le tableau 3 renferme les risques relatifs propres aux strates, entre les facteurs de risque potentiel évalués en 1983 et au moins un trouble évalué en 1987. Pour deux des variables - «difficultés dans les relations interpersonnelles» et «incapacités fonctionnelles» - le risque relatif d'une issue défavorable varie selon la probabilité d'érosion. Par exemple, la variable «difficultés dans les relations interpersonnelles» évaluée en 1983 est une variable de prédiction pour la variable «au moins un trouble» évaluée en 1987. Parmi les personnes avec

une faible probabilité d'érosion, le risque relatif est de 19.57; parmi les personnes avec une probabilité d'érosion élevée, le risque relatif est de 3.12. De même, la variable «incapacités fonctionnelles» évaluée en 1983 est une bonne variable de prédiction pour la variable «au moins un trouble» en 1987 parmi les personnes avec une faible probabilité d'érosion (risque relatif de 6.02), mais pas parmi les personnes avec une probabilité d'érosion élevée (risque relatif de 0.72).

3.2.3 Étape 4

Au cours de l'étape 4, on élabore des poids pour les répondants lors du suivi. Afin de produire ces poids, on stratifie tous les répondants et non-répondants lors du suivi selon leur probabilité d'érosion - une estimation produite plus tôt à l'aide du modèle de régression logistique. On additionne les répondants et les non-répondants dans chaque strate et ce total est divisé par le nombre de répondants afin d'obtenir un poids unitaire dans chaque strate. Ce poids unitaire est alors divisé par une constante afin que la somme des poids soit égale à la taille de l'échantillon lors du suivi: 1 402.

Tableau 3: Risque relatif entre les facteurs de risque potentiel évalués en 1983 et au moins un trouble évalué en 1987, selon la probabilité d'attrition en 1987.

Facteurs de risque potentiel en 1983	Probabilité d'attrition 25% ou moins N = 963	>25% N = 439	Homogénéité X^2(1dl)	valeur p
Enfant				
Au moins un trouble	9.03***	8.30***	0.03	ND
Difficultés dans les relations interpersonnelles	19.57***	3.12	5.00	.05
Requiert une aide professionnelle	4.80*	4.75**	0.00	ND
Échec scolaire antérieur	1.22	1.47	0.07	ND
Incapacités fonctionnelles	6.02***	0.72	5.80	.02
Parents/famille				
Dysfonctionnement familial	2.55	2.67*	0.01	ND
L'un des parents (ou les deux) a (ont) été traité(s) pour troubles nerveux	1.73	1.55	0.05	ND
Score négatif associé aux troubles affectifs	3.20*	1.48	2.31	ND

Tableau 4: Risque relatif non pondéré par opposition à pondéré entre les facteurs de risque potentiel en 1983 et la variable «au moins un trouble» en 1987.

Facteurs de risque potentiel en 1983	Risque relatif Non pondéré (1)	Pondéré (2)	▲ (1)-(2)
Enfant			
Au moins un trouble	8.96***	8.35***	0.61
Difficultés dans les relations interpersonnelles	6.87***	4.35***	2.52
Requiert une aide professionnelle	5.05***	3.55**	1.50
Échec scolaire antérieur	1.44	1.63	-0.19
Incapacités fonctionnelles	2.64*	2.49*	0.15
Parents/famille			
Dysfonctionnement familial	2.82*	2.72***	0.10
L'un des parents (ou les deux) a (ont) été traité(s) pour troubles nerveux	1.71*	1.65*	0.06
Score négatif associé aux troubles affectifs	2.27**	2.46***	-0.19

* p <.05, **p <.01, ***p <.001

▲ risque relatif non pondéré moins risque relatif pondéré

Le tableau 4 montre ce qui arrive au degré d'association entre les facteurs de risque potentiel évalués en 1983 et la variable «au moins un trouble» évaluée en 1987 quand les poids sont appliqués à l'échantillon du suivi. Les

écarts entre les estimations non pondérées et pondérées du risque relatif vont de -0.19 («échec scolaire antérieur», «score négatif au chapitre de l'affect») à 2.52 («difficultés dans les relations interpersonnelles»). Selon les données du tableau 4, le sens prédominant du biais attribuable à l'érosion de l'échantillon s'éloigne de zéro et est plutôt considérable pour deux des variables étudiées (c.-à-d., «difficultés dans les relations interpersonnelles» et «requiert une aide professionnelle»).

4. DISCUSSION

Bien que la recherche longitudinale offre des possibilités d'accroître nos connaissances des troubles mentaux chez l'enfant, des considérations d'ordre méthodologique, comme l'érosion de l'échantillon, sont des facteurs importants pour déterminer l'utilité des données. Dans la présente communication on s'est concentré sur une procédure à variables multiples pour évaluer les effets de l'érosion de l'échantillon lors du suivi afin de déterminer s'il s'était produit ou non une perte sélective et un biais. La procédure comportait (1) la modélisation statistique de l'érosion de l'échantillon à l'aide des données existantes; (2) la production de probabilités à partir d'un modèle de prédiction de l'érosion pour tous les membres de l'échantillon; (3) l'évaluation du besoin d'élaborer des poids unitaires et (4) l'évaluation de l'incidence des poids (si nécessaire) par la comparaison d'analyses non pondérées et pondérées. Dans la présente étude, les données disponibles laissaient supposer que l'élaboration de poids était justifiée. L'évaluation de l'incidence des poids montre que cela n'était pas uniforme. Pour deux des variables relatives aux enfants étudiées - «difficultés dans les relations interpersonnelles» et «requiert une aide professionnelle» - les écarts dans le risque relatif étaient de 2.52 et de 1.50, respectivement. Le sens du biais s'éloignait de zéro. Pour les autres variables étudiées, l'utilisation de poids d'ajustements n'a eu que peu d'incidence sur les estimations statistiques.

La procédure décrite dans la présente communication pour évaluer l'érosion de l'échantillon lors d'études de suivi est relativement simple à utiliser; cependant, elle donne une analyse beaucoup plus complète de l'érosion de l'échantillon que de simples comparaisons des répondants et des non-répondants pour des variables pertinentes. Il est toutefois important de remarquer que les procédures donnent de meilleurs résultats quand le modèle statistique utilisé pour prédire l'érosion de l'échantillon est spécifié correctement. Si, pour toutes les variables pertinentes, on ne peut faire de distinction entre les répondants et les non-répondants, il n'existe alors rien sur quoi on peut se baser pour prédire leur probabilité d'érosion. De plus, on suppose que les résultats du suivi seront les mêmes pour les répondants et les non-répondants qui ont la même probabilité d'érosion.

Comme nous nous tournons de plus en plus vers la recherche longitudinale pour trouver des réponses à propos de la nature de la psychopathologie infantile, les méthodes utilisées pour évaluer l'incidence de l'érosion de l'échantillon lors d'un suivi ainsi que les corrections appropriées à ce problème devraient prendre une plus grande importance. La pondération utilisée comme procédure pour évaluer l'incidence de l'érosion de l'échantillon est assez prometteuse comme moyen de déterminer si l'érosion de l'échantillon a introduit ou non un biais dans une analyse. Il faut effectuer d'autres travaux pour déterminer les conditions et les circonstances dans lesquelles l'utilisation de poids d'ajustement compense effectivement les distorsions introduites par l'érosion sélective de l'échantillon.

REMERCIEMENTS

Ce travail a été appuyé par des fonds provenant du Programme national de recherche et de développement en matière de santé (subvention numéro 6606-3760-42) et par le ministère des Services sociaux et communautaires de l'Ontario et a été réalisé par la Child Epidemiology Unit, Department of Psychiatry, McMaster University et par le Child and Family Centre, Chedoke-McMaster Hospitals, Hamilton (Ontario). Le Dr Boyle reçoit un William T. Grant Foundation Faculty Scholar Award et le Dr Offord une bourse de chercheur émérite - Santé nationale de Santé et Bien-être social Canada.

BIBLIOGRAPHIE

Achenbach, T., et Edelbrock, C. (1981). Behavioral problems and competences reported by parents of normal and disturbed children aged 4 through 16. *Monographs of the Society for Research in Child Development*, 46, 188.

Altman, D.G. (1985). Comparability of randomized groups. *The Statistician*, 34, 125-136.

Aneshensel, C.S., Becerra, R.M., Fielder, E.P., et Schuler, R.H. (1989). Participation of Mexican American female adolescents in a longitudinal panel survey. *Public Opinion Quarterly*, 53, 548-562.

Boyle, M.H., Offord, D.R., Hofmann, H.F., Catlin, G.P., Byles, J.A., Cadman, D.T., Crawford, J.W., Links, P.S., Rae-Grant, N.I., et Szatmari, P. (1987). Ontario Child Health Study: I. Methodology. *Archives of General Psychiatry*, 44, 826-831.

Boyle, M.H., Offord, D.R., Racine, Y.A., et Catlin, G.P. (1991). Ontario Child Health Study Follow-up: evaluation of sample loss. *Journal of the American Academy of Child and Adolescent Psychiatry*, 30, 449-456.

Brook, J.S., Whiteman, M., Gordon, A.S., et Cohen, P. (1986). Dynamics of childhood and adolescent personality traits and adolescent drug use. *Developmental Psychology*, 22, 403-414.

Byles, J.A., Byrne, C., Boyle, M.H., et Offord D.R. (1988). Ontario Child Health Study: Reliability and validity of the general functioning subscale of the McMaster Family Assessment Device. *Family Process*, 27, 97-104.

Criqui, M.H. (1979). Response bias and risk ratios in epidemiologic studies. *American Journal of Epidemiology*, 109, 344-399.

Donovan, J.E., Jessor, R., et Jessor, L. (1983). Problem drinking in adolescence and young adulthood: A follow-up study. *Journal of Studies of Alcohol*, 44, 109-137.

Gersten, J.C., Langner, T.S., Eisenberg, J.G., Simcha-Fagan, O., et McCarthy, E.D. (1976). Stability and change in types of behavioral disturbance of children and adolescents. *Journal of Abnormal Child Psychology*, 4, 111-127.

Greenland, S. (1977). Response and follow-up bias in cohort studies. *American Journal of Epidemiology*, 106, 183-187.

Johnston, L.D., O'Malley, P.M., et Eveland, L.K. (1978). Drugs and delinquency: A search for causal connections. Dans D.B. Kandel (Éd.). *Longitudinal research on drug use: empirical findings and methodology issues*. Washington, DC: Hemisphere-Wiley, 132-156.

Kandel, D.B., et Logan, J.A. (1984). Patterns of drug use from adolescence to young adulthood, I: Periods of risk for initiation, continued use and discontinuation. *American Journal of Public Health*, 74, 660-666.

Kleinbaum, D.G., Morgenstern, H., et Kupper, L.L. (1981). Selection bias in epidemiologic studies. *American Journal of Epidemiology*, 113, 452-463.

Macera, C.A., Jackson, K.L., Farach, C., et Pate, R.R. (1988). The use of proportional hazards regression in investigating dropout rates in a longitudinal study. *Journal of Clinical Epidemiology*, 41, 1175-1180.

Newcomb, M.D., et Bentler, P.M. (1988). Impact of adolescent drug use and social support on problems of young adults: A longitudinal study. *Journal of Abnormal Psychology*, 97, 64-75.

Newcomb, M.D., Maddahian, E., et Bentler, P.M. (1986). Risk factors for drug use among adolescents: concurrent and longitudinal analyses. *American Journal of Public Health*, 76, 525-531.

Offord, D.R., Boyle, M.H., Racine, Y.A., Fleming, J.A., Cadman, D.T., Munroe Blum, H., Byrne, C., Links, P.S., Lipman, E.L., MacMillan, H.C., Rae-Grant, N.I., Sanford, M.N., Szatmari, P., Thomas, H., et Wordward, C.A. (1992). Outcome, prognosis and risk in a longitudinal follow-up study. *Journal of the American Academy of Child and Adolescent Psychiatry*, 31, 916-923.

O'Malley, P.M., Bachman, J.G., et Johnston, L.D. (1984). Period, age and cohort effects on substance use among American youth, 1976-82. *American Journal of Public Health*, 74, 682-688.

Schachar, R., Rutter, M., et Smith, A. (1981). The characteristics of situationally and pervasively hyperactive children: Implications for syndrome definition. *Journal of Child Psychology and Psychiatry*, 22, 375-392.

Schlesselman, J.J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford: Oxford University Press.

Statistique Canada (1982). Dictionnaire du recensement de 1981 (Catalogue n° 99-901). Ottawa. Ministère des approvisionnements et services.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

STRATÉGIE POUR MINIMISER L'IMPACT DE LA NON-RÉPONSE DANS L'ENQUÊTE SUR LA DYNAMIQUE DU TRAVAIL ET DU REVENU

S. Michaud et L. Hunter[1]

RÉSUMÉ

Statistique Canada lancera en 1994 une importante enquête par panel s'adressant aux ménages. Dans le cadre de l'enquête sur la dynamique du travail et du revenu (EDTR), les répondants seront interviewés pendant une période de six ans, à raison de deux interviews par année. La non-réponse et l'érosion de l'échantillon sont considérées comme deux aspects cruciaux à examiner dans la planification de l'enquête. La présente communication examine des méthodes qui sont actuellement évaluées en vue de compenser la non-réponse dans l'EDTR: une méthode de modélisation permettant de compenser la non-réponse par un rajustement de pondération, et l'imputation des vagues manquantes de données.

MOTS-CLÉS: Non-réponse; enquête longitudinale; modélisation; imputation.

1. INTRODUCTION

Statistique Canada lancera en 1994 une importante enquête par panel menée auprès des ménages, appelée enquête sur la dynamique du travail et du revenu (EDTR). Dans cette enquête, les personnes et les familles seront suivies pendant six ans, et l'on recueillera des renseignements sur leurs expériences du marché du travail, leur revenu et la situation familiale. L'EDTR s'appuie sur de solides fondements à Statistique Canada. Ses origines se situent dans plusieurs enquêtes, notamment l'enquête sur la population active (EPA), l'enquête sur les finances des consommateurs (EFC) et l'enquête sur l'activité. L'EPA et l'EFC sont deux enquêtes transversales. À ce titre, elles offrent une série de clichés et sont des outils utiles et efficaces pour suivre des tendances à l'échelle globale. L'enquête sur l'activité est à la fois une enquête longitudinale et une enquête transversale. Deux panels ont été traités jusqu'ici, soit un panel de deux ans (1986-1987) et un panel de trois ans (1988-1990). Pour chaque panel longitudinal, les personnes participant à la première vague étaient interviewées et l'ont veillait à garder leurs coordonnées. Toutes les personnes vivant avec ces personnes, lors des vagues subséquentes, étaient aussi interviewées (mais ne faisaient pas l'objet du même suivi).

Puisque l'EDTR comporte des interviews des mêmes personnes pendant six ans, à raison de deux interviews par année, les taux de non-réponse et l'érosion de l'échantillon sont considérés comme des aspects cruciaux à examiner dans la planification de l'enquête. Des études sont actuellement menées sur différents aspects de la non-réponse touchant l'enquête sur l'activité, dans l'espoir de trouver des approches qui minimiseront les répercussions de la non-réponse sur les données de l'EDTR. Dans la présente communication, nous examinerons certaines des études qui sont actuellement réalisées afin de minimiser l'impact de la non-réponse; ces études portent sur (a) la possibilité d'ajuster un modèle ou des modèles permettant de compenser la non-réponse par un rajustement de pondération et (b) des essais de méthodes d'imputation pour compenser une vague de données manquante. Dans la section 2, nous donnerons d'abord plus de détails sur le plan d'enquête de l'EDTR; nous consacrerons la section 3 aux modèles examinés en vue d'apporter des rajustements de pondération aux fichiers longitudinaux; la section 4 présentera la stratégie d'imputation, tandis que la section 5 exposera d'autres plans aptes à contribuer au traitement de la non-réponse, ainsi que les futurs travaux envisagés.

[1] S. Michaud et L. Hunter, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6.

2. PLAN D'ENQUÊTE DE L'EDTR

Bon nombre d'enquêtes longitudinales ont été effectuées aux États-Unis et dans d'autres pays. Lepkowski a mentionné des facteurs qui pourraient influer sur la stratégie de traitement de la non-réponse dans un contexte longitudinal: la forme de l'analyse (données recueillies surtout pour dresser des comparaisons longitudinales, ou pour établir des estimations transversales, ou encore pour examiner une accumulation longitudinale de données); le type de données recueillies (continues, par catégories ou conditionnelles); et le profil de non-réponse de vague (érosion et non-érosion). L'EDTR comporte certaines autres caractéristiques qui influeront sur la formulation de la stratégie: les données de l'EDTR seront recueillies dans des interviews différées; l'EDTR comporte deux unités d'analyse, c.-à-d. que les données doivent être établies tant au niveau individuel qu'au niveau des ménages; et, enfin, une certaine partie des données recueillies se chevauchent légèrement à la lisière.

2.1 Données recueillies et interviews différées

Comme il a été mentionné ci-dessus, l'EDTR est une enquête longitudinale. Un échantillon de ménages sera prélevé en janvier 1993 à même l'échantillon de l'enquête sur la population active. À compter de janvier 1994, les personnes sélectionnées dans l'échantillon de l'enquête sur la population active de 1993 seront interviewées deux fois par année pendant six ans. Il y aura une interview en janvier, qui visera à recueillir de l'information sur l'activité de l'année précédente (aussi appelée année de référence). L'interview permettra d'obtenir des données comme les dates de début et de fin des emplois occupés, des détails concernant jusqu'à trois employeurs, les périodes sans travail et les absences du travail, ainsi que certains renseignements sur l'incapacité. L'interview de mai visera à recueillir des renseignements sur le revenu et certaines données sur la richesse, pour les personnes qui se trouvaient dans l'échantillon de janvier. L'information sur le revenu concernera la même période de référence que l'information sur l'activité (c.-à-d. l'année civile précédente). Il sera plus facile, croit-on, d'obtenir des données sur le revenu vers le mois de mai, au moment où les gens remplissent leur déclaration de revenu. L'interview de mai peut être considérée comme une interview différée se rattachant à l'échantillon de janvier, puisque l'information est recueillie auprès des mêmes personnes, pour la même période de référence. Un fichier réunissant l'information sur l'activité et le revenu sera diffusé chaque année. Le principe de l'interview différée, dans laquelle on recueille une information différente de celle de l'autre interview, ajoute une nouvelle dimension à la définition de la non-réponse. Habituellement, on considère comme une non-réponse complète le fait de ne pas répondre à une interview. Dans le cas où une interview différée est utilisée, même s'il manque une interview, l'information recueillie ne concerne qu'une partie des données diffusées annuellement. Par conséquent, une interview manquante peut être considérée comme une non-réponse partielle.

2.2 Unités d'analyse de l'EDTR

L'EDTR comportera des interviews de personnes, et une grande quantité d'information sera recueillie et analysée au niveau individuel. Toutefois, les mesures relatives à la pauvreté et certaines études sur le revenu exigent de l'information à l'échelon des ménages. Puisque les ménages ne sont pas une unité stable dans le temps, il a été décidé que l'EDTR suivrait des personnes (plutôt que de tenter de suivre des ménages). Des caractéristiques relatives aux ménages et aux familles économiques seront établies tous les ans en janvier et seront rattachées aux personnes à titre de caractéristiques de ces dernières (p. ex. le répondant vit dans une famille de trois personnes, dont le revenu du ménage est «x»). Ainsi, la non-réponse peut être traitée différemment, si au moins une personne du ménage répond, comparativement au cas où le ménage au complet ne répondrait pas.

2.3 Chevauchements entre les périodes de collecte

Dans la composante de l'EDTR relative à l'activité, les données sont recueillies pour la période allant du premier janvier de l'année de référence jusqu'à la date de l'interview (qui peut avoir lieu aussi tard qu'en février de l'année suivante). Il y aura donc un chevauchement d'environ un mois entre des vagues consécutives d'interviews relatives à l'activité. On disposera ainsi d'outils additionnels pour l'imputation.

2.4 Non-réponse

Pour les fins de l'analyse d'un échantillon longitudinal, Kalton a divisé la non-réponse en deux catégories: l'érosion et la non-érosion. On dit qu'il y a érosion lorsqu'une personne ayant répondu à une ou plusieurs vagues

au début de l'enquête ne répond plus à aucune vague subséquente. Par exemple, dans un panel de trois ans, les personnes qui ne répondraient ni à la deuxième, ni à la troisième vague constitueraient des cas d'érosion. Les cas de non-érosion sont ceux de personnes qui ne répondent pas à une ou plusieurs vagues, mais qui reprennent plus tard leur participation à l'enquête.

L'examen du panel de trois ans de l'enquête sur l'activité révèle qu'une proportion de 16% de la non-réponse était due à l'érosion, tandis qu'une fraction de 4% était formée de cas de non-érosion. Les observations faites dans le cadre d'autres enquêtes montrent que le rapport entre les cas d'érosion et les cas de non-érosion diminue avec le temps (un panel de longue durée a un rapport non-érosion/érosion plus élevé qu'un panel de courte durée).

L'évaluation du fichier de l'enquête sur l'activité a permis de constater que la non-réponse variait beaucoup entre certains groupes:

- Les personnes ayant déménagé affichaient un taux de non-réponse (incluant les cas de répondants introuvables) de près de 20%, tandis que la non-réponse des personnes n'ayant pas déménagé était d'environ 2%. La variable déménagement/non-déménagement est la caractéristique causant, de loin, la variation la plus forte,
- sur la base des caractéristiques observées à la vague 1, la non-réponse était supérieure parmi les personnes qui étaient sans emploi à la vague 1,
- sur la même base, la non-réponse était plus élevée pour le groupe de personnes qui n'étaient pas mariées à la vague 1,
- les personnes qui vivaient en région urbaine à la vague 1 affichaient aussi un taux de non-réponse plus élevé après trois ans,
- les cas de non-érosion semblent attribuables à des personnes ayant des types d'emplois différents de ceux des répondants.

Les différences de caractéristiques entre les répondants et les non-répondants laissent entrevoir quelques possibilités. Premièrement, le rajustement actuel visant à compenser la non-réponse, qui utilise uniquement de l'information relative au plan de sondage (essentiellement certaines données géographiques), n'est peut-être pas le meilleur qui soit. Une technique qui tiendrait compte d'autres caractéristiques en se fondant sur les données d'années antérieures pourrait donner de meilleurs résultats. Deuxièmement, on dispose de beaucoup plus d'information dans le cas de la non-réponse formée de cas de non-érosion. Puisqu'il semble y avoir de nombreuses différences dans les types d'emplois occupés par les répondants et les non-répondants, l'imputation pourrait être un meilleur outil de rajustement visant à compenser la non-réponse, notamment dans les cas de non-érosion. Si la collecte des données se fait sur une longue période, il pourrait être plus avantageux d'imputer les réponses dans les cas de non-érosion que de supprimer l'enregistrement et de faire un rajustement de pondération tenant compte de la non-réponse, en particulier s'il y a peu de vagues manquantes. Toutefois, il a été constaté dans d'autres enquêtes que l'imputation longitudinale peut être un processus difficile, vaste et coûteux. Afin de déterminer la meilleure stratégie de traitement de la non-réponse dans l'EDTR, certaines études d'évaluation ont été entreprises.

3. MODÈLES

Deux approches possibles ont été envisagées pour la compensation de la non-réponse au moyen de rajustements de pondération: rajustements s'appuyant sur des quotients propres à des sous-groupes de la population, et détermination de modèles de régression. L'approche fondée sur les modèles a été retenue parce qu'on croyait que les travaux ainsi réalisés pourraient également servir à d'autres fins. En particulier, il y a deux usages possibles d'un modèle de non-réponse dans le contexte de notre étude longitudinale: la prévision de la non-réponse et l'obtention d'un rajustement de pondération tenant compte de la non-réponse. Bien qu'il soit probable qu'un seul et même modèle ne puisse pas convenir à ces deux usages, nous espérions qu'un ensemble de base de variables communes aux divers modèles pourrait être déterminé, tandis qu'un petit groupe de variables additionnelles répondraient aux besoins propres à chacun des modèles. Par exemple, la caractéristique la plus corrélée à la non-réponse est la variable «a déménagé/n'a pas déménagé depuis la dernière interview» (la non-réponse étant en l'occurrence attribuable à l'impossibilité de retrouver les répondants); de toute évidence,

cette information pourrait être utilisée dans le modèle servant à la pondération, mais elle ne serait pas disponible au moment de l'interview précédente (car l'événement, à ce moment, n'a pas encore eu lieu). Au mieux, nous pourrions espérer trouver, pour les besoins du modèle de prévision, une variable ou un ensemble de variables corrélé avec les déménagements subséquents.

3.1 Le modèle

Une régression logistique a été utilisée pour créer le modèle. Nous avons choisi ce type de modèle parce que la non-réponse est une variable dépendante dichotomique. La régression logistique a été préférée à l'analyse discriminante du fait qu'elle comporte moins d'hypothèses et qu'elle est, essentiellement, aussi efficace que l'analyse discriminante (Harrell 1983).

La fonction de réponse logistique multiple est

$$E\{Y \mid X\} = [1 + \exp(-\beta^T X)]^{-1}, \tag{1}$$

où Y est la variable dépendante,
β est le vecteur colonne des paramètres de régression,
X est la matrice $n \times (p\text{-}1)$ des variables indépendantes.

En développant l'équation (1), on obtient

$$E\{Y \mid X\} = [1 + \exp(-\beta_0 - \beta_1 X_1 - \cdots - \beta_{p-1} X_{p-1})]^{-1}. \tag{2}$$

La variable dépendante, Y_i, dans cette analyse, indiquait si le i^e répondant à l'enquête de 1986 était devenu un non-répondant à l'enquête de 1987. Par conséquent:

$Y_i = 1$ si la i^e personne n'a pas répondu en 1987,
$Y_i = 0$ si la i^e personne a répondu en 1987.

Selon le modèle de régression logistique multiple, les Y_i sont des variables aléatoires de Bernoulli indépendantes pour lesquelles

$$E\{Y_i \mid X_i\} = [1 + \exp(-\beta^T X_i)]^{-1} \tag{3}$$

où X_i est le vecteur des p-1 variables indépendantes associées à la i^e personne.

Si $P(Y=1|X)$ est dénoté par $\pi(X)$, la transformation logit est définie par

$$g(X) = \ln\left[\frac{\pi(X)}{1-\pi(X)}\right]$$
$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \quad .$$

3.2 Les données

Le panel de 1986-1987 de l'enquête sur l'activité a été utilisé pour ajuster et évaluer les modèles de non-réponse. L'ensemble de données était constitué de 66,817 personnes, dont 3,385 (5%) étaient des non-répondants à l'interview de 1987. Des variables démographiques susceptibles d'être reliées à la non-réponse ont été choisies dans le fichier principal de l'enquête sur l'activité de 1986, à titre de variables indépendantes possibles pour le modèle. Une variable additionnelle a été recueillie en 1987 pour toutes les personnes, c'est-à-dire qu'on a déterminé si oui ou non elles avaient changé d'adresse depuis l'interview de 1986.

3.3 Les variables

Les variables examinées en vue d'une inclusion dans le modèle de non-réponse étaient les suivantes:

Information géographique:	Province et type de région (urbaine/rurale) à l'interview de 1986;
Information sur le ménage/le logement:	Taille du ménage, type de logement (maison, autre), mode d'occupation (propriétaire, locataire) à l'interview de 1986;
Information démographique:	Sexe, âge, état matrimonial, fréquentation scolaire (à plein temps, à temps partiel, aucune), niveau de scolarité à l'interview de 1986;
Caractéristiques de l'activité:	Toute période d'emploi, toute période de chômage, toute période hors de la population active, nombre d'emplois, tout emploi de courte durée (< 2 ans), tout emploi de longue durée (2 ans ou plus), toute absence du travail, secteur d'activité de l'emploi ou des emplois, revenu hebdomadaire moyen (pour l'ensemble des emplois), toute période de prestations d'assurance-chômage, toute période de prestations d'aide sociale en 1986;
Déménagement/non-déménagement:	Y a-t-il eu déménagement? (changement d'adresse entre l'interview de 1986 et celle de 1987).

Les variables formées de catégories ont été analysées et traitées de façon que la représentation finale de cette information comporte des groupes de variables binomiales (0-1). Les différences entre les répondants et les non-répondants vis-à-vis des variables indépendantes ont été analysées. Nous avons examiné les corrélations entre toutes les paires de ces variables afin de détecter toute multicolinéarité possible.

3.4 Le fichier échantillon

La procédure PROC LOGIST du SAS a servi à ajuster le modèle de régression logistique. En raison de la taille de l'ensemble de données, le processus exigeait une quantité imposante de ressources informatiques. Par conséquent, il a été décidé de sélectionner un échantillon parmi les ménages du fichier original en vue de l'étape de construction du modèle. Le fichier échantillon était constitué de tous les ménages comportant un non-répondant, et d'un nombre égal de ménages formés de répondants seulement, sélectionnés au hasard. Ce genre d'échantillon était préférable à un échantillon aléatoire simple, car il était plus facile de déterminer les variables associées à la non-réponse en utilisant toute l'information disponible sur la non-réponse. Les paramètres du modèle de régression ont été estimés à l'aide de l'ensemble de données complet.

3.5 Méthodes de régression

En premier lieu, une méthode de régression linéaire pas à pas a été utilisée pour repérer des variables susceptibles d'être utiles à la modélisation. Cette diminution du choix de variables a permis de réduire le nombre de variables devant être introduites dans le processus logistique et, ainsi, de réaliser une importante économie de ressources informatiques.

Les variables données par la procédure STEPWISE ont été introduites dans la procédure PROC LOGIST du SAS, avec les options BACKWARD et FAST. Ces options ont permis à la procédure LOGIST d'utiliser une méthode d'élimination rétrograde approximative pour éliminer les variables non significatives. Différents modèles de régression logistique ont été ajustés en fonction de l'ensemble de données complet, au moyen de combinaisons des variables les plus significatives obtenues à partir du fichier échantillon. L'un des critères du choix du modèle était le nombre de variables. On voulait que le modèle se limite à quelques variables afin qu'il soit simple à utiliser.

Le recours à un modèle de non-réponse dans une enquête longitudinale vise à permettre de rajuster les poids des répondants la deuxième année (1987). Pour ce modèle, la variable dépendante était la non-réponse totale,

tandis que les variables indépendantes étaient des caractéristiques observées l'année précédente (1986), plus l'information de l'année courante (1987) sur les déménagements.

L'option BACKWARD de la procédure PROC LOGIST a été appliquée au fichier échantillon, ce qui a permis de déterminer huit variables liées à la non-réponse.

Homme	(MALE)
Célibataire	(SINGLE)
Locataire	(RENT)
Tout emploi	(ANYEMP)
Scolarité = secondaire	(EDUCSEC)
Déménagement depuis l'interview de 1986	(MOVED)
Taille du ménage	(HHS)
Âge	(AGE)

Avant d'ajuster le modèle en fonction de l'ensemble de données complet, on a examiné, pour les deux variables continues (taille du ménage et âge), la présence d'une linéarité dans le logit. Le tracé des courbes des deux variables a révélé que ni l'une ni l'autre ne semblait linéaire. La non-réponse était élevée pour les âges 16-24, elle était faible pour les âges 25-54, et elle s'élevait quelque peu pour les âges 55+. Nous avons tenté de faire certaines transformations, mais sans succès. Nous avons plutôt décidé de créer des groupes d'âge, et de remplacer la variable continue de l'âge par deux variables binomiales (AGE1, AGE2). Parce que peu de personnes appartenant à l'échantillon venaient de ménages de très grande taille, il a été décidé de regrouper les ménages de 8 membres ou plus et d'attribuer la valeur 8 à la variable ainsi modifiée. Un tracé de la non-réponse en fonction de la nouvelle variable «taille du ménage» a révélé une distribution essentiellement en forme de V. La transformation ABS(HHS - 4.5) a été utilisée pour linéariser la variable. La variable «taille du ménage» transformée a été appelée HHSTRANS.

Tableau 1: Estimations des paramètres pour le modèle final de pondération.

Variable	$\hat{\beta}$	Erreur-type	χ^2
ORDONNÉE À L'ORIG.	-3.81	0.14	702.59
HHSTRANS	0.13	0.06	4.97
MALE	0.25	0.04	41.98
RENT	.23	0.04	29.14
SINGLE	0.11	0.16	0.43
MOVED	2.31	0.04	3,065.95
AGE1	-0.15	0.17	0.75
AGE2	-0.19	0.15	1.65
AGE1*HHSTRANS	0.02	0.07	0.07
AGE2*HHSTRANS	0.05	0.06	0.55
AGE1*SINGLE	0.13	0.18	0.52
AGE2*SINGLE	0.11	0.17	0.40

Quatre modèles ont été ajustés en fonction de l'ensemble de données complet, d'après: (1) les huit variables; (2) toutes les variables sauf RENT; (3) toutes les variables sauf EDUCSEC; (4) toutes les variables sauf EDUCSEC et AGE. Même si les huit variables étaient toutes significatives au moment de l'utilisation du fichier

échantillon, certaines n'apparaissaient plus importantes lors de l'ajustement en fonction de l'ensemble de données complet. Toutefois, il a été décidé de les conserver quand même dans les modèles. Les statistiques permettant d'évaluer la qualité des ajustements ont révélé peu de différences entre les quatre modèles. Les graphique des résidus de Pearson en fonction des valeurs ajustées ont été tracés et examinés. Les résidus du modèle (3) ont indiqué un ajustement légèrement meilleur comportant moins de valeurs extrêmes. Encore une fois à l'aide du fichier échantillon, nous avons examiné les données pour voir s'il existait des interactions bidirectionnelles entre les variables du modèle. Deux ensembles d'interactions ont été ajoutés au modèle: (AGE1 AGE2)*HHSTRANS et (AGE1 AGE2)*SINGLE. Un sommaire des valeurs ajustées pour ce modèle est présenté ci-dessus. Il est à remarquer que les variables «âge» et «célibataire», ainsi que leurs interactions, ne sont pas statistiquement significatives. Néanmoins, nous avons constaté que lorsqu'un modèle était ajusté sans ces variables, les résidus comportaient davantage de valeurs extrêmes.

Au moyen des estimations des paramètres du modèle final, des probabilités prévues de non-réponse ont été calculées pour tous les répondants à l'interview de 1987. Le rajustement de pondération visant à compenser la non-réponse a été obtenu en divisant le poids initial de 1986 par (1 - probabilité prévue). On disposait ainsi de poids rajustés pour 1987. Une stratification a posteriori a ensuite été effectuée afin de rajuster les poids par rapport à des totaux de contrôle de la population. À cette fin, un rajustement par quotient a été appliqué aux poids à l'intérieur de catégories «province-sexe-groupe d'âge», ce qui a donné les poids finals de 1987.

3.6 Évaluation des poids

L'enquête sur l'activité étant une enquête longitudinale, les mêmes personnes sont présentes dans l'échantillon pour les deux années. La seule différence entre la composition du fichier de 1986 et celle du fichier de 1987 tient au fait qu'il manque certaines personnes dans le fichier de 1987 en raison de la non-réponse. Si le rajustement de pondération visant à compenser la non-réponse est adéquat, il ne devrait pas y avoir de différence entre les estimations obtenues pour les répondants de 1986 et celles obtenues pour les répondants de 1987 lorsqu'on fait des totalisations selon les caractéristiques de 1986. Un certain nombre de caractéristiques démographiques et liées à l'activité ont été évaluées. Les estimations ont été faites d'après les poids de 1986, les poids de 1987 rajustés selon le modèle et des poids de 1987 normaux (la non-réponse étant compensée par un rajustement par quotient à des niveaux géographiques peu élevés). Pour chaque caractéristique, un intervalle de confiance à 95% a été calculé pour les estimations d'après les poids de 1986. Les deux estimations de 1987 ont été comparées à celles de 1986 et entre elles, pour évaluer les écarts. Le tableau qui suit montre certains des résultats.

Pour toutes les caractéristiques comparées, seulement une estimation de 1987 se situait en dehors de l'intervalle de confiance de 1986: 49-52 semaines d'emploi selon la pondération normale. Une tendance, toutefois, apparaissait clairement. Les estimations fondées sur les poids rajustés selon le modèle étaient régulièrement plus proches des estimations de 1986 que celles fondées sur la méthode de pondération normale. Les deux estimations de 1987 ont également été comparées d'après les sous-poids (avant le rajustement de la stratification a posteriori), ainsi qu'au niveau des provinces et du pays dans son ensemble. En général, les écarts entre les estimations de 1987 étaient plus élevés lorsqu'on utilisait les sous-poids plutôt que les poids finals. Les écarts, en outre, étaient plus prononcés pour les caractéristiques liées à l'activité que pour les caractéristiques démographiques; les écarts étaient plus grands dans le cas des variables incluses dans le modèle de non-réponse; les écarts étaient plus grands à l'échelon provincial qu'au niveau national. Bien que les écarts soient de faible ampleur, il semble que la méthode fondée sur le modèle donne de meilleurs résultats. On croit qu'avec une non-réponse s'étendant sur un plus grand nombre d'années, les gains seront encore plus appréciables.

Tableau 2: Comparaison des estimations fondées sur les poids finals, d'après les caractéristiques de 1986.

	Estimation de 1986	Int. de conf. à 95% pour l'estimation de 1986	Estimation de 1987 fondée sur le modèle	Estimation normale de 1987
État matrimonial				
Marié(e)	64.6%	(64.1,65.1)	64.8%	65.1%
Célibataire	26.7%	(26.3,27.0)	26.6%	26.4%
Veuf(ve)	3.1%	(2.9,3.3)	3.0%	3.0%
Divorcé(e)	5.7%	(5.4,6.0)	5.5%	5.4%
Scolarité				
0-8e année	14.7%	(14.2,15.2)	14.7%	14.6%
Secondaire	50.3%	(49.7,50.9)	50.0%	50.1%
Postsecondaire partiel	10.1%	(9.8,10.4)	10.2%	10.2%
Cert./dipl. postsecondaire	12.9%	(12.5,13.3)	13.0%	13.0%
Diplôme universitaire	12.0%	(11.6,12.4)	12.1%	12.1%
Activité				
Tout emploi	77.2%	(76.8,77.6)	77.3%	77.6%
Toute période de chômage	17.3%	(16.9,17.7)	17.1%	16.9%
Toute période hors de la population active	40.5%	(40.0,41.0)	40.3%	40.1%
Semaines d'emploi en 1986				
0 semaine	22.8%	(22.4,23.2)	22.7%	22.4%
1-26 semaines	12.0%	(11.7,12.3)	11.8%	11.8%
27-48 semaines	12.2%	(11.9,12.5)	12.1%	12.1%
49-52 semaines	53.0%	(52.4,53.6)	53.4%	53.7%

4. IMPUTATION

L'imputation est souvent la solution de rechange aux rajustements de pondération pour compenser la non-réponse. Même si l'EDTR est avant tout une enquête longitudinale, des estimations transversales seront aussi produites. Ces deux besoins différents se répercutent sur la stratégie d'imputation.

Selon les plans actuels, un fichier transversal sera produit tous les ans. Les fichiers transversaux ne seront pas raccordés, de sorte qu'il n'y aura pas d'obligation d'assurer la cohérence longitudinale avec l'année précédente. Il suffit donc que la méthode d'imputation assure la cohérence interne. Le fichier transversal sera principalement formé de répondants longitudinaux, mais il pourra aussi inclure de nouvelles personnes s'étant jointes au ménage. La catégorie du répondant (nouveau membre du ménage ou répondant longitudinal) influera sur l'information disponible pour les fins de l'imputation.

Un fichier longitudinal sera publié à la fin du cycle (sauf peut-être pour le premier panel, à l'égard duquel les responsables de l'EDTR diffuseront vraisemblablement des fichiers longitudinaux tous les ans jusqu'à ce qu'un

panel complet soit terminé). Le fichier longitudinal constituera un dossier raccordant l'information sur les répondants longitudinaux pour toute la durée du panel. Dans le cas de l'enquête sur l'activité, un important nettoyage a été effectué dans le fichier des répondants pour assurer une certaine cohérence dans l'information longitudinale. Les exigences imposées par la cohérence se répercutent sur les méthodes d'imputation. Même si l'information longitudinale est susceptible de produire une imputation plus robuste, elle rend aussi l'imputation beaucoup plus difficile. Puisqu'une enquête longitudinale est souvent axée sur des mesures de caractéristiques en évolution, la méthode d'imputation doit être mise en oeuvre d'une façon qui minimise les changements artificiels.

4.1 Analyse des non-répondants longitudinaux

Des essais ont été faits au moyen des fichiers longitudinaux de l'enquête sur l'activité. Le panel de deux ans et celui de trois ans ont tous deux été utilisés pour évaluer la possibilité de recourir à l'imputation dans une enquête longitudinale. Voici un sommaire des résultats obtenus:

- La plupart des non-répondants appartenaient à un ménage entièrement non-répondant la première année. La deuxième année comportait aussi un taux de non-réponse relativement élevé, principalement attribuable à des personnes introuvables,

- les variables se chevauchant à la lisière (dates de fin de certains emplois) donnaient de puissants prédicteurs de l'activité globale au cours de l'année,

- les comparaisons entre les répondants au cycle de trois ans et les cas de non-érosion (personnes ayant répondu la première et la troisième année, mais non la deuxième année) ont révélé des différences dans les types d'emplois occupés par les personnes des deux groupes. Le tableau 3 montre un exemple des résultats obtenus de la comparaison des types d'emplois occupés la première et la troisième année du cycle. Un emploi longitudinal est un emploi qui était occupé à la fois l'année 1 et l'année 3 (même employeur et même profession),

- parmi les répondants, pour les emplois longitudinaux, les caractéristiques de l'emploi (comme l'appartenance à un syndicat, la catégorie de travailleur) présentaient une très forte corrélation entre deux années. Les variables comme le traitement/salaire ou les heures travaillées sont davantage touchées par l'erreur de réponse, de sorte qu'elles n'étaient pas aussi corrélées que prévu entre deux années. La corrélation entre deux années consécutives de variables comme les interruptions d'emploi, les mises à pied ou les absences était beaucoup plus faible,

- les comparaisons entre les répondants au cycle de trois ans et les cas de non-érosion, pour des questions non liées à l'activité (comme l'incapacité ou l'état de santé), ont révélé que ces caractéristiques étaient assez stables.

Tableau 3: Comparaisons de l'activité générale, pour les répondants au cycle de trois ans de l'enquête sur l'activité, par rapport aux cas de non-érosion.

	Répondants	Cas de non-érosion
Pas d'emploi l'année 1 et l'année 3	18%	16%
Emplois l'année 1 mais pas l'année 3	6%	12%
Emplois l'année 3 mais pas l'année 1	4%	4%
Emplois l'année 1 et l'année 3, mais pas d'emploi longitudinal	27%	49%
Emplois l'année 1 et l'année 3, au moins un emploi longitudinal	45%	19%

4.2 Stratégie d'imputation

D'après ces résultats, la stratégie d'imputation expérimentée sur les données de l'enquête sur l'activité a été la suivante:

- Nous avons exclu de l'imputation les enregistrements qui représentaient des non-répondants complets dans l'enquête sur l'activité. Il en est résulté une diminution de plus de 50% de la quantité d'imputations à faire,

- nous avons divisé les enregistrements exigeant une imputation en deux groupes: les enregistrements ne comportant pas d'information longitudinale et les enregistrements comportant une information longitudinale. Pour les enregistrements sans information longitudinale, une imputation hot-deck a été effectuée pour toute l'information relative à l'activité. Les classes d'imputation ont été établies d'après l'information démographique propre à la personne, ainsi que certaines données de base sur le ménage,

- l'imputation relative aux enregistrements comportant une information longitudinale a été plus complexe; un équilibre a dû être trouvé entre la cohérence interne et la cohérence longitudinale. Une stratégie de compromis a été testée; les variables exigeant une imputation ont été divisées en trois catégories: (1) caractéristiques de la personne (p. ex. incapacité et santé), (2) information sur l'activité exigeant une cohérence longitudinale et (3) autre information sur l'activité,

- les caractéristiques de la personne (1) ont été imputées d'après les réponses de l'année précédente (ou des deux années pour les cas de non-érosion),

- pour l'information sur l'activité, une imputation hot-deck a été effectuée. Les variables utilisées pour former les classes d'imputation comprenaient à la fois des variables nécessaires pour assurer la cohérence de l'information longitudinale (p. ex. si un emploi devait être occupé au début de l'année, et certaines caractéristiques globales de cet emploi) et d'autres variables prédictives comme le groupe d'âge, le sexe, l'état matrimonial, le nombre d'emplois occupés l'année précédente, etc.,

- l'information sur le donneur a été utilisée pour attribuer les composantes de l'activité à l'année manquante dans le cas de l'information (2) et de l'information (3); la cohérence interne de l'enregistrement a ainsi été assurée,

- l'information sur l'activité exigeant une cohérence longitudinale (2) a été imputée de nouveau à l'aide des données longitudinales (p. ex. le nom de l'employeur, le secteur d'activité détaillé et la profession qui avaient été imputés d'après les valeurs du donneur ont été remplacés par les valeurs de l'année précédente relatives à cet emploi longitudinal). On a ainsi maintenu au minimum le nombre de changements artificiels qui auraient été introduits par le report des valeurs du donneur dans des zones de données devant afficher une cohérence longitudinale.

Des essais sont toujours en cours. Il a fallu beaucoup de temps pour élaborer le processus d'imputation. Même lorsque les classes d'imputation sont relativement larges, une quantité appréciable de fusions est nécessaire pour permettre que tous les non-répondants fassent l'objet d'une imputation. L'information sur les cas de non-érosion offre de nombreux outils pour l'imputation, et l'analyse initiale laisse entrevoir des résultats prometteurs. Toutefois, la stratégie mise au point ne donne pas de bons résultats dans les cas peu nombreux caractérisés par une évolution très complexe de la situation sur le marché du travail. Il faudra peut-être examiner une stratégie d'imputation différente pour ce petit nombre de cas.

5. STRATÉGIE GLOBALE ET TRAVAUX FUTURS

Le dilemme entre la pondération et l'imputation pour traiter la non-réponse est une question avec laquelle ont dû composer les responsables de plusieurs enquêtes. Sur la base des études qui sont actuellement en cours, le traitement de la non-réponse dans l'EDTR sera formé d'une combinaison de méthodes de pondération et d'imputation. La non-réponse à la totalité des vagues sera compensée par une pondération. La non-réponse à une interview différée (réponse à l'interview sur l'activité mais pas à celle sur le revenu, ou vice-versa) fera

l'objet d'une imputation. Jusqu'ici, il est proposé de traiter les cas de non-érosion par l'imputation. Quant aux cas d'érosion, le choix entre la pondération et l'imputation sera probablement effectué quand on saura (1) si le non-répondant fait partie ou non d'un ménage qui est entièrement non-répondant, (2) le nombre d'années de données obtenues jusqu'à ce moment et (3) les taux de non-réponse.

L'imputation longitudinale n'est pas une tâche évidente. Les analyses effectuées jusqu'ici montrent qu'il n'existe pas de méthode unique qui soit la «meilleure» pour imputer en même temps toutes les variables. L'imputation de vague exige le recours à des modèles différents pour des ensembles de variables différents. Jusqu'à maintenant, l'imputation a été limitée aux données sur l'activité. Il s'agit d'une simplification du processus qui devra être mis en oeuvre dans l'EDTR, car lorsqu'une seule des deux interviews (activité ou revenu) sera manquante, il s'agira d'une non-réponse «à des questions particulières». Puisque les données de l'EDTR seront recueillies par des interviews assistées par ordinateur, nous espérons que si une interview sur l'activité est manquante, une information minimum sur l'activité pourra être demandée au moment de l'interview sur le revenu, afin de rendre plus robuste l'imputation. Si l'interview sur le revenu est manquante, on envisage la possibilité d'établir une correspondance avec des sources administratives comme les dossiers fiscaux, afin de faciliter le processus d'imputation.

Rao a proposé d'utiliser un estimateur de variance jacknife dans le cadre d'une imputation hot-deck. Ainsi, on pourra examiner l'impact de l'imputation sur les estimations. La technique d'imputation sera par ailleurs évaluée au moyen d'études additionnelles effectuant la comparaison des taux de transition pour les cas de non-érosion.

Les modèles de pondération semblent très prometteurs. Même si les différences obtenues avec la méthode des modèles étaient faibles dans le cas de notre étude de la non-réponse ayant porté sur un intervalle d'un an, nous nous attendons à ce que les gains soient supérieurs sur une période plus longue. Le plan des travaux à venir comprend des essais sur la stabilité du modèle pour le panel de trois ans de l'enquête sur l'activité. Certains aspects opérationnels doivent être examinés si la compensation de la non-réponse due à l'érosion est faite au moyen d'un rajustement de pondération. Par exemple, lorsque la troisième année de données sera ajoutée, il n'y aura pour certains des non-répondants que les données de la première année, tandis que pour d'autres, on disposera des données des deux premières années. La façon exacte dont on procédera pour intégrer ces aspects complexes au modèle n'a pas encore été décidée. Les essais à ce sujet devraient porter sur le panel de trois ans. Un estimateur jacknife pour l'évaluation du modèle est également élaboré, afin de permettre des évaluations appropriées des estimations fondées sur le modèle.

Dans la présente communication, nous avons examiné la non-réponse du point de vue de la réaction au fait accompli. L'équipe de l'EDTR effectue également des recherches sur la façon de réduire le fardeau de réponse et d'obtenir la coopération des répondants. On tente également de mettre en oeuvre une stratégie de dépistage, afin de réduire au minimum la perte de répondants.

BIBLIOGRAPHIE

Harrell, F.E. (1986). *The LOGIST Procedure, SUGI Supplemental Library Guide*, Version 5 Edition, Cary, NC: SAS Institute Inc.

Hosmer, D.W. Jr., et Lemeshow, S. (1989). *Applied Logistic Regression*, John Wiley & Sons.

Hunter, L., Michaud, S., et Torrance, V. Modelling for non-response in a longitudinal survey.

Kalton, G. (1986). Handling wave non-response in panel surveys. *Journal of Official Statistics*, 2, 3, 303-314.

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys. *Panel Surveys*, John Wiley & Sons, 348-374.

Rao, J.N.K. (1992). Jacknife variance estimators under hot-deck imputation. Working paper.

Singh, M.P., Drew, J.D., Gambino, J.G., et Mayda, F. *Methodology of the Canadian Labour Force Survey 1984-1990*, publication de Statistique Canada, catalogue 71-256.

The Labour Market Activity Survey. 1986-87 Longitudinal File, Microdata User's Guide, Special Surveys Group, Statistique Canada.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

IMPUTATION POUR LA NON-RÉPONSE DE VAGUE DANS L'ENQUÊTE «SURVEY OF INCOME AND PROGRAM PARTICIPATION» (SIPP)

J.M. Lepkowski, D.P. Miller, G. Kalton et R. Singh[1]

RÉSUMÉ

Quand des participants à une enquête longitudinale ne répondent pas lors d'une vague, on peut compenser pour la non-réponse à l'aide d'imputations faites à partir de données provenant des mêmes, ou d'autres participants. On a simulé la non-réponse de vague parmi les personnes qui ont répondu à l'ensemble des sept vagues du panel SIPP de 1987. Les vagues de données manquantes simulées ont été remplacées à l'aide d'une méthode simple et d'une méthode modifiée d'imputation par report, ainsi que de méthodes «hot-deck» longitudinales. Des comparaisons directes des valeurs imputées et des valeurs réelles sont effectuées, et l'effet de l'imputation sur l'estimation de la durée des périodes de participation aux programmes est examiné.

MOTS-CLÉS: Hot-deck; report; données manquantes simulées; estimation de la durée des périodes.

1. INTRODUCTION

La non-réponse de vague est une forme étendue de la non-réponse à des questions particulières, en vertu de laquelle une unité de l'échantillon ne répond pas à une ou plusieurs vagues du même panel. Trois méthodes de base permettent de compenser pour la non-réponse de vague: la pondération, l'imputation et une combinaison des deux (Kalton 1986; Lepkowski 1989). La pondération consiste à appliquer des rajustements compensatoires pour la non-réponse de vague aux unités de l'échantillon ne comportant pas de vagues manquantes. Un poids unique peut être attribué aux unités de l'échantillon répondant à toutes les vagues de l'enquête (comme on le fait dans l'enquête SIPP), à titre de compensation pour toute unité de l'échantillon qui ne répond pas à une ou plusieurs vagues. On peut aussi attribuer plusieurs poids à titre de compensation pour des profils particuliers de non-réponse de vague. Par exemple, des poids peuvent être attribués afin de compenser pour la non-réponse due à l'érosion à chaque vague. Ainsi, une enquête par panel à sept vagues comporterait sept poids, un pour la non-réponse de la première vague, et six autres poids pour compenser les pertes enregistrées à chaque vague successive. Des plans encore plus complexes pourraient être imaginés, mais l'utilisation d'un poids unique est favorisée, car elle simplifie la manipulation des données pour l'analyste.

L'imputation consiste à remplacer une vague manquante entière par des données provenant soit de la même unité de l'échantillon, soit d'une autre unité. Des nombres élevés de vagues peuvent faire l'objet d'une imputation pour certains profils de non-réponse de vague, de sorte qu'il se peut que la majorité des données soient imputées pour un sujet particulier. Des combinaisons de l'imputation et de la pondération sont possibles; ainsi, on peut compenser pour une bonne partie, mais non la totalité, des vagues manquantes par une imputation, et pour le reste des vagues manquantes par l'attribution de poids. En procédant ainsi, on évite d'avoir à appliquer des poids multiples et à imputer une grande quantité de données pour un sujet particulier.

L'enquête SIPP est une enquête par panel permanente menée auprès de la population civile américaine ne vivant pas en institution (Short 1985). Un nouveau panel est créé chaque année, et les membres des ménages de l'échantillon original sont suivis pendant sept ou huit interviews réalisées tous les quatre mois. Les interviewers du Bureau of the Census recueillent des données, à chaque vague, sur un nombre important de sources de

[1] J.M. Lepkowski et D.P. Miller, Institute for Social Research, P.O. Box 1248, Ann Arbor (MI), É.-U. 48106. G. Kalton, Westat, Inc., Rockville (MD), É.-U. R. Singh, U.S. Bureau of the Census, Washington (DC), É.-U.

revenus. Le panel de 1987, sur lequel porte la présente recherche, a couvert sept vagues de collecte de données qui ont servi à produire 28 rapports mensuels consécutifs sur les revenus et la participation aux programmes pour les membres de l'échantillon ayant participé à l'ensemble des sept vagues, et qui ont procuré des données en moins grand nombre pour les personnes n'ayant pas répondu à une ou plusieurs vagues. Lorsqu'un panel prend fin, un fichier longitudinal liant les interviews individuelles d'une vague à l'autre est créé; c'est ce fichier qui a constitué la source de données de la présente étude.

Notre recherche s'intéresse principalement à l'estimation de la durée des périodes de réception de prestations du programme AFDC sur la période de 28 mois couverte par le panel de 1987. Plus précisément, la présente communication examine la nature des profils de non-réponse de vague pour les répondants ayant répondu à la totalité et à une partie des vagues du panel, ainsi que les méthodes permettant de compenser pour la non-réponse de vague (Kalton 1990; Lepkowski 1989). La présente communication décrit des travaux en cours de réalisation. Elle examine l'exactitude relative de plusieurs techniques d'imputation simples pour la non-réponse de vague. Elle ne traite pas des questions liées à la non-réponse dont on doit tenir compte (Fay 1986 et Fay 1989), et se limite aux méthodes de compensation qui conviennent à la non-réponse dont on peut ne pas tenir compte.

L'enquête SIPP se fonde actuellement sur une stratégie d'analyse longitudinale qui utilise un poids longitudinal unique. Toutes les unités de l'échantillon ayant une ou plusieurs vagues de non-réponse sont incluses dans l'ensemble de données, mais se voient attribuer un poids de zéro. Les autres répondants, qui ont répondu au panel entier, reçoivent un poids longitudinal non nul qui compense pour les unités exclues. Cette méthode a été critiquée du fait qu'elle exclut une quantité importante de données recueillies auprès de personnes ayant fourni une réponse partielle. Par exemple, les unités de l'échantillon affichant une non-réponse provisoire touchant une seule vague sont exclues, malgré la présence de six des sept vagues de données.

Les profils de la non-réponse de vague sont examinés dans la prochaine section. Une méthode de simulation de ces profils utilisant les répondants au panel entier est décrite à la section 3, tandis que les méthodes d'imputation utilisées dans notre recherche sont décrites à la section 4. La section 5 présente les résultats de l'imputation dans le cas de la participation au programme AFDC, y compris l'impact de l'imputation de vague sur l'estimation du nombre de périodes de prestations et de la durée des périodes. La section 6 décrit brièvement des recherches additionnelles portant sur l'imputation des montants de prestations.

2. NON-RÉPONSE DE VAGUE DANS L'ENQUÊTE SIPP DE 1987

Le tableau 1 présente les profils et les fréquences de la non-réponse (et de la réponse) de vague dans l'enquête SIPP de 1987. Près de 80% de toutes les personnes faisant partie de l'échantillon (l'unité d'analyse dans la présente étude) ont répondu à l'ensemble des sept vagues, ou à toutes les vagues pour lesquelles elles étaient des membres admissibles de la population d'inférence. Un faible pourcentage (2.6 %) des membres du panel sont devenus inadmissibles à une vague particulière du panel et sont demeurés inadmissibles pour toutes les vagues subséquentes. Le taux d'inadmissibilité est relativement constant d'une vague à l'autre, les pertes attribuables au décès, à l'entrée dans les forces armées ou en institution, ou encore au déménagement à l'étranger étant à peu près égales à chacune des vagues. Les cas affichant une non-réponse à une vague quelconque ou comportant des vagues inadmissibles sont écartés dans la pondération longitudinale de l'enquête SIPP.

Parmi les profils de non-réponse de vague, les plus fréquents sont ceux liés à l'érosion, représentant les trois cinquièmes de toute la non-réponse de vague. L'érosion la plus forte survient à la deuxième ou à la troisième vague. Un peu plus du quart de la non-réponse de vague est formée de profils de non-réponse provisoire, dans lesquels une vague caractérisée par une non-réponse est précédée et suivie d'une vague ayant obtenu une réponse. La majorité de la non-réponse provisoire concerne une seule vague, bien qu'on note quelques cas qui peuvent atteindre quatre vagues de non-réponse provisoire.

Une non-réponse de type *Z* est enregistrée lorsque des données ne peuvent être obtenues d'une personne de l'échantillon faisant partie d'un ménage dans lequel d'autres personnes de l'échantillon ont répondu. L'analyse qui suit se limite aux personnes qui ont répondu à la première vague du panel de 1987. Toutefois, il existe un

certain nombre de personnes de l'échantillon ayant affiché une non-réponse de type Z à la première vague. Nous les avons incluses dans le tableau même si, à des fins de simplicité, elles ont été exclues des opérations de simulation et d'imputation. Enfin, un petit nombre de personnes de l'échantillon ont eu une non-réponse de vague avant de devenir inadmissibles. Par souci d'exhaustivité, la recherche inclut les profils de non-réponse de vague pour ces personnes dans les opérations de simulation et d'imputation.

3. SIMULATION DE LA NON-RÉPONSE DE VAGUE

3.1 Modèle de sélection de l'échantillon pour les répondants au panel entier

Kalton et Miller (1986) ont simulé des profils de non-réponse de vague parmi les répondants qui ont répondu en entier aux trois premières vagues du panel SIPP de 1984. Ils ont employé un algorithme SEARCH qui, grâce à un algorithme de séparation binaire (Sonquist, Baker et Morgan 1973), a identifié des sous-groupes de personnes de l'échantillon à l'intérieur desquels il existait une variation importante des profils de non-réponse de vague. Ils ont ensuite prélevé, dans ces sous-groupes, des sous-échantillons de répondants au panel entier. Cette méthode permettait le prélèvement d'un pourcentage minimal (61.6%) de répondants au panel entier dans chaque groupe. Les répondants au panel entier ont été sélectionnés, pour les besoins de la simulation, selon une méthode reflétant la distribution de la population d'où ils provenaient.

Tableau 1: Profils et fréquence de la non-réponse de vague dans l'enquête SIPP de 1987.

Profil de non-réponse	Fréquence	%
Échantillon total	30 769	100.0
Membres du panel	24 448	79.5
Ayant répondu aux 7 vagues (1111111)*	23 653	76.9
Non-réponse pour cause d'inadmissibilité	795	2.6
Décès	357	1.2
Entrée en institution	167	0.5
Entrée aux forces armées	62	0.2
Déménagement à l'étranger	206	0.7
Raison non mentionnée	3	0.0
Non-membres du panel	6 321	20.5
Non-réponse due à l'érosion	3 887	12.6
1222222	1 453	4.7
1122222	713	2.3
1112222	571	1.9
1111222	480	1.6
1111122	338	1.1
1111112	332	1.1
Non-réponse provisoire	1 714	5.6
Une seule vague (e.g., 1211111; 112111)	1 323	4.3
Deux vagues (e.g., 1221111; 1122111)	298	1.0
Trois vagues ou plus	93	0.3
Non-réponse provisoire et due à l'érosion	373	1.2
Non-interview de type Z à la vague 1	271	0.9
Vagues de non-réponse et inadmissibles combinées	76	0.2

* 1 = répondant à la vague, 2 = non-répondant à la vague

Il serait difficile de reprendre cette méthode dans une étude portant sur sept vagues de données. En outre, l'approche de Kalton et Miller utilise un échantillon de tous les non-répondants de vague, ce qui diminue la précision des analyses subséquentes. Nous avons mis au point une méthode alternative utilisant les poids

transversaux existants de la vague 1 de l'enquête SIPP et les poids longitudinaux pour «sélectionner» les répondants au panel entier. La figure 1 présente un échantillon hypothétique pour illustrer le mode de sélection ayant servi à la simulation. Supposons qu'un échantillon de $n = 24$ sujets a été prélevé selon le taux $f = 1/100$, et que la population et l'échantillon sont formés à moitié de femmes. Chaque personne de l'échantillon a un poids de base W_1 de 100. Seulement huit des 12 hommes répondent, tandis que 10 des 12 femmes le font (le fait de répondre étant représenté par l'indicateur $R_i = 1$). Plus précisément, $Pr\{R_i = 1 \mid X_i = H\} = 0.67$, tandis que $Pr\{R_i = 1 \mid X_i = F\} = 0.83$. Pour retrouver la distribution originale de la population (et de l'échantillon), des poids compensatoires rajustés pour la non-réponse (W_2) sont calculés, et l'on obtient 150 pour les hommes et 120 pour les femmes.

La non-réponse de vague peut être traitée de façon semblable en vertu d'une hypothèse de réponse manquante au hasard. Supposons que $F_i = 1$ dénote la réponse à toutes les vagues du panel. Parmi les personnes de l'échantillon ayant répondu, $Pr\{F_i = 1 \mid R_i = 1\} = 13/18$. Encore une fois, le taux des répondants au panel entier diffère entre les hommes et les femmes, et des poids compensatoires, W_3, de 240 et de 150 sont calculés pour les hommes et les femmes respectivement. En vertu d'une hypothèse de réponse manquante au hasard, la non-réponse à la vague 1 et la non-réponse de vague sont compensées par des poids qui varient selon des sous-groupes affichant des taux de réponse différents.

Le poids transversal de la vague 1, W_2, et le poids du panel entier, W_3, peuvent être utilisés pour trouver une méthode de simulation fondée sur la pondération comme solution de rechange à la stratégie de simulation fondée sur l'échantillonnage employée par Kalton et Miller. Un modèle de la non-réponse pour le panel entier, étant donné un ensemble de caractéristiques X, est donné par

$$Pr\{F_i = 1 \mid R_i = 1, X\} = \exp\{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji}\} \; / \; \left[1 + \exp\{\beta_0 + \sum \beta_j X_{ji}\}\right].$$

Cette probabilité conditionnelle peut être estimée pour chaque répondant au panel entier, et son inverse peut être utilisé à titre de compensation pour la non-réponse du panel entier parmi les personnes de l'échantillon de la vague 1.

Malheureusement, la présente recherche ne disposait pas des ressources nécessaires au développement d'un tel modèle. En remplacement, nous avons tiré parti de l'étude de la non-réponse pour le panel entier menée précédemment par le U.S. Bureau of the Census en vue de l'établissement des poids du panel entier. Nous avons fait l'hypothèse que le rapport entre le poids transversal W_2 et le poids du panel entier W_3 est la probabilité conditionnelle de la réponse au panel entier qui est nécessaire à une pondération appropriée. Plus précisément, nous soutenons que pour $W_2 = Pr\{R_i = 1 \mid X\}^{-1}$ et $W_3 = Pr\{F_i = 1 \mid R_i = 1, X\}^{-1}$, la probabilité conditionnelle de la réponse au panel entier est W_2 / W_3. Ainsi, un poids compensatoire pour la réponse au panel entier, en conditionnant (dans notre cas) selon le X inconnu, est donné par W_3 / W_2.

Nous utilisons, pour les besoins de la simulation, le rapport W_3 / W_2 comme poids pour les répondants au panel entier. Des profils de non-réponse de vague sont attribués aux répondants au panel entier de façon aléatoire, selon la distribution observée dans les données. Afin de retrouver la distribution échantillonnale des caractéristiques X ayant servi à la «sélection» des répondants au panel entier, et d'autres caractéristiques Z n'ayant pas servi à cette sélection, nous attribuons aux résultats un poids de $\{1 - W_3 / W_2\}^{-1}$. Cette méthode n'est pas sujette à des pertes de précision attribuables à la sélection de l'échantillon (comme dans les travaux de Kalton et Miller 1986), bien qu'une certaine perte de précision due à la pondération soit enregistrée.

Figure 1: Échantillon hypothétique illustrant la non-réponse pour les unités et le panel entier.

i	X_i	R_i	F_i	Poids de base W_1	Poids transversal W_2	Poids du panel entier W_3
1	H	1	1	100	150	240
2	H	1	1	100	150	240
3	H	1	1	100	150	240
4	H	1	1	100	150	240
5	H	1	1	100	150	240
6	H	1	0	100	150	0
7	H	1	0	100	150	0
8	H	1	0	100	150	0
9	H	0	...*	100	0	0
10	H	0	...	100	0	0
11	H	0	...	100	0	0
12	H	0	...	100	0	0
13	F	1	1	100	120	150
14	F	1	1	100	120	150
15	F	1	1	100	120	150
16	F	1	1	100	120	150
17	F	1	1	100	120	150
18	F	1	1	100	120	150
19	F	1	1	100	120	150
20	F	1	1	100	120	150
21	F	1	0	100	120	0
22	F	1	0	100	120	0
23	F	0	...	100	0	0
24	F	0	...	100	0	0
Total	--	18	13	2 400	2 400	2 400

* Sans objet

3.2 Attribution des profils de non-réponse de vague

Chaque répondant au panel entier s'est vu attribuer au hasard un profil de non-réponse de vague. L'attribution n'a pu être faite entièrement au hasard, car les membres du panel ayant des périodes d'inadmissibilité se sont également vus attribuer des profils. De nombreux profils de non-réponse de vague ne peuvent être attribués à des membres du panel ayant des périodes d'inadmissibilité. Soit $\phi_{ij} = Pr\{$la personne de l'échantillon ayant le i-ième profil de réponse réel a le j-ième profil de non-réponse de vague$\}$. Les valeurs ϕ_{ij} ont pu être calculées d'après un modèle d'indépendance entre le profil de réponse au panel en entier et le profil de non-réponse de vague, sauf pour la présence de «zéros structurels», correspondant à des combinaisons de profils de réponse et de profils de non-réponse de vague qui sont impossibles. La figure 2 illustre le problème. Supposons que 0 représente l'inadmissibilité à une vague donnée. Le profil de non-réponse de vague correspondant à une érosion à la vague 2, 1222222, peut être attribué à n'importe quel profil de réponse du panel, étant donné que lors d'une période de non-réponse de vague, une personne de l'échantillon pourrait devenir inadmissible à notre insu. En revanche, le profil de non-réponse de vague 1122222 ne peut être attribué au profil de réponse du panel 1000000, étant donné que nous ne pouvons attribuer un profil qui comporte la réalisation d'une interview à la vague 2 à une personne dont on connaissait l'inadmissibilité au moment de cette vague.

Figure 2: Attribution de profils de non-réponse de vague aux profils de réponse du panel entier.

Profil de non-réponse de vague (j)	n_j	ϕ_j	Profil de réponse du panel (i)				
			1111111	1000000	1100000	. . .	1111110
1222222	1 453	0.047	ϕ_{11}	ϕ_{21}	ϕ_{31}	. . .	ϕ_{71}
1122222	713	0.023	ϕ_{12}	0	ϕ_{32}	. . .	ϕ_{72}
1112222	571	0.019	ϕ_{13}	0	0	. . .	ϕ_{73}
. . .	. . .	. . .	. . .	. . .	. . .		. . .
Total	6 321	1.000	$\phi_{1\cdot}$	$\phi_{2\cdot}$	$\phi_{3\cdot}$	. . .	$\phi_{7\cdot}$

Un modèle de quasi-indépendance (Agresti 1989) permet le calcul d'estimations des probabilités conjointes pour chaque combinaison de profil de non-réponse de vague et de profil de réponse du panel. Une fois les probabilités conjointes ϕ_{ij} calculées d'après le modèle de quasi-indépendance, des profils de non-réponse de vague ont été attribués au hasard aux répondants au panel entier et aux autres membres du panel. La distribution pondérée des profils de non-réponse de vague simulés correspondait à celle observée pour l'ensemble de l'échantillon et pour environ 20 sous-groupes (p. ex. femmes âgées de 18 à 24 ans).

4. MÉTHODES D'IMPUTATION

4.1 Méthodes d'imputation par report

Une méthode simple d'estimation par report consiste à répéter, pour le même enregistrement, les valeurs des mois non manquants à titre d'imputation pour les mois manquants plus tard dans la période de 28 mois. Puisque tous les cas de non-réponse à la vague 1 ont été exclus de la présente analyse (y compris la non-réponse de type *Z* à la vague 1), l'imputation par report simple commence à la vague 2. Si la vague 2 est manquante, la valeur du mois 4 (c.-à-d. le dernier mois de la vague 1) est reportée pour les quatre mois de la vague 2. Si la vague 3 est manquante, la valeur du mois 8 (c.-à-d. le dernier mois de la vague 2) est reportée pour les quatre mois de la vague 3. Si les données de la vague 2 servant à l'imputation de la vague 3 étaient le résultat d'une imputation d'après le mois 4, les données des quatre mois de la vague 3 sont celles du mois 4. Le processus est répété séquentiellement pour toutes les vagues qui restent. L'imputation par report simple n'introduit aucun changement dans l'état de participation aux programmes. Les mois de toutes les vagues de non-réponse reçoivent la même valeur, celle reportée du dernier mois de la dernière vague non manquante.

Une autre méthode, utilisant aussi bien le report prospectif que le report rétrospectif, a été appliquée à la non-réponse provisoire. Un nombre aléatoire était choisi entre un et le nombre total de mois devant faire l'objet d'une imputation. La sélection aléatoire n'était pas uniforme; elle accordait plutôt des probabilités plus élevées (d'après la distribution empirique des mois auxquels le changement survenait) aux mois terminant une vague. Ce processus reflétait bien le problème de la «lisière» (voir, par exemple, Singh, Weidman et Shapiro 1989; Coder et Ruggles 1988; ou Murray, Michaud, Egan et Lemaitre 1991) dans les données imputées. Tous les mois suivant le mois choisi au hasard faisaient l'objet d'une imputation par report «rétrospectif» d'après le premier mois de la vague non-manquante suivante, tandis que tous les mois allant jusqu'au mois choisi, inclusivement, étaient soumis à une imputation par report «prospectif» d'après le dernier mois de la dernière vague non manquante.

4.2 Méthodes longitudinales

Une méthode d'imputation «hot-deck» longitudinale de base a été utilisée. Les données ont été triées selon sept variables: sexe, âge (cinq catégories), race (quatre catégories), origine hispanique ou non, rapport entre le revenu familial et le seuil de pauvreté à la première vague (cinq catégories), strate d'échantillonnage et code de demi-échantillon. Une matrice triangulaire supérieure de valeurs hot-deck a été créée (voir figure 3) à partir de laquelle jusqu'à 24 mois pouvaient faire l'objet d'une imputation pour une personne de l'échantillon. Selon la vague à laquelle commençait la non-réponse (p. ex. vague 2, vague 3, etc.), les valeurs hot-deck étaient obtenues de la vague correspondante de la matrice.

Deux valeurs de l'enregistrement receveur ont été examinées: l'état de participation au premier mois (par exemple, bénéficiaire de coupons alimentaires au mois 1) et l'état de participation au dernier mois précédant la non-réponse de vague. La combinaison de ces deux éléments pour le donneur était appariée aux valeurs du receveur pour la vague devant faire l'objet d'une imputation. Ainsi, les dons ont été faits à partir de personnes de l'échantillon semblables aux receveurs en ce qui a trait aux variables de tri, avec lesquelles il y avait une correspondance exacte du mois visé (c.-à-d. de la vague), de l'état de participation au premier mois et de l'état de participation au dernier mois. Pour l'attribution initiale des valeurs de la matrice, avant que n'ait lieu l'imputation, les données ont été soumises au processus d'imputation dans l'ordre inverse, sans que des dons soient faits, remplissant les cellules de la matrice avant l'imputation.

Figure 3: Caractéristiques d'appariement d'un processus hot-deck longitudinal propre à un programme, et valeurs hot-deck hypothétiques.

Caractéristiques d'appariement			Valeurs hot-deck à imputer pour chaque mois			
Vague	Coupons alimentaires					
	Mois 1	Mois t-1	5 6 7 8	9 10 11 12	...	25 26 27 28
2	Oui	Oui	X X X X	O O O O	...	O O O O
	Oui	Non	O O O O	O O O O	...	O O O O
	Non	Oui	X X X X	X X X X	...	O O O O
	Non	Non	O O O O	X X X X	...	X X X X
3	Oui	Oui	(Sans objet)	X X X X	...	X X X X
	Oui	Non		O O O O	...	X X O O
	Non	Oui		X X O O	...	O O O O
	Non	Non		O O O X	...	X O O O
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
7	Oui	Oui	(Sans objet)	(Sans objet)	...	X X X X
	Oui	Non			...	O O O O
	Non	Oui			...	O O O O
	Non	Non			...	X X X X

Deux autres méthodes hot-deck ont aussi été utilisées. Dans l'une d'elles, l'état de participation au premier et au dernier mois à l'égard d'un programme était utilisé pour établir un appariement exact dans la matrice, mais les valeurs imputées visaient un autre programme. Par exemple, les donneurs et les receveurs étaient appariés selon qu'ils étaient bénéficiaires ou non de coupons alimentaires au premier et au dernier mois, mais l'état de participation au programme AFDC pour les mois manquants était imputé du donneur au receveur. Cette méthode a permis d'étudier la qualité des appariements pour le genre d'imputation conjointe ou simultanée qui

caractérise les applications «hot-deck» dans lesquelles des variables multiples sont imputées simultanément d'après un processus unique de tri et d'appariement. Dans la deuxième méthode, le mode d'appariement a été modifié, c'est-à-dire qu'on a tenu compte de l'état de participation à l'égard de deux programmes simultanément. L'état de participation au dernier mois pour chaque programme a servi à obtenir un appariement exact dans la matrice. On parle dans ce cas d'une méthode «hot-deck» conjointe, car celle-ci dépend de la distribution conjointe des états de participation pour les deux programmes faisant simultanément l'objet d'une imputation.

Ainsi, pour la participation au programme AFDC, des imputations ont été faites selon trois méthodes hot-deck longitudinales. Le hot-deck AFDC a permis une imputation selon un appariement basé sur l'état de participation au programme AFDC au premier mois et au mois le plus récent. Le hot-deck des coupons alimentaires a permis une imputation selon un appariement basé sur la réception de coupons alimentaires au premier et au dernier mois. Enfin, le hot-deck conjoint a utilisé un appariement basé sur la participation au programme AFDC et la réception de coupons alimentaires au dernier mois.

5. IMPUTATION DE L'ÉTAT DE PARTICIPATION AUX PROGRAMMES

5.1 Imputation des mois

La figure 4 définit la notation permettant de comparer les valeurs réelles et les valeurs imputées pour chaque mois. Par exemple, «a» désigne le nombre de mois pour lesquels la valeur imputée concorde avec la valeur réelle lorsque cette dernière est «oui». La proportion a/(a+c) est le taux d'exactitude parmi les mois où il y a participation au programme («oui»), tandis que la proportion d/(b+d) donne l'exactitude pour les mois de non-participation («non»).

Figure 4: Relation entre les valeurs mensuelles réelles et imputées.

Valeur imputée	Valeur réelle		Total
	Oui	Non	
Oui	a	b	a+b
Non	c	d	c+d
Total	a+c	b+d	h

Le tableau 2 indique l'exactitude des imputations pour les cinq méthodes d'imputation, séparément pour les mois de participation («oui») et de non-participation («non») au programme AFDC, ainsi que la proportion des mois auxquels les valeurs «oui» ou «non» auraient été imputées correctement par chance seulement (en vertu d'un modèle simple d'indépendance entre les valeurs imputées et les valeurs réelles). Les estimations sont présentées seulement pour le nombre total de mois faisant l'objet d'une imputation. Les mois dont les données étaient manquantes en raison d'une non-réponse due à l'érosion à compter de la vague 2, d'une non-réponse provisoire d'une seule vague et d'autres profils de non-réponse de vague ont aussi été examinés, mais les résultats ne sont pas présentés ici car ils étaient essentiellement les mêmes.

La probabilité d'obtenir une imputation exacte par chance seulement est élevée quand la valeur réelle pour le mois est «oui». Pourtant, chaque méthode d'imputation produit de meilleurs résultats que la chance, traitant correctement dans tous les cas sauf un (le hot-deck des coupons alimentaires) plus de la moitié des imputations. Par contre, la probabilité de prédire la non-participation («non») par chance seulement est très faible. Toutes les méthodes d'imputation ont un pouvoir de prédiction de loin supérieur à la chance. La méthode simple et la méthode modifiée d'imputation par report donnent les meilleurs résultats, et ont des rendements qui se situent à peu près à égalité. Les méthodes hot-deck offrent un rendement considérablement inférieur et, encore une fois, c'est le hot-deck des coupons alimentaires qui a le niveau d'exactitude le plus bas.

Il est clair que l'imputation de l'état de participation au programme AFDC est la moins exacte quand la réception de coupons alimentaires est le critère d'appariement. En outre, puisque le hot-deck conjoint donne

des résultats presque aussi bons que le hot-deck AFDC, il semble que la caractéristique critique d'appariement soit le dernier mois (qui est utilisé à la fois dans le hot-deck AFDC et le hot-deck conjoint) plutôt que le premier mois.

Tableau 2: Probabilité d'imputation exacte selon la méthode, pour l'état de participation au programme AFDC.

Méthode d'imputation	Pr{exact|oui}		Pr{exact|non}	
	a/(a+c)	Écart par rapport à la chance	d/(b+d)	Écart par rapport à la chance
Tous les mois				
Chance/espérance	0.957	-	0.043	-
Report simple	0.987	+0.030	0.868	+0.825
Report modifié	0.988	+0.031	0.867	+0.824
Hot-deck AFDC	0.988	+0.031	0.701	+0.658
Hot-deck coupons alim.	0.974	+0.017	0.489	+0.446
Hot-deck conjoint	0.987	+0.030	0.697	+0.654

5.2 Imputation du nombre de périodes de participation

La méthode d'imputation par report ne crée pas de périodes de participation. La figure 5 présente une comparaison sommaire des nombres de périodes réelles et imputées pour chaque personne. Nous nous attendons à ce que, par chance seulement, la plupart des personnes se situent dans la cellule A (concordance dans le cas où il n'y a réellement aucune période de participation), ou la plupart des gens ne sont pas des prestataires du programme AFDC. Les méthodes ne varient sans doute pas beaucoup, par ailleurs, pour ce qui est de leur capacité de produire une imputation exacte lorsqu'il y a réellement des périodes de participation, c.-à-d. dans le cas représenté par les cellules de type D (concordance dans le cas où il y a réellement une ou plusieurs périodes). Les méthodes d'imputation par report omettent des périodes, n'imputant aucune période lorsqu'il en existe réellement, comme dans le cas représenté par les cellules de type B (périodes omises dans le cas où il y a réellement une ou plusieurs périodes), ou imputant moins de périodes qu'il n'en survient réellement, comme dans le cas représenté par les cellules de type F (périodes omises dans le cas où il y a réellement deux périodes ou plus). En revanche, les méthodes d'imputation par report ne peuvent créer des périodes, et ne peuvent contenir des personnes dans les cellules de type C (périodes créées dans le cas où il n'y a réellement aucune période) ou de type E (périodes créées dans le cas où il y a réellement une ou plusieurs périodes). Puisque ces méthodes ne créent pas de périodes, on peut s'attendre à ce qu'elles sous-estiment le nombre total de périodes.

Figure 5: Nombres de périodes réels et imputés: classification des cellules pour l'état de participation au programme AFDC par personne.

Nombre imputé de périodes	Nombre réel de périodes				
	0	1	2	3	4
0	A	B			
1	C	D	F		
2			D		
3		E		D	
4					D

Le tableau 3 présente les taux de ces divers types de concordances du nombre de périodes, de périodes omises et de périodes créées. Bien que les imputations par report présentent des taux plus élevés en ce qui touche les concordances, elles sont biaisées du fait de leur incapacité à créer de nouvelles périodes. Les imputations par report omettent complètement les périodes de courte durée (c.-à-d. celles qui durent moins d'une vague), et ne créent aucune autre période pour compenser. Par contre, les imputations hot-deck omettent certaines périodes de courte durée, en repèrent d'autres, et peuvent aussi créer des périodes pour compenser la perte de périodes de courte durée. L'effet net devrait être une exactitude supérieure des méthodes hot-deck, dans l'ensemble, en ce qui touche le nombre de périodes, car ces méthodes sont capables de créer des périodes, ce qui n'est pas le cas des méthodes d'imputation par report.

Tableau 3: Nombres de périodes réels et imputés pour la participation au programme AFDC, selon la méthode d'imputation.

Méthode d'imputation	Concordance | 0 période (A)	Concordance | 1 pér. ou + (D)	Périodes omises | 1 pér. ou + (B)	Périodes créées (C)	Périodes créées | 1 pér. ou + (E)	Périodes omises | 2 pér. ou + (F)
Tous les membres de l'échantillon						
Report simple	0.967	0.018	0.013	0.000	0.000	0.002
Report modifié	0.967	0.018	0.013	0.000	0.000	0.002
Hot-deck AFDC	0.956	0.019	0.010	0.013	0.001	0.002
Hot-deck coup. alim.	0.932	0.019	0.009	0.034	0.014	0.002
Hot-deck conjoint	0.954	0.020	0.009	0.013	0.002	0.002

Le tableau 4 présente plusieurs taux sommaires des divers types d'erreurs. Les méthodes d'imputation par report ne peuvent pas, par définition, comporter d'imputations «fausses positives» ou de périodes créées, mais elles comportent des imputations fausses négatives et des périodes omises. L'effet net est que les imputations par report sous-estiment le nombre total de périodes. Les imputations hot-deck comportent environ le même taux d'imputations fausses positives et fausses négatives, mais un léger excédent de périodes omises. Donc, contrairement aux attentes, les méthodes d'imputation hot-deck ont tendance elles aussi à sous-estimer le nombre de périodes. Compte tenu du faible nombre de périodes de participation au programme AFDC concernées, le degré de sous-estimation pour les deux types d'imputation devrait vraisemblablement être faible.

Tableau 4: Taux d'erreur pour les imputations du nombre de périodes de participation au programme AFDC selon la méthode d'imputation: tous les membres de l'échantillon.

Méthode d'imputation	Fausses positives C/ (A+C)	Fausses négatives B/ (A+B)	Périodes omises | 1 pér. ou + (B+F)/ (B+F+D+E)	Périodes créées (C+E)/ (C+E+D+F)
Report simple	0.000	0.013	0.443	0.000
Report modifié	0.000	0.013	0.443	0.000
Hot-deck AFDC	0.012	0.010	0.264	0.545
Hot-deck coup. alim.	0.036	0.010	0.164	0.779
Hot-deck conjoint	0.013	0.010	0.247	0.562

La sous-estimation du nombre de périodes par les méthodes d'imputation par report est également apparente dans le tableau 5, qui présente la distribution du nombre de périodes pour les données complètes (c.-à-d. les données réelles pour les répondants au panel entier) et par méthode d'imputation. Les méthodes d'imputation par report montrent une tendance marquée à sous-estimer le nombre total de périodes; elles ont un nombre de personnes plus élevé dans le cas d'une seule période et un nombre de personnes moins élevé dans le cas de deux périodes que les données complètes. Le hot-deck AFDC et le hot-deck conjoint présentent une distribution des périodes qui s'approche davantage de celle des données complètes.

5.3 Estimation de la durée des périodes

La capacité d'estimer avec exactitude le nombre total de périodes est liée à un autre problème d'estimation, celui de la durée des périodes. Le tableau 6 présente les estimations de la durée des périodes pour les données complètes et pour chaque méthode d'imputation. On peut y voir les estimations de Kaplan-Meier de la durée des périodes de participation au programme AFDC obtenues à l'aide des poids appropriés (voir Miller, Lepkoswki et Kalton 1992, pour une discussion sur la méthode d'estimation). Les estimations de Kaplan-Meier relatives aux données complètes représentent les «valeurs non biaisées» qui doivent être estimées à l'aide de données obtenues par les diverses méthodes d'imputation. Les résultats sont conformes à ceux présentés dans la dernière section. Les méthodes d'imputation par report montrent une tendance à surestimer la durée des périodes. Par exemple, la proportion réelle de membres de l'échantillon ayant des périodes AFDC qui ont duré plus de six mois est de 0.539. Les deux méthodes d'imputation par report donnent des estimations supérieures des proportions ayant des périodes plus longues. Par contre, les méthodes hot-deck donnent des estimations qui, à six mois, sont légèrement inférieures à la valeur réelle. Il y a, tout au plus, une légère tendance des méthodes hot-deck à sous-estimer la durée des périodes. La sous-estimation caractérisant les méthodes hot-deck ne semble pas aussi prononcée que la surestimation attribuable aux méthodes d'imputation par report.

Tableau 5: Distribution en pourcentage des périodes de participation au programme AFDC, selon la méthode.

Méthode d'imputation	Nombre de périodes			
	1	2	3	4
Données complètes	79.1	17.7	2.7	0.5
Report simple	81.2	16.0	2.2	0.6
Report modifié	81.2	16.0	2.2	0.6
Hot-deck AFDC	78.3	18.3	2.7	0.5
Hot-deck coup. alim.	76.1	20.6	3.0	0.5
Hot-deck conjoint	78.7	18.3	2.5	0.5

Tableau 6: Estimations de Kaplan-Meier de la durée des périodes de participation au programme AFDC, selon la méthode d'imputation.

Méthode d'imputation	Pourcentage des périodes de participation au programme AFDC durant plus de ...				
	1 mois	3 mois	6 mois	12 mois	24 mois
Données complètes	85.4	74.0	53.9	35.6	27.2
Report simple	85.6	75.0	55.2	38.2	31.4
Report modifié	85.4	74.7	54.8	38.2	31.5
Hot-deck AFDC	85.9	75.0	52.7	35.4	26.5
Hot-deck coup. alim.	87.9	78.1	52.0	35.5	26.6
Hot-deck conjoint	86.0	76.0	52.9	36.3	27.5

6. CONCLUSION

Les opérations d'imputation dont il a été fait état s'inscrivent dans une recherche en cours portant sur l'emploi de méthodes d'imputation longitudinales simples pour compenser l'absence de vagues entières dans une enquête par panel. Le travail s'est limité à l'imputation de données relatives à la réception de prestations d'un petit nombre de programmes. Les résultats indiquent que les méthodes d'imputation par report sont meilleures pour

l'imputation de l'état de participation à chacun des mois, mais qu'en raison de leur incapacité à créer des périodes, elles sont biaisées en ce qui a trait à l'estimation du nombre de périodes et de la durée des périodes. Les méthodes hot-deck affichent de piètres résultats pour ce qui est de l'imputation de l'état de participation mensuel, mais sont un peu meilleures quant à l'imputation du nombre de périodes et de la durée des périodes.

Après s'être intéressée à la participation, la recherche se tournera vers les montants reçus. Nous nous proposons d'imputer les montants en utilisant une méthode semblable à celle illustrée à la figure 3, en faisant un appariement d'après les catégories de revenus pour le premier et le dernier mois. L'appariement à des fins d'imputation conjointe se fera selon les catégories de revenus, et peut-être les catégories de revenus pour un programme et l'état de participation pour un autre. Des mesures comme les écarts quadratiques moyens et la différence entre les données complètes et les données imputées pour ce qui est du revenu annuel total serviront à évaluer l'exactitude. Des corrélations produit-moment simples seront utilisées pour examiner l'atténuation des associations.

BIBLIOGRAPHIE

Coder, J., et Ruggles, P. (1988). *Welfare Recipiency as Observed in the SIPP.* Document de travail 8818. U.S. Bureau of the Census, Washington, DC.

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

Kalton, G., Hill, D., et Miller, M. (1990). *The Seam Effect in Panel Surveys*. Document de travail SIPP n° 9011. U.S. Bureau of the Census, Washington, DC.

Kalton, G., et Miller, M. (1986). Effects of adjustments for wave nonresponse on panel survey estimates. *Proceedings of the Survey Research Methods Sections, American Statistical Association.*

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys. In *Panel Surveys*, D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh, Éds. New York: J.W. Wiley and Sons, Inc., 348-374.

Miller, D.P., Lepkowski, J.M., et Kalton, G. (1992). Estimating duration of food stamp spells from the SIPP. *Proceedings of the Survey Research Methods Section, American Statistical Association*, (à venir).

Murray, T.S., Michaud, S., Egan, M., et Lemaître, G. (1991). Invisible seams? The experience with the Canadian Labour Market Survey. *Bureau of the Census 1991 Annual Research Conference Proceedings.*

Short, K.S. (1985). *Survey of Income and Program Participation: Uses and Applications*. Document de travail SIPP 8501. U.S. Bureau of the Census, Washington, DC.

Singh, R., Weidman, L., et Shapiro, G. (1989). *Quality of the SIPP Estimates.* Document de travail SIPP 8901, U.S. Bureau of the Census, Washington, DC.

SESSION 4

Utilisation d'une structure longitudinale pour l'estimation

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

LISSAGE LONGITUDINAL DE VARIANCES D'INDICES DE PRIX

R. Valliant[1]

RÉSUMÉ

Dans la présente communication, nous élaborons des variances généralisées pour des indices de prix en appliquant des méthodes non paramétriques de lissage de nuages de points à des séries chronologiques d'estimations ponctuelles de variance. Le but visé est de formuler des variances lissées qui sont approximativement sans biais, qui produisent des couvertures acceptables des intervalles de confiance et qui sont plus stables que les estimations ponctuelles de variance. Des méthodes de lissage sont appliquées à des séries chronologiques d'estimations ponctuelles de variance dans une étude de simulation qui fait appel à des données obtenues dans le cadre du programme de l'indice des prix à la consommation des États-Unis.

MOTS CLÉS: Fonction de variance généralisée; indice de prix de Laspeyres; estimateur de variance par linéarisation; «loess»; «super smoother».

1. INTRODUCTION

Les indices sont caractérisés par des fluctuations saisonnières et irrégulières qui s'ajoutent aux tendances fondamentales. La littérature statistique regorge de méthodes permettant de décomposer et de lisser de telles séries chronologiques. Les estimateurs ponctuels de variance, obtenus par linéarisation, par répétition ou par une autre méthode, peuvent être l'objet des mêmes types de variations saisonnières et irrégulières que les indices eux-mêmes. La nature variable des estimations ponctuelles de variance a été illustrée par Leaver (1990) dans le cas des indices. Dans la présente communication, nous examinons la possibilité d'élaborer des variances généralisées pour les indices de prix, en appliquant des méthodes non paramétriques de lissage de nuages de points à des séries chronologiques d'estimations ponctuelles de variance. Le but visé est de formuler des variances lissées qui sont approximativement sans biais, qui produisent des couvertures acceptables des intervalles de confiance et qui, surtout, sont plus stables que les estimations ponctuelles de variance.

La méthode utilisée ici est quelque peu différente de celle qui est parfois employée dans les enquêtes sur les ménages pour estimer des fonctions de variance généralisées (FVG). Cette méthode est décrite dans Wolter (1985), et certains éléments théoriques sous-jacents sont présentés dans Valliant (1987). L'idée générale est d'utiliser des modèles pour obtenir une approximation des variances. Pour un ensemble donné de variables d'enquête dont les variances suivent toutes le même modèle, les paramètres du modèle sont estimés par la méthode des moindres carrés. Les estimations des paramètres sont alors fournies aux utilisateurs au lieu des estimations individuelles des variances, ce qui permet de condenser les publications contenant les résultats des enquêtes. Idéalement, les modèles produiront aussi des estimations plus stables de la variance. Des applications des FVG dans deux enquêtes particulières sont présentées dans Hanson (1978) et Johnson et King (1987). Dans le cas des indices de prix, il peut être difficile de trouver plusieurs indices dont les variances obéissent au même modèle. Toutefois, le lissage des variances d'un indice particulier dans le temps est une solution de rechange pratique. Pour un indice donné, il s'agit d'un processus en deux étapes, qui consiste à estimer les variances à un certain nombre de points du temps et à lisser la série des estimations ponctuelles de variance. Comme nous allons le montrer, cette méthode permet d'obtenir des estimations de variance plus stables, qui sont approximativement sans biais et qui produisent des couvertures d'intervalles de confiance voisines du niveau théorique.

[1] R. Valliant, U.S. Bureau of Labor Statistics, Room 4915, 2 Massachusetts Avenue N.-E., Washington (DC), É.-U. 20212.

Dans la section 2, nous définissons l'indice de prix de Laspeyres d'une population, une classe d'estimateurs d'indices et un modèle de superpopulation qui sert à étudier la variance des estimateurs d'indices. À la section 3, une approximation de la variance d'un estimateur des variations de prix à long terme est examinée. La section 4 présente les méthodes qui ont été testées pour l'estimation des variances généralisées. Une étude de simulation, décrite à la section 5, a été réalisée à l'aide des données de l'indice des prix à la consommation des États-Unis, pour évaluer la performance des estimateurs de variance proposés en situation pratique. Enfin, les conclusions sont énoncées à la section 6.

2. ESTIMATEURS D'INDICES ET MODÈLE DE SUPERPOPULATION

La population est divisée en H strates, et la strate h contient N_h établissements. L'établissement *(hi)* contient M_{hi} articles, et le nombre total d'articles dans l'ensemble des établissements de la strate h est $M_h = \sum_{i=1}^{N_h} M_{hi}$. Au temps t le prix de l'article j dans l'établissement *(hi)* est p_{hij}^t, et le prix relatif entre le temps t et le temps 0 de la période de base est $r_{hij}^{t,0} = p_{hij}^t / p_{hij}^0$. La quantité de l'article *(hij)* achetée au cours de la période de base est q_{hij}^0. La valeur, pour une population finie, de l'indice de prix de Laspeyres à base fixe à long terme comparant la période t à la période 0 est

$$\begin{aligned} I^{t,0} &= \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} p_{hij}^t q_{hij}^0 \Big/ \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} p_{hij}^0 q_{hij}^0 \\ &= \sum_h \sum_i \sum_j W_{hij}^0 r_{hij}^{t,0}, \end{aligned} \tag{1}$$

où $W_{hij}^0 = p_{hij}^0 q_{hij}^0 / \sum_{h,i,j} p_{hij}^0 q_{hij}^0$ est la fraction de la valeur ou du coût total de la période de base représentée par l'article *(hij)*. Pour référence future, il est également utile de définir l'indice de strate $I_h^{t,0} = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} W_{hij}^0 r_{hij}^{t,0} / W_h^0$ où $W_h^0 = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} W_{hij}^0$. Sur la base d'indices à long terme, l'indice à court terme de la population permettant de comparer les périodes t_2 et t_1 $(t_1 < t_2)$ est défini comme $I^{t_2,t_1} = I^{t_2,0} / I^{t_1,0}$. Les variations mensuelles, trimestrielles, semestrielles et annuelles sont couramment publiés dans le cadre des programmes d'établissement des indices.

Afin d'analyser les propriétés des estimateurs d'indices, nous allons considérer le modèle de superpopulation défini ci-après, qui a aussi été utilisé dans Valliant (1991).

$$\begin{aligned} r_{hij}^{t,0} &= \alpha_{th} + \omega_{thi} + \epsilon_{thij} \\ \epsilon_{thij} &= \rho_h \epsilon_{t-1,hij} + \xi_{thij}, \end{aligned} \tag{2}$$

où $E(\omega_{thi}) = E(\omega_{t,hi}\omega_{t,h'i'}) = 0$ pour tous les t, h, i, et $(t_1 hi) \neq (t_2 h' i')$; $E(\omega_{thi}^2) = \sigma_{wh}^2$; $E(\xi_{thij}) = E(\xi_{t,hij}\xi_{t,h'i'j'}) = 0$ pour tous les t, h, i, j et $(t_1 hij) \neq (t_2 h' i' j')$; $E(\xi_{thij}^2) = \sigma_{\xi h}^2$; et $-1 < \rho_h < 1$. Par convention, définissons $\alpha_{0h} \equiv 1$ et $\epsilon_{0hij} \equiv 0$. Si l'on remonte dans le temps uniquement jusqu'à la période de base et pas au-delà, (2) signifie que $\epsilon_{thij} = \sum_{k=0}^{t-1} \rho_h^k \xi_{t-k,hij}$. Si l'on utilise cette expression et les propriétés de ξ_{thij}, la structure de covariance qu'entraîne le modèle (2) est

$$\operatorname{cov}(r_{hij}^{t_2,0}, r_{h'i'j'}^{t_1,0}) = \begin{cases} \sigma_{wh}^2 + (1-\rho_h^{2t_2})\Delta_h^2 & t_2 = t_1, h = h', i = i', j = j' \\ \rho_h^{t_2-t_1}(1-\rho_h^{2t_1})\Delta_h^2 & t_1 < t_2, h = h', i = i', j = j' \\ \sigma_{wh}^2 & t_2 = h_1, h = h', i = i', j \neq j' \\ 0 & \text{dans les autres cas} \end{cases} \tag{3}$$

où $\Delta_h^2 = \sigma_{\xi h}^2 / (1 - \rho_h^2)$. L'expression (3) signifie que les prix relatifs pour un article particulier sont corrélés dans le temps. À une période donnée, les articles d'un établissement particulier sont également corrélés, tandis que les autres articles ne le sont pas.

Le plan d'échantillonnage examiné ici est celui d'une enquête à groupe de renouvellement dans laquelle les établissements sont sélectionnés à titre d'unités du premier degré. Les établissements demeurent dans l'échantillon pendant une période donnée, puis ils en sont retirés et sont remplacés par de nouvelles unités. À chaque temps $t(t=1,...,T)$, nous avons un échantillon s_{th} de n_h établissements parmi les N_h établissements de la strate h et un échantillon s_{thi} de m_{hi} articles parmi les items M_{hi} articles de l'établissement sélectionné *(thi)*. Un plan d'échantillonnage à deux degrés, qui est souvent l'objet d'une approximation en pratique, est un plan dans lequel les établissements sont sélectionnés avec des probabilités proportionnelles à $W_{hi}^0 = \sum_{j=1}^{M_{hi}} W_{hij}^0$. Les articles des établissements sont ensuite sélectionnés avec des probabilités proportionnelles à W_{hij}^0. Des mesures substituts de la taille, jugées étroitement reliées à W_{hi}^0 et à W_{hij}^0, comme les valeurs des ventes ou les nombres d'emplois courants, sont souvent utilisées en pratique. À chaque période, la taille totale de l'échantillon d'établissements est supposée constante à $n = \sum_h n_h$ le nombre total d'articles sélectionnés dans la strate h étant $m_h = \sum_{i \in s_{th}} m_{hi}$. À chaque période, une proportion δ_h des établissements est retirée de l'échantillon dans la strate h, et un nombre égal d'établissements est ajouté. La taille du chevauchement, $s_{uth} = s_{th} \cap s_{uh}$, des échantillons entre le temps t et le temps $u(t \geq u)$ est $\max\{0, n_h[1-(t-u)\delta_h]\}$.

La classe d'estimateurs considérée ici a été présentée dans Valliant et Miller (1989) pour l'échantillonnage à un seul degré et généralisée dans Valliant (1991). Pour l'indice à long terme, définissons

$$\hat{I}^{t,0} = \sum_h \bar{z}_{th}^{\,t} \prod_{u=1}^{t-1} \left[\frac{\bar{z}_{uth}^{\,u}}{\bar{z}_{u+1,h}^{\,u}} \right]^{\gamma_h^{tu}}, \tag{4}$$

où $\bar{z}_{kth}^{\,u} = \sum_{i \in s_{uh}} \lambda_{hi} \bar{r}_{khi}^{\,u,0}$, $\bar{r}_{khi}^{\,u,0} = \sum_{j \in s_{uhi}} r_{hij}^{u,0} / m_{hi}$ pour $k=u$ ou $u+1$ $(u=1,...,t-1)$ et γ_h^{tu} est un nombre réel. Le terme λ_{hi} est un coefficient qui ne dépend pas des variables aléatoires du modèle $r_{hij}^{t,0}$. Pour le plan à deux degrés avec probabilités proportionnelles à la taille mentionné ci-dessus, par exemple, on a $\lambda_{hi} = W_h^0 / n_h$. Nous limitons notre analyse aux cas où

$$\sum_{i \in s_{uh}} \lambda_{hi} = W_h^0$$

pour lesquels on a alors $E(\bar{z}_{kh}^{\,u} - W_h^0 I_h^{u,0}) = 0$. Quand les échantillons d'établissements de chacune des strates sont tous de grande taille, $\hat{I}^{t,0}$ est approximativement non biaisé selon le modèle, en vertu de (2). Les estimateurs à court terme sont définis sous forme de quotients des estimateurs à long terme. La variation de prix entre les temps t_1 et $t_2(t_1 < t_2)$ est estimée par $\hat{I}^{t_2,t_1} = \hat{I}^{t_2,0} / \hat{I}^{t_1,0}$.

Un certain nombre d'estimateurs de la classe (4) sont énumérés dans Valliant (1991). Trois présentent un intérêt particulier. Si $\gamma_h^{tu} \equiv 1$, l'expression (4) donne l'estimateur produit, qui peut être exprimé sous la forme

$$\hat{I}_1^{t,0} = \sum_h \prod_{u=1}^{t} \left[\frac{\bar{z}_{uth}^{\,u}}{\bar{z}_{u,h}^{\,u-1}} \right]$$

avec $\bar{z}_{1h}^{\,0} \equiv 1$. Si $\gamma_h^{tu} \equiv 0$, l'expression (4) se réduit à l'estimateur d'indice simple

$$\hat{I}_2^{t,0} = \sum_h \bar{z}_{th}^{\,t}.$$

Une troisième possibilité pour γ_h^{tu} est la valeur qui réduit au minimum la variance approximative de $\hat{I}^{t,0}$ en vertu du modèle (2). L'optimum est une expression complexe en général, mais dans le cas spécial où il y a un nombre constant d'articles sélectionnés par établissement, $m_{hi} = \bar{m}_h$, et où λ_{hi} est une constante pour tous les établissements sélectionnés de la strate h, l'optimum se réduit à

$$\gamma_h^{*tu} = \frac{1}{2} \frac{\alpha_{uth}}{\alpha_{th}} \rho_h^{t-u} \left[1 + \bar{m}_h \frac{g_{uth}}{1-g_{uth}} \right]^{-1}$$

pour $1 \leq u \leq t-1$ et $g_{uth} = \sigma_{uth}^2 / \left[\sigma_{uth}^2 + (1-\rho_h^{2u}) \Delta_h^2 \right]$.

3. VARIANCES APPROXIMATIVES EN VERTU DU MODÈLE

Quand la taille de l'échantillon d'établissements n_h est élevée dans chaque strate, l'estimateur de l'indice à long terme peut être approché, comme il est démontré dans l'annexe A de Valliant (1991), par l'expression suivante

$$\hat{I}^{t,0} \approx \sum_h \left\{ \bar{z}_{th}^{\,t} + \sum_{u=1}^{t-1} \gamma_h^{tu} \frac{\alpha_{th}}{\alpha_{uh}} \left(\bar{z}_{uh}^{\,u} - \bar{z}_{u-1,h}^{\,u} \right) \right\}. \tag{5}$$

Avant que nous présentions la variance de cette approximation, il est instructif de comparer l'expression (5) à une expression semblable visant les estimateurs composites. Dans les enquêtes à passages répétés, une forme courante d'estimateur composite (Cantwell 1990) est donnée par

$$\hat{x}_c^{\,t} = (1-k)\hat{x}_{s_t}^{\,t} + k\left(\hat{x}_c^{\,t-1} + \phi_{s_{t,t-1}}^{t} \right),$$

où $\hat{x}_c^{\,t}$ est l'estimateur composite au temps t d'un total, k est un coefficient de pondération situé entre 0 et 1, $\hat{x}_{s_t}^{\,t}$ est un total estimatif basé sur l'échantillon au temps t seulement, et $\phi_{s_{t,t-1}}^{t} = \hat{x}_{s_{t,t-1}}^{\,t} - \hat{x}_{s_{t,t-1}}^{\,t-1}$ est une estimation de la variation entre $t-1$ et t d'après les unités appartenant à l'échantillon aussi bien au temps t qu'au temps t $t-1$. Une substitution répétée donne

$$\hat{x}_c^{\,t} = \tilde{\hat{x}}_c^{\,t} + \sum_{u=1}^{t-1} k^{t-u} \left(\tilde{\hat{x}}_c^{\,u} - \hat{x}_{s_{u+1,u}}^{\,u} \right),$$

où $\tilde{\hat{x}}_c^{\,t} = (1-k)\hat{x}_{s_t}^{\,t} + k\hat{x}_{s_{t,t-1}}^{\,t}$. Ainsi, l'estimateur composite peut être exprimé sous forme d'un estimateur $\tilde{\hat{x}}_c^{\,t}$ au temps t, plus une somme d'estimateurs de 0. Si l'on en juge d'après l'expression (5), il en va à peu près de même pour un estimateur de la classe γ. Par conséquent, la méthode de lissage des variances examinée plus bas devrait aussi s'appliquer à certains types d'estimateurs composites.

En utilisant (5) et les résultats énoncés dans l'annexe de Valliant (1991), nous pouvons exprimer la variance approximative de l'estimateur à long terme de la façon suivante

$$\operatorname{var}(\hat{I}^{t,0}) \approx \sum_h \left\{ \sum_{u=1}^{t-1} a_{tuh} (\iota_h^{t,u})^2 + 2 \sum_{u=1}^{t-1} b_{tuh} \iota^{t,u} + c_{th} \right\}, \tag{6}$$

où

$$\iota_h^{t,u} = \alpha_{th}/\alpha_{uh} = E(I_h^{t,0})/E(I_h^{u,0}),$$

$$a_{tuu} = (\gamma_h^{tu})^2 \left[\sum_{i \in C_{uh}} \frac{\lambda_{hi}^2}{m_{hi}} v_{uhi} + \sum_{i \in D_{uh}} \frac{\lambda_{hi}^2}{m_{hi}} v_{uhi} \right],$$

$$b_{tuh} = \gamma_h^{tu} \rho_h^{t-u} (1-\rho_h^{2u}) \Delta_h^2 \left[\sum_{i \in s_{tuh}} \frac{\lambda_{hi}^2}{m_{hi}} - \sum_{i \in s_{t,u+1,h}} \frac{\lambda_{hi}^2}{m_{hi}} \right],$$

$$c_{th} = \sum_{i \in s_{th}} \frac{\lambda_{hi}^2}{m_{hi}} v_{thi},$$

où $v_{uhi} = v_{uh}\left[1 + (m_{hi}-1)g_{uh}\right]$, $v_{uh} = \sigma_{uh}^2 + (1-\rho_h^{2u})\Delta_h^2$, $C_{uh} = s_{uh} - s_{u+1,h}$, c.-à-d. la partie de s_{uh} qui n'est pas contenue dans $s_{u+1,h}$, et $D_{uh} = s_{u+1,h} - s_{uh}$.

Une expression semblable à celle présentée en (6) peut aussi être obtenue pour la variance approximative de l'estimateur d'indice à court terme $\hat{I}^{t,t-1}$.

4. FONCTIONS DE VARIANCE GÉNÉRALISÉES POUR DES INDICES

D'après l'expression (6), la variance approximative est un polynôme du deuxième degré en $\iota_h^{t,u}$, soit les indices à court termes de superpopulation des strates. Cette relation est analogue à celle qui existe entre un estimateur $\hat{T}$ du total de la population T, dans un échantillonnage à deux degrés, et sa variance approximative établie dans Valliant (1987) pour une classe particulière de modèles dans lesquels la variance d'une unité était une fonction quadratique de la moyenne de l'unité:

$$\text{var}(\hat{T}) \approx aE(T)^2 + bE(T). \tag{7}$$

Les termes a et b sont des coefficients qui dépendent de diverses quantités, notamment les corrélations intra-grappes, la taille de la population et le nombre d'unités sélectionnées à l'intérieur des grappes, et les coefficients de l'estimateur $\hat{T}$. Pour ajuster le modèle de fonction de variance généralisée *(FVG)* défini par (7), la méthode habituelle consiste à sélectionner un groupe de variables qui ont toutes les mêmes coefficients a et b, à calculer des estimations ponctuelles de variance pour chacune des variables, puis à estimer a et b par un processus quelconque des moindres carrés. L'application de cette démarche à l'expression (6) se heurterait à d'importantes difficultés pratiques. Dans (7), il n'y a que deux coefficients de régression à estimer: a et b. Dans (6), il y en a $2(t-1)+1$. Le nombre de coefficients augmente donc avec t. Les composantes des coefficients, a_{cah}, b_{cah}, et c_{ch}, sont elles aussi complexes, de sorte que la détermination de différents indices obéissant tous au modèle (6) poserait des difficultés.

Une autre méthode consiste à travailler avec un indice particulier et à tenter de modéliser le comportement de sa variance dans le temps. Si $\iota_h^{t,u}$ est une fonction lisse du temps, p. ex. un polynôme en $t-u$, la variance (6) sera aussi une fonction lisse du temps, disons $f(t)$. Si un estimateur de variance sans biais, ou approximativement sans biais, est utilisé pour $\hat{I}^{t,0}$, son espérance peut aussi être décrite par $f(t)$. Avec l'accumulation des données au fil du temps, une série chronologique d'estimations ponctuelles de variance est constituée et la fonction $f(t)$ peut être ajustée par une méthode de lissage de nuage de points, sans qu'on ait besoin de savoir la forme explicite de la fonction. Un certain nombre de telles méthodes de lissage sont disponibles, et nous en examinerons deux qui se sont révélées utiles dans d'autres situations.

Les deux méthodes de lissage utilisées ici sont le «super smoother» (Friedman 1984) et le «loess» (Cleveland 1979, Cleveland, Cleveland, McRae et Terpenning 1990). La description détaillée des deux algorithmes est relativement complexe, de sorte que nous nous contenterons ici d'en tracer les grandes lignes. Les deux méthodes utilisent des ajustements linéaires locaux dans des voisinages entourant chaque point t. Un paramètre crucial des deux algorithmes est l'étendue, c.-à-d. la taille du voisinage autour de t, qui sert à estimer $f(t)$. Dans le cas du «loess», l'étendue est fixe, tandis que le «super smoother» peut utiliser des étendues variables. Le «loess» incorpore explicitement des caractéristiques visant à réduire les effets des valeurs extrêmes et, des deux, il est celui qui a tendance à produire la courbe d'estimations apparaissant la plus lisse. Les étendues variables utilisées par le «super smoother» permettent à ce dernier de s'adapter plus facilement aux variations de la courbure de $f(t)$. Le «super smoother» a aussi l'avantage d'exiger moins de temps de calcul que le «loess».

5. UNE ÉTUDE EMPIRIQUE

Une étude de simulation, utilisant une population formée par le BLS à partir de données recueillies dans le cadre du programme de l'indice des prix à la consommation des États-Unis, a été réalisée pour évaluer l'utilité de la méthode proposée de calcul des *FVG*. La population, composée d'établissements et d'articles, a été décrite en détail dans Valliant (1991). Nous ferons ici un bref rappel de ses principales caractéristiques. Les 659 établissements formant la population ont été divisés en cinq strates. Chaque établissement contenait un nombre moyen d'articles tout juste inférieur à dix, et l'on disposait du prix de chaque article pour 42 mois consécutifs.

Deux ensembles de 500 échantillons stratifiés à deux degrés ont été prélevés, le nombre d'établissements sélectionnés attribués à chaque strate étant grosso modo proportionnel à W_h^0. Les tailles globales des échantillons d'établissements dans les deux ensembles étaient respectivement n = 50 et 100. Les échantillons ont été prélevés de telle façon que 20 % des établissements de l'échantillon étaient renouvelés à chaque période de 12 mois. À cette fin, on prélevait d'abord un vaste échantillon systématique, à origine choisie au hasard, d'établissements dans chaque strate avec probabilités proportionnelles à W_{hi}^0. Pour les échantillons de taille n = 50, la taille du grand échantillon initial était de 84, tandis que pour les échantillons de taille n = 100, elle était de 168. Ces échantillons initiaux étaient de taille suffisante pour couvrir l'ensemble de la période de 42 mois, en tenant compte du renouvellement des établissements. L'échantillon initial de chaque strate était ensuite trié selon un ordre déterminé au hasard. Pour une période particulière t, l'échantillon d'établissements de la strate était formé des établissements $1+(t-1)n_h\delta_h,...,n_h+(t-1)n_h\delta_h$, où δ_h était la proportion des établissements renouvelés au cours d'un mois. Tant pour le cas n = 50 que pour le cas n = 100, δ_h = 1/60 ce qui correspondait à un renouvellement annuel de $12(n_h/60)=n_h/5$ établissements, ou 20 %. De chaque établissement de l'échantillon, $\bar{m}_h$ = 2 articles étaient prélevés systématiquement avec probabilité proportionnelle à W_{hij}^0.

Pour chaque échantillon, les estimateurs produits à long terme $\hat{I}_1^{t,0}$ (t=1,...,42), et les estimateurs à court terme de la variation sur 1 mois et de la variation sur 12 mois ont été calculés. Le cas spécial de $\lambda_{hi}=W_h^0/n_h$ a été utilisé, ce qui produit un estimateur non biaisé selon le plan, pour le plan d'échantillonnage appliqué à l'étude de simulation. Des résultats plus détaillés de cette simulation sont fournis dans Valliant (1992).

Les estimations ponctuelles de variance ont été obtenues par la méthode de linéarisation, et ont été décrites en détail dans Valliant (1991). Il convient de souligner que les résultats présentés ici ne dépendent pas de l'emploi d'une méthode particulière d'estimation ponctuelle de variance. Les estimations obtenues par la méthode BRR (balanced repeated replication), le jackknife ou une autre méthode conviendraient tout aussi bien, pourvu que ces estimateurs de variance soient convergents ou approximativement sans biais. Chaque échantillon a fait l'objet d'une estimation de variance par la méthode de linéarisation pour chacune des estimations des indices à long terme et à court terme et pour chacune des périodes mentionnées ci-dessus. Deux *FVG* - «super smoother» et «loess» - ont ensuite été calculées pour chaque série d'estimations d'indice. Par exemple, pour l'estimation produit de l'indice à long terme, une série de 42 estimations ponctuelles de variance a été produite pour chaque échantillon. Les valeurs du «super smoother» et du «loess» ont été calculées pour chaque échantillon, par l'application de ces méthodes à la série de 42 estimations par linéarisation correspondant à chaque estimation d'indice. Les calculs de simulation ont été exécutés en double précision à l'aide du *Turbo Pascal* de Borland. Les *FVG* ont été calculées à l'aide du logiciel *S-Plus for DOS* de Statistical Sciences Inc.

Des statistiques sommaires ont ensuite été calculées pour l'ensemble des 500 échantillons. Les racines carrées des erreurs quadratiques moyennes empiriques ont été calculées sous la forme $\left[\sum(\hat{I}-I)^2/500\right]^{1/2}$, la sommation portant sur les 500 échantillons, $\hat{I}$ représentant l'un des estimateurs à long terme ou à court terme et I étant l'indice de la population défini à la section 2. Les racines carrées de la moyenne des estimations de variance ont été calculées pour chaque période sous la forme $\sqrt{\bar{v}}$ où $\bar{v}=\sum_{s=1}^{500} v_s/500$ et v_s est l'un des trois types d'estimation de variance (linéarisation, «super smoother» ou «loess») à une période particulière pour l'échantillon s. Cette opération a été effectuée séparément pour les estimations produits visant les variations de prix à long terme et sur 1 mois.

Des résultats sommaires pour l'ensemble des échantillons et des périodes sont présentés au tableau 1. Les rapports (en pourcentage) entre la racine carrée de l'estimation moyenne de variance et la racine de l'erreur quadratique moyenne (RMSE) empirique sont quelque peu inférieurs à 100 dans tous les cas, c'est-à-dire qu'aussi bien l'estimation ponctuelle de variance ($\hat{v}$) que la *FVG* sont des sous-estimations, mais le problème est minime. Dans tous les cas, les *FVG* sont plus stables que $\hat{v}$. Par exemple, au tableau 1, l'écart-type de la *FVG* du «super smoother» correspond à 61 % de celui de $\hat{v}$ pour la variation sur 1 mois quand n=100. Pour le même cas, la *FVG* du «loess» a un écart-type représentant 57 % de celui de $\hat{v}$. Les plus grands gains de stabilité sont obtenus pour les variations de prix sur 1 mois, tandis que les gains les plus faibles se produisent

dans le cas de la variation à long terme. Les estimations du «loess» sont généralement plus précises que celles du «super smoother», tandis que l'amélioration par rapport à l'estimation par linéarisation est un peu inférieure pour la plus grande taille d'échantillon. Le tableau 1 indique aussi les couvertures empiriques des intervalles de confiance à 95 % pour l'ensemble des 42 périodes. Des intervalles de confiance d'approximation normaux ont été calculés de la façon habituelle, sous la forme $\hat{I} \pm 1.96\sqrt{v}$ où $\hat{I}$ est l'indice à long terme ou à court terme et v est l'une des estimations de variance. Bien que toutes les estimations de variance donnent une couverture inférieure au niveau théorique de 95 %, le plus petit pourcentage figurant au tableau 1 est 92.0 %, et les *FVG* soutiennent très bien la concurrence de $\hat{v}$.

Tableau 1: Résultats de la simulation pour l'estimateur produit, d'après 500 échantillons par grappes avec moyenne établie sur 42 périodes. Tous les chiffres sont des pourcentages. $\hat{v}$ désigne l'estimation de variance obtenue par linéarisation.

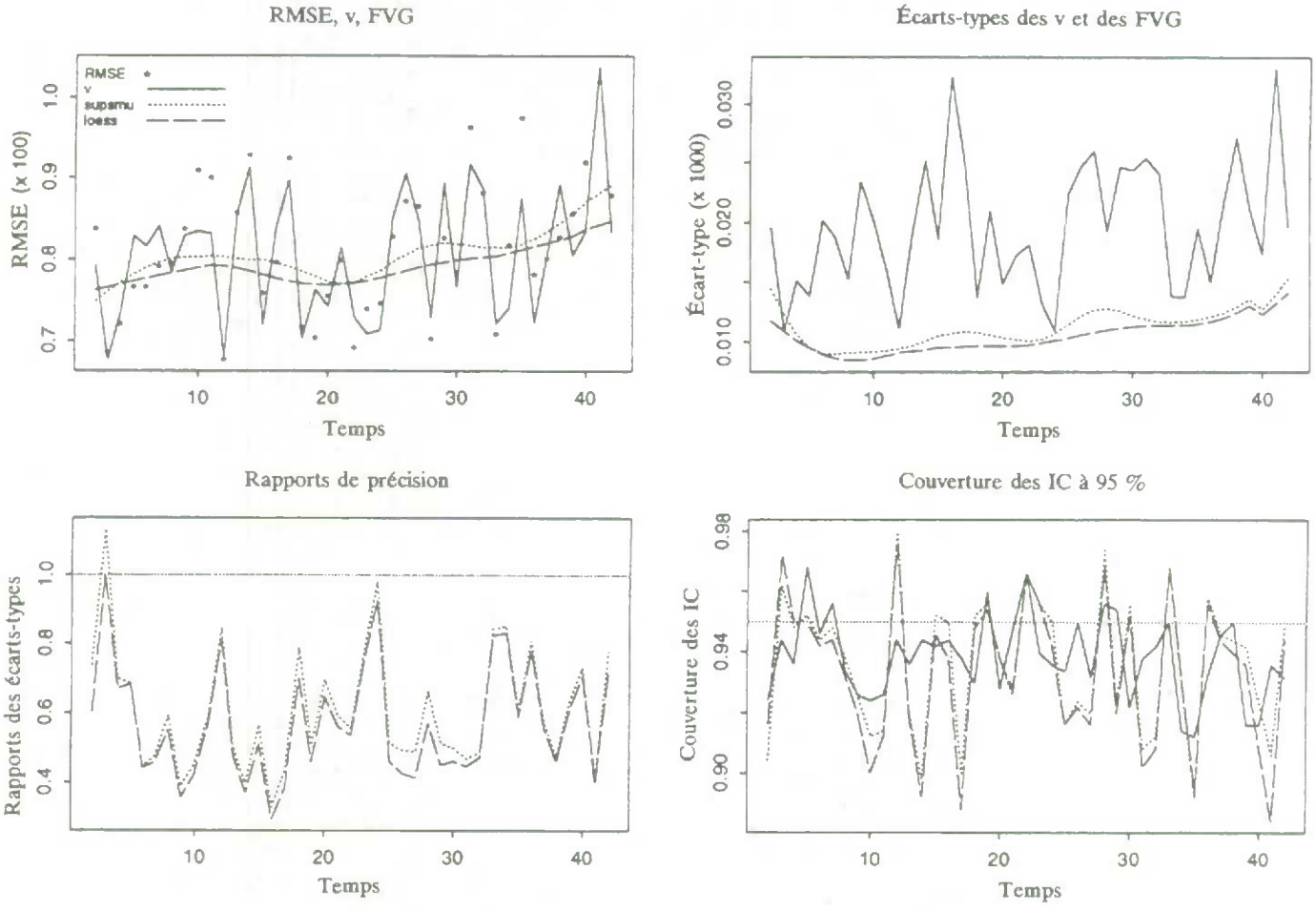
	$\sqrt{v}$/RMSE			Écart-type (*GVF*) /Écart-type ($\hat{v}$)		Couverture des IC à 95%		
	$\hat{v}$	Supsmu	Loess	Supsmu	Loess	$\hat{v}$	Supsmu	Loess
n=50								
LT	99.5	99.0	98.0	90.8	86.8	93.7	93.1	92.8
1-mois	99.3	99.4	96.8	56.8	50.8	93.3	93.6	92.9
12-mois	95.0	94.8	93.9	74.3	71.0	92.0	92.3	92.1
n=100								
LT	99.8	99.2	98.5	96.1	92.1	94.2	93.4	93.3
1-mois	99.3	99.8	98.1	61.0	57.0	93.8	93.6	93.3
12-mois	96.1	95.8	96.0	79.1	79.6	93.2	93.4	93.3

La figure 1 présente des graphiques de statistiques sommaires pour les 500 échantillons, selon les périodes, pour n=100. Les graphiques ne sont fournis que pour l'estimateur produit des variations sur 1 mois. Le graphique supérieur gauche de la figure présente les erreurs quadratiques moyennes (RMSE) empiriques et la racine carrée de la moyenne de chaque *FVG* en fonction du temps. Les *FVG* sont beaucoup plus lisses que $\hat{v}$, comme on pouvait s'y attendre. Bien qu'aucune méthode de lissage ne subisse démesurément l'influence des valeurs extrêmes parmi les $\hat{v}$, le «super smoother» suit de plus près que le «loess» les fluctuations des courbes $\hat{v}$, Il s'ensuit que le «super smoother» présente en général un écart-type plus élevé que le «loess», comme le montre le graphique supérieur droit de la figure. Le graphique inférieur gauche montre le rapport entre l'écart-type des *FVG*, pour les 500 échantillons, et l'écart-type de $\hat{v}$. Encore une fois, on constate que les deux *FVG* sont plus précises que l'estimation par linéarisation, les gains étant particulièrement élevés pour la variation sur 1 mois. Le graphique inférieur droit de la figure montre la couverture des intervalles de confiance à 95 % en fonction du temps. Les *FVG* produisent des couvertures raisonnablement bonnes, presque égales à celles des estimations ponctuelles de variance. Les périodes pour lesquelles les *FVG* donnent une couverture notablement plus faible que $\hat{v}$ sont celles pour lesquelles les fonctions de lissage ne suivent pas étroitement les fluctuations vers le haut de $\hat{v}$.

La modélisation des séries chronologiques est un domaine qui est l'objet de nombreux travaux, et il existe de nombreux autres choix de méthodes de lissage de séries chronologiques qui pourraient fonctionner tout aussi bien que celles examinées ici. Une analyse récente est présentée dans Kohn, Ansley et Wong (1992). Une autres possibilité, que nous n'avons pas explorée, serait de calculer une moyenne pondérée d'une variance lissée et de l'estimation ponctuelle de variance à chaque période. Cette solution pourrait être avantageuse si l'on croit

que les estimateurs ponctuels de variance sont davantage exempts de biais que les estimations lissées, en raison de l'incapacité de la variance approximative (6) d'être une fonction lisse du temps.

Figure 1: Tracés sommaires des résultats de la simulation pour l'estimateur produit de la variation sur 1 mois, d'après 500 échantillons de taille n=100 établissements. La légende du graphique supérieur gauche s'applique à chacun des autres graphiques. v désigne l'estimateur par linéarisation; supsmu désigne le «super smoother».



6. CONCLUSION

Dans les enquêtes permanentes qui produisent des séries chronologiques d'estimations, les méthodes étudiées ici pour le lissage d'estimations de variance apparaissent très utiles. Dans le cas des enquêtes permanentes où le plan d'échantillonnage et la taille de l'échantillon demeurent les mêmes pendant de longues périodes, les utilisateurs s'attendent à ce que les variances suivent des courbes lisses dans le temps, caractéristique que n'ont pas, en général, les estimations ponctuelles de variance. Un tel souhait des utilisateurs peut sembler, au premier abord, déraisonnable du point de vue statistique, car les erreurs quadratiques moyennes réelles peuvent varier avec le temps. Toutefois, pour les indices de prix, nous avons montré, en utilisant la théorie des grands échantillons et des simulations connexes, qu'on peut obtenir des estimations de variance lissées, approximativement sans biais, qui sont plus stables que les estimations ponctuelles de variance pour les variations de prix aussi bien à court terme qu'à long terme, et qui donnent par ailleurs des couvertures d'intervalles de confiance voisines des niveaux théoriques. Par conséquent, dans la situation étudiée ici, les variances lissées ont une justification statistique, et sont en outre beaucoup plus attrayantes du point de vue de la présentation aux utilisateurs des données. Par conséquent, les variances lissées méritent d'être prises en considération, tant pour l'analyse interne que pour la publication de données.

AVIS

Les opinions exprimées dans cette communication sont celles de l'auteur et n'engagent en rien le Bureau of Labor Statistics.

BIBLIOGRAPHIE

Cantwell, P. (1990). Formules de variance pour estimateurs composites dans les plans de renouvellement. *Techniques d'enquête* 16, 163-174.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

Cleveland, R.B., Cleveland, W.S., McRae, J.E., et Terpenning, I. (1990). STL: A seasonal decomposition procedure based on loess. *Journal of Official Statistics*, 6, 3-32.

Friedman, J.H. (1984). A variable span smoother. Rapport technique n° 5. Laboratory for Computational Statistics, Stanford University.

Hanson, R.H. (1978). *The current population survey: Design and methodology*. Papier technique 40, Washington DC, U.S. Bureau of the Census.

Johnson, E.G., et King, B.F. (1987). Generalized variance functions for a complex sample survey. *Journal of Official Statistics*, 3, 235-250.

Kohn, R., Ansley, C., et Wong, C.-M. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika*, 79, 335-346.

Leaver, S.G. (1990). Estimating variances for the U.S. consumer price index for 1978-1986. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 290-295.

Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82, 499-508.

Valliant, R. (1991). Variance estimation for price indexes from a two-stage sample with rotating panels. *Journal of Business and Economic Statistics*, 9, 409-422.

Valliant, R. (1992). Smoothing variance estimates for price indexes over time. *Journal of Official Statistics*, 9, sous presse.

Valliant, R., et Miller, S.M. (1989). A class of multiplicative models for Laspeyres price indexes. *Journal of Business and Economic Statistics*, 7, 387-394.

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

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

NOUVEAUX DÉVELOPPEMENTS DANS L'ESTIMATION COMPOSITE POUR L'ENQUÊTE «CURRENT POPULATION SURVEY»

P.J. Cantwell et L.R. Ernst[1]

RÉSUMÉ

En janvier 1994, le Bureau of the Census des É.-U. apportera plusieurs modifications à l'enquête «Current Population Survey (CPS)». Ces changements risquent d'influencer le calcul de la valeur des caractéristiques de l'activité ou encore la structure du biais rattachée aux huit groupes de renouvellement. Nous nous proposons ici d'examiner s'il est souhaitable ou nécessaire de modifier l'estimateur composite actuel. Nous traitons plusieurs questions clés qui concernent l'estimateur: 1) Quel est le «meilleur» estimateur pour janvier 1994 et les mois suivants? 2) Devrait-on changer de plan de renouvellement plus tard? 3) Quelle forme devrait avoir l'estimateur composite à longue échéance? En considérant des mesures d'erreur appropriées, nous soumettons des recommandations à l'égard des deux premières questions. Quant à la troisième, nous examinons brièvement une méthode selon laquelle des estimateurs composites distincts sont élaborés pour différentes caractéristiques sans nuire à la cohérence des totaux pour les sous-groupes.

MOTS-CLÉS: Situation vis-à-vis de l'activité; plan de renouvellement; estimateur composite AK; biais de conditionnement; totaux de contrôle démographiques.

1. INTRODUCTION

1.1 Modifications à la CPS et cadre de cette communication

L'enquête «Current Population Survey (CPS)», parrainée par le Bureau of Labor Statistics et exécutée par le Census Bureau, sert à mesurer la situation vis-à-vis de l'activité aux États-Unis. À l'heure actuelle, environ 20% des interviews sont réalisées à l'un ou l'autre des deux centres d'interview téléphonique assistée par ordinateur (ITAO). Cette proportion est censée augmenter d'ici l'an 2000. Le reste des interviews (environ 80%) sont des interviews papier et crayon effectuées dans les régions.

En janvier 1994, on prévoit éliminer la formule classique «papier-crayon» et la remplacer par la formule de l'interview personnelle assistée par ordinateur (IPAO), avec ordinateur portatif. Le questionnaire doit aussi être entièrement révisé. Le nouveau questionnaire permettra d'évaluer plusieurs nouvelles caractéristiques et de redéfinir celles déjà existantes. La modification des procédures risque d'influencer de diverses manières le calcul de la valeur des caractéristiques de l'activité. Pour se préparer à ce changement, le Census Bureau a commencé à utiliser en janvier 1992 un groupe expérimental de ménages appelé «panel de chevauchement ITAO-IPAO» (CII). Ce panel sera en fonction jusqu'en décembre 1993.

Exception faite des modifications requises, l'instrument du panel CII est le même que celui qui sera utilisé dans la CPS régulière en janvier 1994 et dans les mois suivants. En outre, lorsque le panel CII sera entièrement en fonction, les 15 000 ménages de l'échantillon seront interviewés (chaque mois) selon le même plan de renouvellement que dans la CPS. Notons que durant sa période d'opération, le panel CII est entièrement indépendant de l'enquête régulière; aucune des données tirées de ce panel ne sert à l'établissement d'estimations ou n'est publiée et aucun des membres de ce panel ne demeurera dans l'échantillon en 1994.

---

[1] P. Cantwell, agent principal de recherche, et L. Ernst, chef adjoint, Statistical Research Division, Bureau of the Census, Washington (DC), É.-U. 20233.

Comme ces changements doivent se faire en une seule fois en janvier 1994, on étudie aussi la possibilité de modifier l'estimateur composite courant. Les sous-sections 1.2 et 1.3 contiennent des renseignements sur le plan de renouvellement actuel de la CPS et l'estimation composite. Dans la section 2, nous tentons de déterminer quel serait le meilleur estimateur pour janvier 1994 et les mois suivants. Nous étudions la variance et le biais de conditionnement rattachés à l'estimation de valeurs mensuelles et de variations d'un mois à l'autre selon trois scénarios: i) conserver l'estimateur composite AK habituel, ii) ne pas utiliser l'approche composite en janvier 1994, et iii) utiliser l'approche composite avec le panel CII.

Dans une perspective à plus long terme, la section 3 examine divers plans de renouvellement, selon lesquels les ménages sont interviewés durant une période de six ou huit mois consécutifs. Dans la section 4, nous considérons des changements plus profonds pour l'estimateur. Nous examinons d'autres formes d'estimateur composite et décrivons une méthode (proposée par Fuller) qui définit des estimateurs composites différents pour chaque caractéristique et qui utilise les totaux de contrôle correspondants pour assurer la cohérence des totaux entre les divers sous-groupes.

1.2 Plan de renouvellement actuel de la CPS

Dans la CPS, les ménages échantillonnés sont interviewés quatre mois consécutifs, passent les huit mois suivants hors de l'échantillon, puis reviennent dans l'échantillon pour quatre autres mois. Pour plus de renseignements sur ce plan 4-8-4 ou sur la CPS en général, voir le document technique n° 40, U.S. Bureau of the Census (1978). À titre de comparaison, les participants à l'Enquête sur la population active (EPA) de Statistique Canada demeurent dans l'échantillon six mois consécutifs avant d'en être retirés.

L'échantillonnage répété offre plusieurs avantages au point du vue du coût. Certains frais généraux sont engagés une seule fois pour chaque ménage; les interviews téléphoniques -- beaucoup moins coûteuses que les interviews sur place -- conviendraient mieux après le contact initial. Toutefois, notre attention porte uniquement sur les effets du biais et de la variance. Les coefficients de chevauchement des plans de renouvellement de la CPS et de l'EPA -- 75 et 83% respectivement -- réduisent la variance des estimations de variation.

1.3 Estimation composite

Pour tirer profit davantage du chevauchement des ménages pendant des mois consécutifs, la CPS utilise un estimateur composite. Pour une caractéristique particulière, posons $x_{h,i}$ comme le total estimé de cette caractéristique pour le mois h relativement au groupe de renouvellement qui est interviewé pour la *i-ième* fois ce mois-là. Il y a huit estimations de ce genre à chaque mois. On définit une estimation par quotient simple au moyen de l'expression $Y_h = (1/8) \Sigma x_{h,i}$. (Il s'agit d'un estimateur par quotient car le poids rattaché à chaque répondant est calculé à la suite de plusieurs ajustements.)

Il y a une vingtaine d'années, la CPS utilisait à l'origine un «estimateur composite simple». Définissons $\Delta = (1/6) \{ \Sigma x_{h,i} - \Sigma x_{h-1,i-1} \}$ -- les sommations étant étendues aux groupes de renouvellement communs aux mois h et $h-1$ -- pour estimer l'écart entre les mois. Au lieu de calculer l'estimation par quotient du total pour le mois h, on peut aussi bien faire la somme du total estimé du mois précédent et de l'écart estimé Δ (c'est ce qu'on appelle ici une «estimation de variation»). L'estimateur composite simple pour le mois h se définit donc $Y_h' = (1\text{-K}) Y_h + \text{K} (Y_{h-1}' + \Delta)$, soit comme une combinaison linéaire de l'estimation par quotient simple et de l'estimation de variation pour le mois h. Auparavant, la CPS utilisait cet estimateur avec K = 0.5.

Si on perfectionne l'estimateur composite, en n'introduisant qu'un seul paramètre de plus, on obtient l'estimateur composite AK, élaboré par Gurney et Daly (1965). Soit $\beta = (1/8) \{ \Sigma x_{h,i} - (1/3) \Sigma x_{h,j} \}$, où $i = 1,5$ et $j = 2,3,4; 6,7,8$. L'estimateur composite AK est défini $Y_h'' = (1\text{-K}) Y_h + \text{K} (Y_{h-1}'' + \Delta) + \text{A } \beta$. L'introduction du paramètre A a pour conséquence de rapprocher les coefficients des groupes de renouvellement de ceux des estimateurs linéaires à variance minimum et d'atténuer les effets du biais de conditionnement (décrit dans la sous-section 2.1). La CPS a commencé à utiliser cet estimateur, avec A = 0.2 et K = 0.4, vers le milieu des années 1980.

Pour en savoir davantage sur l'évolution de l'estimation composite, se référer à Gurney et Daly (1965), Wolter (1979), Kumar et Lee (1983), Breau et Ernst (1983), Adam et Fuller (1992) ou encore, Yansaneh et Fuller (1992).

2. QUEL ESTIMATEUR POUR JANVIER 1994?

2.1 Modifications prévues pour 1994 et autres considérations

Nous avons exposé dans la sous-section 1.1 plusieurs modifications implantées en une seule fois et qui doivent être apportées à la CPS en janvier 1994. Quant aux modifications permanentes dont doivent faire l'objet les estimateurs, elles seront apportées à une date ultérieure, probablement au début de 1996. Nous espérons que le changement de questionnaire prévu pour janvier 1994 affectera le moins possible la série chronologique d'estimateurs mais nous sommes conscients que la nouvelle méthodologie peut provoquer une variation brusque à ce moment-là.

Nous avons considéré trois estimateurs pour les besoins de cette communication. Ce sont les suivants:

(E1) Composite -- forme habituelle -- en janvier 94 et dans les mois suivants, avec A = 0.2 et K = 0.4.

(E2) Estimation par quotient au départ (une simple moyenne des estimations tirées des huit groupes de renouvellement) en janvier 94; composite - février + janvier 94 (A = 0.2, K = 0.4); composite - mars + février + janvier 94; etc.

(E3) Composite - janvier 94 + estimateur composite de décembre 93 tiré du panel CII (A = 0.2 et K = 0.4); composite - février + janvier 94 (A = 0.2, K = 0.4); composite - mars + février + janvier 94; etc. Pour estimer la variation d'un mois à l'autre en janvier 94, soustraire l'estimation de décembre 93 tirée du panel CII de l'estimation de janvier 94.

Plusieurs facteurs peuvent influer sur le choix de l'estimateur à utiliser en 1994. Premièrement, l'ancien et le nouveau questionnaires ne sont pas parfaitement compatibles, c'est-à-dire que de nouvelles questions seront ajoutées et le choix des réponses sera modifié dans certaines questions existantes. Ce facteur doit être pris en considération sérieusement si nous utilisons un estimateur composite qui intègre des données de 1993 et de 1994 et si nous nous intéressons à la variation d'un mois à l'autre en janvier 94.

Deuxièmement, il est possible que la nouvelle méthodologie suscite une variation brusque de la valeur des caractéristiques de l'activité, notamment du nombre de chômeurs (C), en janvier 1994. N'importe quel estimateur, y compris l'estimateur par quotient simple, peut être biaisé. Cependant, même s'il n'y a pas de variation réelle de valeur entre décembre 93 et janvier 94, l'application d'une nouvelle méthodologie peut modifier la valeur probable de l'estimateur par quotient ou d'autres estimateurs.

Troisièmement, nous avons des raisons de croire que la nouvelle méthodologie amènera une modification de la structure du biais de conditionnement. Bailar (1975) définit et analyse les effets qui découlent du conditionnement du panel dans la CPS et que l'on désigne par l'appellation «effet de conditionnement». Pour expliquer brièvement cet effet, disons que pour n'importe quel mois donné et n'importe quelle caractéristique à estimer, les valeurs probables des estimations des huit groupes de renouvellement ne sont généralement pas égales mais reflètent le nombre d'interviews antérieures ou d'autres facteurs. En reprenant la notation exposée dans la section précédente, nous pouvons définir l'*indice* de biais pour le *i-ième* mois de présence dans l'échantillon par l'expression $E(x_{h,i}) / E(\Sigma x_{h,j}/8)$, de sorte qu'une valeur d'indice supérieure à 1 suppose qu'il y a surestimation dans ce mois par rapport aux sept autres mois. Dans notre étude, nous supposons l'estimateur par quotient simple non biaisé pour créer un point de comparaison.

Par une analyse récente des réponses données dans la CPS entre 1980 et 1987, Adams (1991) a établi des estimations des indices de biais pour la population active en chômage (C) et la population active civile (PAC) suivant la méthodologie actuelle. Par ailleurs, dans une analyse de la mise en application progressive de l'ITAO, Shoemaker (1992) obtient des indices très différents de ceux de Adams pour la population active en chômage. (Il n'existe pas encore d'indices ITAO pour la PAC.) Comme toutes les interviews seront assistées par

ordinateur en 1994, il est très probable que les nouveaux indices se rapprochent de ceux calculés suivant la méthodologie de l'ITAO.

De plus, le nouvel instrument d'enquête renfermera un certain nombre de questions qui seront posées chaque mois aux travailleurs découragés. On pense que ces questions ont généralement pour conséquence d'accroître légèrement la valeur de C. Avant 1970, ces questions étaient posées uniquement aux 1er et 5e mois de présence dans l'échantillon. Depuis 1970, elles sont posées aux 4e et 8e mois. Bailar (1975) présente des estimations des indices de biais correspondants pour les deux périodes. Nous pouvons ainsi mesurer l'effet de ces questions sur le biais de conditionnement et prévoir la valeur des indices (appelés «indices QTD», pour «questions sur les travailleurs découragés») si ces questions reviennent à chaque mois.

Le tableau 1 donne, pour les huit mois de présence dans l'échantillon, les indices de biais de trois types (courants, QTD et ITAO) pour C (population active en chômage) et les indices de deux types (courants et QTD) pour la PAC.

Tableau 1: Indices de biais utilisés dans l'étude de divers estimateurs.

Mois de présence dans l'échantillon	1	2	3	4	5	6	7	8
Population active en chômage								
Indices courants	1.070	1.004	.988	1.002	1.004	.956	.965	1.012
Indices QTD	1.125	1.030	.993	.949	1.037	.987	.954	.925
Indices ITAO	.98	1.07	1.08	.98	.94	1.03	.96	.96
Population active civile								
Indices courants	1.016	1.002	.996	1.002	.998	.992	.993	1.000
Indices QTD	1.017	1.004	.999	.997	1.002	.994	.994	.993

Notons que la mise en application de la nouvelle méthodologie créera une combinaison exceptionnelle de biais pour les répondants de janvier 1994. Par exemple, un groupe de renouvellement sera interviewé pour la quatrième fois ce mois-là, mais ce sera la première fois avec le nouveau questionnaire. De fait, d'autres erreurs systématiques, que nous ne saurions mesurer à l'avance, pourraient survenir dans ces conditions.

Quatrièmement, comme d'habitude, les trois quarts des ménages qui auront participé à la CPS de décembre 93 demeureront dans l'échantillon en janvier 94. Un estimateur comme E1 peut être amélioré par les estimations corrélées qui proviennent de groupes de renouvellement communs. En revanche, le panel de chevauchement ITAO-IPAO (CII), qui est en opération jusqu'en décembre 1993, ne comptera aucun ménage dans l'échantillon en 1994.

Enfin, la structure de corrélation des groupes de renouvellement communs peut changer. Si $x_{h,i}$ et $x_{h+r,j}$ sont des estimations du même groupe de renouvellement à r mois d'intervalle, les valeurs sont corrélées en fonction de r. En calculant C (population active en chômage) lors d'études antérieures (Breau et Ernst, 1983; Adam et Fuller, 1992), des coefficients de corrélation d'environ 0.50 furent obtenus lorsque $r = 1$; ces coefficients tombent à environ 0.20 lorsque $r = 15$. Les valeurs correspondantes pour la PAC (population active civile) étaient beaucoup plus élevées; d'environ 0.80 au départ ($r = 1$), elles tombaient à 0.55 ($r = 15$). Dans notre étude, nous nous sommes servis de coefficients légèrement lissés par rapport à ceux calculés dans les ouvrages mentionnés.

Comme le nouvel instrument d'enquête repose plus largement sur l'interview avec rétro-information, il se peut que certains coefficients de corrélation augmentent en 1994. Pourtant, nous avons supposé que les corrélations inter-mensuelles pour 1994 ne varieraient pas avec l'introduction du nouveau questionnaire. Toutefois, la

corrélation *entre une estimation calculée en 1993 et une autre calculée en 1994*, pour un même groupe de renouvellement, diminuera vraisemblablement -- ne serait-ce que légèrement -- à cause du changement de méthodologie. Dans notre analyse, nous avons supposé une baisse de 10% des coefficients de corrélation habituels dans ces conditions.

2.2 Qualités des estimateurs

Pour décider de l'estimateur à utiliser en janvier 1994, les directeurs de l'enquête examinent les estimateurs en lice sous trois aspects: conditions d'utilisation; variance; et biais. En ce qui concerne le dernier aspect, nous avons calculé dans notre étude l'écart par rapport à la valeur probable de l'estimateur composite stationnaire (c.-à-d., de l'estimateur composite dans la forme qu'il prend une fois que tous les effets de la transition se sont fait sentir). Comme nous l'avons déjà mentionné, l'application d'une nouvelle méthodologie pourra entraîner une variation exceptionnelle de la valeur des caractéristiques de l'activité. Puisqu'on ne connaîtra jamais la valeur réelle de ces caractéristiques, il est à espérer que ce changement se produira d'un seul coup (janvier) et que l'estimateur composite habituel prendra sa forme stationnaire le plus rapidement possible. Autrement dit, on recherche un estimateur dont le biais évoluera le plus rapidement possible vers celui d'un estimateur composite à long terme.

Avant de présenter les résultats des calculs, examinons brièvement les trois estimateurs définis dans la sous-section 2.1. L'estimateur E1 correspond à l'estimateur composite habituel; il combine des données de 1993 avec des données de janvier 1994 et des mois suivants. Comme il «tire profit» des corrélations entre des groupes de renouvellement communs, il devrait être l'estimateur qui a la variance la moins élevée dans la plupart des cas. Cependant, combiner des données de questionnaires différents peut être complexe en pratique. En outre, les effets de conditionnement pourraient être instables pendant plusieurs mois en 1994 à cause du changement de méthodologie. On pourrait ainsi se retrouver avec un estimateur de la variation d'un mois à l'autre dont l'espérance est non nulle.

Le deuxième estimateur, E2, est simple dans la pratique. En débutant avec un estimateur par quotient simple en janvier 1994, puis en enchaînant avec un estimateur composite pour les premiers mois de l'année seulement, on introduit une coupure nette par rapport aux effets de conditionnement de l'ancienne méthode. En revanche, en négligeant les groupes de renouvellement qui auront fait partie de l'échantillon à la fin de 1993, on renonce à une diminution probable de la variance. De plus, pour les trois séries de structures de biais que nous considérons, il s'écoulera plusieurs mois avant que les nouveaux biais équivalent à ceux d'un estimateur composite stationnaire.

Dans le cas du troisième estimateur, E3, qui combine des données de janvier 1994 avec des données du panel CII, on se sert du même instrument avant comme après le 1er janvier 1994. Le problème de la compatibilité des questionnaires ne se pose donc pas. De plus, les nouveaux biais équivaudront presque sur-le-champ à ceux de l'estimateur composite stationnaire. (Rappelons-nous cependant la remarque de la sous-section 2.1 sur la combinaison exceptionnelle de biais pour les répondants de janvier.) Malheureusement, E3 présente deux inconvénients majeurs: i) l'effectif du panel CII équivaut à environ le quart de l'effectif de l'échantillon de l'enquête régulière, et ii) il n'existe aucun point de chevauchement entre le panel CII de 1993 et l'échantillon de la CPS de 1994. Ces inconvénients se traduisent par une hausse notable des variances pertinentes.

2.3 Résultats des calculs pour janvier 1994 et les mois suivants

Pour simplifier l'étude, nous avons commencé par comparer les variances des trois estimateurs en lice. Les problèmes que pose E3 ont été évoqués dans la sous-section précédente. De fait, pour janvier 1994, cet estimateur a une variance d'environ 35% supérieure à celle de l'estimateur composite stationnaire pour l'estimation de la valeur mensuelle de C (l'augmentation est de 50% pour l'estimation de la PAC) et une variance environ 4 fois plus élevée (7 fois dans le cas de la PAC) que celle de l'estimateur composite stationnaire pour l'estimation de la variation d'un mois à l'autre. Compte tenu de ces chiffres, nous nous sommes limités à la comparaison de E1 et E2.

Afin d'évaluer les estimateurs E1 et E2 pour les six premiers mois de 1994, nous avons calculé leurs variances respectives ainsi que l'écart entre leur valeur et la valeur espérée de l'estimateur composite stationnaire. Les

tableaux 2 et 3 présentent les variances et les écarts quadratiques moyens (EQM = variance + carré de l'écart), chacune de ces valeurs étant divisée par la variance de l'estimateur composite stationnaire: deux colonnes pour la valeur mensuelle et deux autres pour la variation d'un mois à l'autre. Pour faciliter la comparaison, les valeurs pour E1 (continuer d'utiliser l'approche composite) sont placées au-dessus de celles pour E2 (estimateur par quotient en janvier 94) dans chaque cellule.

Pour tous les calculs, nous nous sommes servis des indices de biais courants pour les mois précédant 1994. Quant aux mois de 1994, nous avons utilisé les trois types d'indices (courants, QTD, ITAO) pour examiner chaque cas. Pour les tableaux 2 et 3, nous nous sommes servis uniquement des indices QTD.

Pour représenter l'effet de la nouvelle méthodologie de la CPS, nous avons supposé une hausse de la valeur des caractéristiques de l'activité pour janvier 1994. Dans les cas étudiés, nous supposons que la valeur espérée de l'estimateur par quotient subit une hausse de 10% pour la population active en chômage (tableau 2) et de 1% pour la population active civile (tableau 3). Ces «hausses», établies d'après des données provisoires, influent sur le biais et la variance. Nous n'avons pas supposé d'autres variations de valeur pour les autres mois de 1994.

Les tableaux 2 et 3 montrent que l'estimateur E1 se compare favorablement à l'estimateur composite stationnaire (ECS) en ce qui concerne l'estimation de la valeur mensuelle; l'estimateur E2 est un peu moins efficace par rapport à l'ECS (6% dans le tableau 2, 20% dans le tableau 3). En ce qui concerne la PAC (tableau 3), les deux estimateurs étudiés sont beaucoup moins brillants par rapport à l'ECS pour l'estimation de la variation d'un mois à l'autre (hausses de 23 et de 30% respectivement). Cela s'explique par le fait qu'on suppose une baisse de 10% des coefficients de corrélation courants entre décembre et janvier, alors que ces coefficients sont ordinairement élevés (80% pour deux groupes à un mois d'intervalle). Si nous utilisons d'autres types d'indices de biais ou d'autres pourcentages de variation pour notre analyse, la comparaison des variances des deux estimateurs s'en trouve peu modifiée.

Tableau 2: Variances et écarts quadratiques moyens (par comparaison à l'estimateur composite stationnaire) pour *C* -- indices de biais QTD, variation de *10%*; valeurs pour E1 au-dessus de celles pour E2.

Mois	Valeur mensuelle		Variation d'un mois à l'autre	
	Variance	Écart quadratique moyen	Variance	Écart quadratique moyen
Janvier 1994	1.0154 1.0673	2.0700 2.8724	1.0283 1.0087	22.30 24.99
Février 1994	1.0044 1.0246	1.1731 1.3134	1.0022 1.0028	1.3097 1.5292
Mars 1994	1.0011 1.0087	1.0281 1.0549	1.0008 1.0014	1.0500 1.0856
Avril 1994	1.0002 1.0025	1.0045 1.0099	1.0003 1.0012	1.0082 1.0147
Mai 1994	1.0000 1.0004	1.0007 1.0016	1.0001 1.0007	1.0013 1.0029
Juin 1994	1.0000 1.0000	1.0001 1.0002	1.0000 1.0001	1.0002 1.0005

Tableau 3: Variances et écarts quadratiques moyens (par comparaison à l'estimateur composite stationnaire) pour *PAC* -- indices de biais QTD, variation de *1%*; valeurs pour E1 au-dessus de celles pour E2.

Mois	Valeur mensuelle		Variation d'un mois à l'autre	
	Variance	Écart quadratique moyen	Variance	Écart quadratique moyen
Janvier 1994	1.0289 1.1966	1.5145 2.4065	1.2259 1.2962	35.06 40.85
Février 1994	1.0080 1.0722	1.0857 1.2657	1.0073 1.0133	1.2477 1.6122
Mars 1994	1.0020 1.0234	1.0144 1.0543	1.0026 1.0114	1.0411 1.1072
Avril 1994	1.0003 1.0063	1.0023 1.0113	1.0010 1.0062	1.0071 1.0215
Mai 1994	1.0000 1.0010	1.0004 1.0018	1.0002 1.0032	1.0011 1.0057
Juin 1994	1.0000 1.0001	1.0001 1.0002	1.0000 1.0006	1.0002 1.0010

Les écarts par rapport à l'estimateur composite stationnaire sont très élevés en janvier. Cela s'explique, dans le cas de la valeur mensuelle, par l'utilisation d'indices de biais QTD dans les exemples des tableaux 2 et 3. De plus, dans le cas de la variation d'un mois à l'autre, s'ajoute l'effet de l'hypothèse de la variation de la valeur espérée de l'estimateur par quotient -- 10% dans le tableau 2, 1% dans le tableau 3. (La variation d'un mois à l'autre est supposée nulle dans le modèle stationnaire.) Les écarts sont beaucoup plus faibles, dans le cas de la valeur mensuelle, si on utilise des indices de biais ITAO en 1994 (pour C) et sont presque nuls si on conserve les indices courants. Pour ce qui est de la variation d'un mois à l'autre, les écarts pour janvier 94 ne sont que légèrement plus faibles si on utilise les indices ITAO ou les indices courants à cause de la forte influence de la variation de valeur.

Les calculs que nous avons effectués avec d'autres ensembles de paramètres révèlent plusieurs tendances dignes d'intérêt. Même si les variances et les écarts quadratiques moyens (EQM) calculés pour E2 sont généralement plus élevés que les variances et les EQM calculés pour E1, l'écart est souvent mince. En outre, peu importe la valeur de la variance ou du EQM en janvier, les valeurs calculées pour février se rapprochent déjà beaucoup plus de celles pour l'estimateur composite stationnaire. Dès mars ou avril, l'écart n'est plus que de 1 ou 2% en règle générale. Si nous pensons qu'il peut se produire un accroissement plus considérable des autres erreurs non dues à l'échantillonnage dans la période de transition, les estimateurs E1 et E2 auront somme toute une influence mineure dans l'ensemble du projet.

Compte tenu de ces observations, le choix de l'un ou l'autre de ces deux estimateurs reposera probablement plus sur des considérations pratiques ou des considérations relatives au traitement. Si tel est le cas, nous penchons actuellement pour l'estimateur E2: on débute avec un estimateur par quotient simple en janvier et on reprend un estimateur composite en février. Par surcroît, il y aurait possibilité de modifier les coefficients A et K de manière à ramener plus rapidement les variances ou les EQM au niveau de ceux de l'estimateur composite stationnaire.

3. DEVRAIT-ON CHANGER LE PLAN DE RENOUVELLEMENT?

3.1 Interview des ménages sur 6 ou 8 mois consécutifs

Le plan de renouvellement actuel de la CPS (4-8-4), dont nous avons parlé dans la sous-section 1.2, a pour effet de réduire la variance des estimations de la variation d'un mois à l'autre et d'une année à l'autre, cela grâce au fait qu'environ 75% des ménages ayant participé à l'enquête dans un mois donné sont interviewés de nouveau le mois suivant et 50% de ce même groupe sont réinterviewés un an plus tard. Quelle réduction de variance pourrait-on obtenir si on interviewait les ménages sur six ou huit mois consécutifs par exemple? De passage au Census Bureau récemment, Wayne Fuller discutait des résultats que lui et ses collègues, Adam et Yansaneh, avaient obtenus en examinant ces divers plans de renouvellement. Bien que le taux de chevauchement d'un mois à l'autre passerait à 83 ou à 87%, selon qu'on choisit le plan à six ou à huit mois consécutifs, il n'y aurait pas de chevauchement entre les mois à un an d'intervalle. Les comparaisons de Fuller supposent l'utilisation d'un estimateur linéaire à variance minimum.

Dans notre analyse, nous avons comparé les variances d'estimateurs composites AK en faisant varier plusieurs paramètres. Trois plans de renouvellement ont été analysés: le 4-8-4, le plan à 6 mois consécutifs et celui à huit mois consécutifs. Pour chacun d'eux, nous avons calculé la variance d'estimateurs de la valeur mensuelle, de la variation d'un mois à l'autre et de la moyenne annuelle. Pour les caractéristiques étudiées -- le nombre de personnes en chômage (C) et l'effectif de la population active civile (PAC) -- nous nous sommes servis des mêmes ensembles de coefficients de corrélation que dans la section 2. Enfin, bien que nous ne nous soyons pas intéressés aux estimateurs linéaires généraux dans notre analyse, nous avons calculé des variances d'estimateurs composites AK en faisant prendre à A et à K les valeurs 0, 0.1, 0.2, ..., 0.9.

3.2 Résultats pour les trois plans de renouvellement

Nous avons calculé le changement de variance qu'entraîne pour des estimateurs composites AK l'utilisation d'un plan de renouvellement à six ou à huit mois consécutifs, *en comparaison du plan actuel (4-8-4) avec A = 0.2 et K = 0.4*. Les plans à six ou à huit mois consécutifs ont des effets mineurs sur la variance de l'estimation de la valeur mensuelle. En ce qui a trait à la caractéristique C, on constate une hausse de la variance de l'ordre de 0 à 10% pour la plupart des paires de valeurs A,K; quant à la PAC, les plans à six ou à huit mois consécutifs peuvent amener une réduction de la variance pouvant atteindre 11% avec des valeurs de A et de K voisines de 0.7 ou 0.8.

Un effet plus sensible des plans à six ou à huit mois consécutifs est la réduction observée dans la variance de l'estimation de la variation d'un mois à l'autre lorsque les corrélations sont plus fortes. Dans l'estimation de la caractéristique C, les plans à 6 ou à 8 mois entraînent une réduction de la variance de l'ordre de 3 à 7% pour la plupart des paires de valeurs A,K. À cause de la corrélation plus forte qui existe entre les groupes de renouvellement lorsqu'on estime la PAC, le taux de réduction de la variance grimpe à 23% avec le plan à 6 mois (26% avec le plan à 8 mois) si K = 0.8 (mais A faible, 0.1 ou 0.2).

Il convient de mentionner ici que le fait de comparer des résultats optimaux pour les plans à 6 et à 8 mois est injuste par rapport au plan 4-8-4. Après tout, les valeurs en usage actuellement (A = 0.2 et K = 0.4) sont un peu le résultat d'un compromis; elles sont satisfaisantes pour C, passables pour PAC. (Voir l'analyse à la sous-section 4.1.) Si nous estimions PAC en conservant le plan 4-8-4, nous pourrions réduire les variances de l'estimation de la valeur mensuelle et de l'estimation de la variation d'un mois à l'autre dans des proportions pouvant atteindre 12 et 17% respectivement en choisissant d'autres paires de valeurs A,K.

Malheureusement, lorsqu'on opte uniquement pour un plan à mois consécutifs, on accroît sensiblement la variance de l'estimation de la moyenne annuelle. Même avec la paire optimale de valeurs A,K, la variance s'accroît de 14% pour C (de 18% pour PAC) selon le plan à 6 mois et de 28% (de 36%) selon le plan à 8 mois.

Même si les estimations de la valeur mensuelle et de la variation d'un mois à l'autre sont toujours jugées plus importantes que les estimations de la moyenne annuelle, celles-ci ont de l'importance pour les 40 plus petits États (y compris le District de Columbia). Tandis que les 11 plus grands États ont droit à des estimations mensuelles de la population active, les 40 autres doivent se contenter d'estimations annuelles à cause de la faible

taille de leur échantillon mensuel. Selon les corrélations qui peuvent exister, la taille effective de l'échantillon est moins grande pour les plans à mois consécutifs que pour les plans 4-8-4 lorsqu'il s'agit de calculer une moyenne pour 12 mois, si l'on tient compte du fait que les plans à mois consécutifs supposent un taux de chevauchement des ménages plus élevé que les plans 4-8-4. Comme le démontrent les calculs, cette situation amène une hausse de la variance.

Comme les coefficients de variation des estimations de la moyenne annuelle pour les petits États sont plus élevés que ceux des estimations mensuelles nationales, les autorités du Bureau of Labor Statistics et du Bureau of the Census ont jugé que l'administration d'interviews sur six ou huit mois consécutifs aurait des conséquences néfastes pour les estimations des États moins importants. On a donc décidé de conserver le plan de renouvellement actuel, soit le 4-8-4, pour les années 1990.

4. QUESTIONS DE FOND SUR L'ESTIMATION COMPOSITE

4.1 Choix des coefficients de l'estimateur composite

Avant de modifier la formule de l'estimation composite, on étudiera des données de 1994 pour déterminer la structure du biais qui découlera de la nouvelle méthodologie et évaluer les autres conséquences possibles. Les paires de valeurs A,K optimales, ou les coefficients généraux optimaux, varient selon l'ensemble de biais de renouvellement et de coefficients de corrélation. Nous espérons pouvoir apporter des modifications à la formule vers 1996.

Les coefficients utilisés actuellement dans l'estimateur de la CPS -- K = 0.4 et A = 0.2 -- sont quelque peu le résultat d'un compromis. Bien qu'ils soient presque des coefficients optimaux pour le calcul de C, on obtiendrait, pour des caractéristiques à corrélation plus forte comme la PAC et la population active occupée, une plus forte réduction de la variance avec des valeurs de K voisines de 0.7 ou 0.8 et des valeurs de A aussi plus élevées. Cependant, la population active en chômage est souvent considérée comme la caractéristique la plus importante de toutes. D'autres facteurs influent aussi sur le choix des valeurs de K et de A. Tandis que des valeurs de K élevées ont généralement un effet réducteur sur la variance dans l'estimation de la variation d'un mois à l'autre, elles ont l'effet contraire dans l'estimation de la moyenne annuelle, qui est un paramètre important pour les petits États (comme nous l'avons mentionné dans la sous-section 3.2).

L'estimateur composite AK est lui-même le résultat d'un compromis sur le plan du stockage de données. L'estimateur par quotient simple utilise uniquement des données du mois courant. À l'inverse, l'estimateur linéaire à variance minimum nécessite l'enregistrement de données de groupes de renouvellement de nombreux mois antérieurs. Bien que l'estimateur AK utilise des données de nombreux mois antérieurs, ces données sont résumées dans l'estimateur composite du mois précédent. Il suffit donc de sauvegarder les estimations de groupes de renouvellement de ce mois et du mois précédent, sans oublier l'estimateur composite AK du dernier mois, pour produire le nouvel estimateur.

L'utilisation d'un estimateur linéaire général est une solution attrayante en ceci que, sous réserve de l'enregistrement obligatoire de données antérieures, nous pouvons chercher à obtenir un estimateur à variance minimum. En outre, nous pouvons calculer les coefficients optimaux pour l'estimation de la variation par rapport au dernier mois en faisant la soustraction des coefficients optimaux pour les estimations du mois courant et du mois précédent, utilisant le plus grand nombre de mois possible pour chaque estimation. Toutefois, comme d'autres l'ont souligné, il faudrait réviser les estimations de la population active du mois précédent. Le Bureau of Labor Statistics voit d'un mauvais oeil cette pratique parce qu'elle est difficile à expliquer aux utilisateurs de données.

Il est possible d'améliorer la variance de l'estimateur AK actuel sans accroître pour autant les exigences de stockage de données. Breau et Ernst (1983) examinent des coefficients *composites généralisés* dans un estimateur de la forme $Y_h = \Sigma\, a_i x_{h,i} - K \Sigma\, b_i x_{h,i} + K Y_{h-1}$. Dans leur résumé, ils notent que l'utilisation de cet estimateur peut amener une réduction appréciable de la variance dans le calcul de la moyenne annuelle et une réduction plus modeste dans le calcul de la valeur mensuelle et de la variation d'un mois à l'autre. De plus, cette réduction est presque équivalente à celle obtenue avec un estimateur linéaire à variance minimum.

En contrepartie, des études antérieures montrent que, comme on passe de l'estimateur composite AK à l'estimateur composite généralisé à l'estimateur linéaire à variance minimum, la réduction de la variance s'accompagne souvent d'une hausse du biais de conditionnement dans les estimations. Cet inconvénient explique le peu d'empressement que l'on met à proposer le remplacement de l'estimateur AK par l'un des deux autres estimateurs. Une fois que nous aurons évalué, en 1994, les effets de la nouvelle méthodologie sur l'évolution du biais, nous aurons peut-être de meilleures raisons de modifier la forme de l'estimateur. Entre-temps, comme nous cherchons particulièrement à limiter la variance des estimations de la moyenne annuelle, nous poursuivons notre étude de l'estimateur composite généralisé et de l'estimateur linéaire à variance minimum.

4.2 Des coefficients différents pour chaque caractéristique

Une autre solution, examinée par Wayne Fuller au Census Bureau, pour améliorer les estimations de la population active consiste à utiliser des coefficients différents pour chaque caractéristique clé. Cette solution avait été écartée dans le passé à cause des problèmes de cohérence de données. La règle à l'heure actuelle est de produire des estimations de totaux qui concordent avec les composantes de ces totaux et qui soient cohérentes d'une période à l'autre. Par exemple, la somme de l'effectif de la population active occupée et de l'effectif de la population active en chômage doit correspondre à l'effectif de la population active civile; la valeur estimée d'une moyenne annuelle doit être égale à la moyenne des valeurs estimées de ses composantes.

Fuller a proposé une nouvelle méthode par laquelle différentes paires de valeurs A,K ou différents coefficients permettent d'établir des estimations composites de totaux pour les principales caractéristiques de l'activité. Ces valeurs estimées constituent alors un nouvel ensemble de totaux de contrôle, comme pour les classes «âge-origine raciale-sexe», qui servent au redressement par pondération. Une fois que tous les redressements ont été effectués, la somme des poids des individus doit égaler l'estimation composite du total de population active et équivaloir à l'agrégat démographique pertinent.

La méthode pourrait aller comme suit. Accomplir toutes les étapes du redressement (non-réponse, totaux de contrôle démographiques, etc.). Déterminer l'estimateur composite de C au moyen de la paire optimale de valeurs A,K pour cette caractéristique. Déterminer l'estimateur composite de EO (effectifs occupés) au moyen de la paire optimale de valeurs A,K pour cette caractéristique. Additionner les estimations obtenues dans chaque cas pour calculer la valeur estimée de PAC. Soustraire cette valeur des totaux de contrôle démographiques pour connaître le nombre de personnes de 16 ans et plus qui sont inactives. Ces totaux de population active servent désormais de totaux de contrôle tout en demeurant compatibles avec les estimations de population active. À ce stade-ci, il convient d'effectuer à nouveau un redressement en fonction des totaux de contrôle démographiques, car le recours à l'approche composite peut avoir modifié quelque peu les chiffres.

Cette méthode aurait plusieurs aspects intéressants. Les poids inclus dans les fichiers de données renfermeraient les effets de l'estimation composite. Les utilisateurs de données de la CPS pourraient ainsi reproduire des estimations finales avant l'opération de désaisonnalisation, ce qui ne se fait pas actuellement. En outre, on n'aurait pas besoin de stocker autant de données.

Plusieurs questions se posent. Premièrement, pour quel couple de caractéristiques de l'activité devrait-on utiliser l'estimation composite (C et EO?), et quel couple de caractéristiques devrait-on estimer par déduction? La réponse à ces questions devrait-elle dépendre uniquement de l'importance des caractéristiques, ou devrait-on aussi prendre en considération des aspects statistiques?

Deuxièmement, à quel niveau devrait-on utiliser les totaux de contrôle qui proviennent des estimations de l'activité: totaux nationaux, fréquences marginales des classes démographiques (âge-origine raciale-sexe), ou cellules démographiques proprement dites (classement recoupé)? Si nous nous mettons à utiliser l'approche composite pour les cellules à faible effectif, nous simplifions le procédé de balayage mais en même temps, nous accroissons la variabilité des poids et le risque d'obtenir des fréquences par cellule négatives.

Finalement, on choisirait les coefficients de l'estimateur composite dans le but d'améliorer les estimations nationales de l'activité. Quel effet aurait cette action sur 1) les estimations relatives aux sous-groupes *qui ne sont pas touchés par l'approche composite* (par ex., le nombre de Noirs occupés, si nous n'appliquons pas l'approche composite pour l'origine raciale), 2) les estimations de l'activité au niveau de l'État et 3) les variables autres que

la population active? Ces questions, ainsi que d'autres, doivent être approfondies avant que nous puissions recommander des modifications majeures pour le système d'estimation composite de la CPS.

REMERCIEMENTS

Les auteurs remercient Wayne Fuller de les avoir encouragés dans leur recherche et de leur avoir communiqué ses précieux commentaires.

BIBLIOGRAPHIE

Adam, A., et Fuller, W. (1992). Covariances of estimators for the current population survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, sous presse.

Adams, D. (1991). Memorandum for documentation, CPS Month-In-Sample (MIS) bias index research, 10/21/91. U.S. Bureau of the Census, Washington, DC.

Bailar, B. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

Breau, P., et Ernst, L. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 397-402.

Gurney, M., et Daly, J. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 242-257.

Kumar, S., et Lee, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 403-408.

Shoemaker, H. (1992). Memorandum for documentation. CATI phase-in analysis: A look at month-in-sample bias indexes for unemployed (CC_ALYS-8), 10/30/92, U.S. Bureau of the Census, Washington, DC.

U.S. Bureau of the Census (1978). The current population survey: Design and methodology. Papier technique n° 40, Washington, DC: U.S. Government Printing Office (Department of Commerce).

Wolter, K. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Yansaneh, I., et Fuller, W. (1992). Alternative estimators for the current population survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, sous presse.

* Cette communication expose les résultats généraux de recherches qui ont été faites par le personnel du Census Bureau. Les opinions qui y sont exprimées sont celles des auteurs et elles ne reflètent pas nécessairement la position du Census Bureau.

SESSION 5

Études longitudinales dans des recherches sur la santé

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

ANALYSE DE DONNÉES LONGITUDINALES DICHOTOMIQUES

G.A. Darlington[1]

RÉSUMÉ

Lorsqu'on observe des réponses dichotomiques dans le temps, on doit tenir compte de la dépendance entre les observations relatives à une personne. Deux méthodes de modélisation s'appliquant à des données longitudinales dichotomiques sont examinées. Chacune de ces méthodes convient lorsqu'on cherche principalement à déterminer la relation entre la probabilité marginale d'un événement et un ensemble de variables explicatives et lorsqu'on veut aussi étudier la dépendance entre la structure de corrélation et les variables explicatives. Les méthodes de modélisation sont illustrées par des exemples.

MOTS CLÉS: Données longitudinales dichotomiques, variables explicatives, infection par le VIH.

1. INTRODUCTION

La régression logistique est couramment utilisée pour modéliser la relation entre des réponses dichotomiques indépendantes et un ensemble de covariables (Cox 1970). Lorsque les réponses dichotomiques correspondent à des observations relatives à des personnes dans le temps, la dépendance entre les observations relatives à la même personne doit être prise en considération. Par conséquent, il faut modifier la méthode de régression logistique pour tenir compte des réponses corrélées.

D'importants travaux initiaux sur l'analyse de données longitudinales dichotomiques, utilisant des fonctions de vraisemblance de travail et des estimations robustes de la variance, ont été réalisés par Zeger, Liang et Self (1985). Dans la présente communication, nous utiliserons la méthode de Zeger et coll. (1985) pour examiner des modèles destinés à des situations où l'on s'intéresse principalement à la probabilité marginale de succès, mais où l'on veut aussi étudier la dépendance entre la structure de corrélation et les variables explicatives. L'introduction de paramètres visant à tenir compte de la dépendance entre les observations relatives à une personne sera examinée aux sections 2 et 3. Des exemples numériques tirés d'études sur le SIDA sont présentés à la section 4.

2. MODÉLISATION DE L'AUTOCORRÉLATION COMME FONCTION DE COVARIABLES

Soit Y_{it} une réponse dichotomique observée pour la personne i au temps t, $i = 1, ..., K$, $t = 1, ..., n_i$. Soit z_i le vecteur s x 1 des variables explicatives indépendantes du temps pour la personne i, et soit $\pi_i = Pr(Y_{it} = 1 | z_i)$. Zeger et coll. (1985) font d'abord l'hypothèse que logit $(\pi_i) = \beta' z_i$, où β est un vecteur s x 1 de paramètres. Ils supposent aussi que corr $(Y_{it}, Y_{it-1} | z_i) = \rho$. Autrement dit, ils supposent que l'autocorrélation avec décalage de un est constante.

Pour de nombreuses études longitudinales, on s'attend à ce que ρ se situe dans l'intervalle $0 < \rho < 1$. Cette restriction permettra une représentation logistique simple de l'autocorrélation avec décalage de un incluant une dépendance à l'égard de covariables. Supposons que

[1] G.A. Darlington, division d'épidémiologie et de statistique, Fondation ontarienne pour la recherche en cancérologie et le traitement du cancer, 620 University Avenue, Toronto (Ontario), Canada, M5G 2L7.

$$\text{logit}(\rho_i) = \tau' z_i,$$

où $\rho_i = \text{corr}(Y_{it}, Y_{it-1}|z_i)$ et τ est un vecteur s x 1 de paramètres. Si l'on fait l'hypothèse d'une dépendance de Markov, la fonction de vraisemblance s'écrit

$$L(\beta,\tau) = \prod_{i=1}^{K} \left[\pi_i^{y_{i1}} (1-\pi_i)^{1-y_{i1}} \prod_{t=2}^{n_i} \pi_{it}^{y_{it}} (1-\pi_{it})^{1-y_{it}} \right], \tag{2.1}$$

où

$$\pi_i = Pr(Y_{it}=1|z_i) = \frac{e^{\beta' z_i}}{1+e^{\beta' z_i}},$$

$$\pi_{it} = Pr(Y_{it} = 1|Y_{it-1}, z_i) = E(Y_{it}|Y_{it-1}, z_i)$$

$$= \pi_i + \rho_i (Y_{it-1} - \pi_i)$$

et

$$\rho_i = \frac{e^{\tau' z_i}}{1+e^{\tau' z_i}}.$$

Notons que la fonction de vraisemblance, telle qu'elle est formulée en (2.1), ne dépend pas du modèle choisi. Notons en outre, d'après (2.1), que l'intervalle non restreint pour ρ_i est

$$\max\left[-\frac{\pi_i}{1-\pi_i}, -\frac{1-\pi_i}{\pi_i} \right] < \rho_i < 1.$$

Rappelons que la fonction de vraisemblance (2.1) est fondée sur une hypothèse de dépendance de Markov. Compte tenu de la possibilité que la dépendance soit plus générale, nous appellerons cette fonction de vraisemblance, comme l'ont fait Zeger et coll. (1985), la fonction de vraisemblance de travail, puisqu'elle suppose qu'on travaille avec une hypothèse de dépendance de Markov. Le logarithme de la fonction de vraisemblance de travail avec hypothèse de Markov est donc donné par

$$l(\beta,\tau) = \sum_{i=1}^{K} l_i$$

$$= \sum_{i=1}^{K} [y_{i1}\, \beta' z_i - \log\{1 + \exp(\beta' z_i)\} \tag{2.2}$$

$$+ \sum_{t=2}^{n_i} \{y_{it} \log \pi_{it} + (1-y_{it}) \log (1-\pi_{it})\}].$$

Les estimations des paramètres $\hat{\beta}$ et $\hat{\tau}$ peuvent être obtenues en maximisant ce logarithme de fonction de vraisemblance. Un processus itératif est nécessaire pour calculer ces estimations.

Tester la constance de l'autocorrélation avec décalage de un équivaut à tester si toutes les composantes de τ, sauf le terme constant, sont égales à zéro. Comme dans Zeger et coll. (1985), les estimations $\hat{\beta}$ et $\hat{\tau}$ sont, sous réserve de certaines conditions de régularité, convergentes pour β et τ et asymptotiquement normales en vertu de conditions plus générales, puisque les

$$\sum_{i=1}^{K} \frac{\partial l_i}{\partial \theta} = 0,$$

sont des équations d'estimations non biaisées, où $\theta = (\beta', \tau')'$.

Ainsi, sous réserve de certaines conditions de régularité (Inagaki 1973), dans le cas où Y_{it} est une série chronologique dichotomique stationnaire telle que

$$\text{logit}[Pr(Y_{it} = 1 | z_i)] = \beta' z_i$$

et

$$\text{logit}[\text{corr}(Y_{it}, Y_{it-1} | z_i)] = \tau' z_i,$$

$i = 1, ..., K$, $t = 1, ..., n_i < \infty$, les résultats d'Inagaki (1973) indiquent que si $\theta = (\beta', \tau')'$ et $\hat{\theta} = (\hat{\beta}', \hat{\tau}')'$, alors $\sqrt{K}(\hat{\theta} - \theta)$ est asymptotiquement normal avec moyenne 0 et matrice de covariance $W^{-1} V W^{-1}$ où

$$K^{-1} \sum_{i=1}^{K} E_T \left(\frac{\partial^2 l_i}{\partial \theta \partial \theta'} | z_i \right) \to W, \tag{2.3}$$

$$K^{-1} \sum_{i=1}^{K} E_T \left[\left(\frac{\partial l_i}{\partial \theta} \right) \left(\frac{\partial l_i}{\partial \theta} \right)' | z_i \right] \to V, \tag{2.4}$$

lorsque $K \to \infty$ et $E_T(\cdot)$ représente l'espérance par rapport à la vraie distribution sous-jacente.

Des estimations robustes des erreurs-types des estimations des paramètres peuvent être obtenues de l'estimation de la matrice de covariance $K^{-1} \hat{W}^{-1} \hat{V} \hat{W}^{-1}$ où

$$\hat{W}_{uv} = \frac{1}{K} \sum_{i=1}^{K} \left(\frac{\partial^2 l_i}{\partial \theta_u \partial \theta_v} \right) |_{\hat{\theta}} \tag{2.5}$$

et

$$\hat{V}_{uv} = \frac{1}{K} \sum_{i=1}^{K} \left\{ \left(\frac{\partial l_i}{\partial \theta_u} \right) \left(\frac{\partial l_i}{\partial \theta_v} \right) \right\} |_{\hat{\theta}}. \tag{2.6}$$

Notons que si l'hypothèse de dépendance de Markov est valide, $W^{-1} V W^{-1}$ devient $-W^{-1}$, et donc des estimations fondées sur le modèle des erreurs-types des estimations des paramètres peuvent être obtenues de $-K^{-1} \hat{W}^{-1}$.

3. MODÉLISATION D'UNE PROBABILITÉ DE TRANSITION

Dans le modèle de Zeger et coll. (1985), la distribution marginale de Y_{it} est donnée par logit $(\pi_i) = \beta' z_i$. Farewell (1982) utilise également cette spécification marginale, mais plutôt que de supposer une forme pour la corrélation, il suppose que

$$\text{logit}\ \{Pr(Y_{it} = 1 \mid Y_{it-1} = 1, z_i)\} = \gamma' z_i,$$

où γ est un vecteur s x 1 de paramètres. Notons la présence, dans la structure de dépendance, d'une asymétrie qui ne convient peut-être pas à toutes les applications.

Pour ce modèle, la fonction de vraisemblance de travail avec hypothèse de Markov est donnée par (2.1), où, encore une fois, $\pi_{it} = \pi_i + \rho_i (Y_{it-1} - \pi_i)$, mais

$$\rho_i = \text{corr}(Y_{it}, Y_{it-1} | z_i) = \frac{e^{\gamma' z_i} - e^{\beta' z_i}}{1 + e^{\gamma' z_i}}.$$

Puisque la fonction de vraisemblance doit être maximisée aux valeurs $0 < \hat{\pi}_{it} < 1$ et $0 < \hat{\pi}_i < 1$, ρ_i est confiné à l'intervalle

$$\max \left(-\frac{\pi_i}{1 - \pi_i}, -\frac{1 - \pi_i}{\pi_i} \right) < \rho_i < 1.$$

Le logarithme de la fonction de vraisemblance de travail avec hypothèse de Markov est donné par (2.2). On peut obtenir les estimations des paramètres en maximisant cette fonction par rapport à β et γ. En vertu de l'hypothèse de dépendance de Markov, les résultats habituels de l'estimation du maximum de vraisemblance s'appliquent. Si l'hypothèse de dépendance de Markov n'est pas valide, les estimations $\hat{\beta}$, $\hat{\gamma}$ obtenues par la maximisation de (2.2) peuvent conserver une certaine valeur puisque les équations d'estimation sont non biaisées. Ainsi, sous réserve de certaines conditions de régularité présentées par Inagaki (1973), si $\theta = (\beta', \gamma')'$ et si les n_i sont bornés, $i = 1, ..., K,$, alors $\hat{\theta}$ est convergent et $\sqrt{K}(\hat{\theta} - \theta)$ est asymptotiquement normal avec moyenne 0 et matrice de covariance $W^{-1}VW^{-1}$, où W et V sont donnés par (2.3) et (2.4) respectivement. L'estimation robuste de la matrice de covariance est donnée par $K^{-1}\hat{W}^{-1}\hat{V}\hat{W}^{-1}$, où $\hat{W}$ et $\hat{V}$ sont présentés en (2.5) et en (2.6) respectivement.

4. EXEMPLES NUMÉRIQUES

4.1 Étude sur le SIDA de New York

Les résultats d'une étude ayant porté sur les facteurs de risque de l'infection par le VIH parmi les usagers de drogues intraveineuses à New York sont présentés par Des Jarlais et coll. (1987) et par Marmor et coll. (1987). Nous examinerons un sous-ensemble des données ayant produit ces résultats afin de déterminer si l'usage de drogues intraveineuses diffère entre les personnes séropositives et les personnes séronégatives.

Deux cent vingt-cinq hommes de New York faisant usage de drogues intraveineuses ont été interviewés tous les ans pendant une période de quatre ans. La fréquence mensuelle de consommation de drogues intraveineuses a été déterminée. Toutes les personnes de ce sous-ensemble étaient soit séropositives, soit séronégatives pendant toute la durée de la période de quatre ans.

Les méthodes décrites aux sections 2 et 3 ont été utilisées pour examiner les changements de comportement liés au fait d'être séropositif ou séronégatif. Ainsi, la réponse est définie comme ceci: $Y_{it} = 1$ si la personne i, au temps t, fait un usage fréquent de drogues intraveineuses, où $i = 1, ..., 225$, $t = 1, ..., n_i, n_i \leq 4$, et où «usage fréquent» signifie plus de cinq injections par mois; dans les autres cas, $Y_{it} = 0$. La covariable, indépendante du temps, est un indicateur précisant si la personne i est séropositive $(z_i = 1)$ ou séronégative $(z_i = 0)$.

Les résultats de l'ajustement du modèle décrit à la section 2 sont présentés au tableau 1. En comparant les estimations des paramètres aux estimations robustes ou fondées sur le modèle de l'erreur-type, on observe que l'autocorrélation avec décalage de un est constante et que la probabilité de faire un usage fréquent de drogues intraveineuses est plus grande pour les personnes séropositives.

Tableau 1: Résultats des estimations pour le modèle qui suppose que

$$\text{logit}\,[Pr(Y_{it} = 1 \mid z_i)] = \beta_0 + \beta_1 z_i \text{ et}$$
$$\text{logit}\,[\text{corr}\,(Y_{it}, Y_{it-1} \mid z_i)] = \tau_0 + \tau_1 z_i.$$

Paramètre	Estimation	E.-t.-$_{FM}$[1] estimée	E.-t.$_R$[2] estimée
β_0	-0.120	0.178	0.238
β_1	0.879	0.279	0.411
τ_0	-0.494	0.487	0.653
τ_1	-0.567	0.818	1.247

[1] D'après l'estimation fondée sur le modèle de la matrice de covariance.

[2] D'après l'estimation robuste de la matrice de covariance.

Tableau 2: Résultats des estimations pour le modèle simplifié qui suppose que

$$\text{logit}\,[Pr(Y_{it} = 1 \mid z_i)] = \beta_0 + \beta_1 z_i \text{ et}$$
$$\text{logit}\,[\text{corr}\,(Y_{it}, Y_{it-1} \mid z_i)] = \tau.$$

Paramètre	Estimation	E.-t.-$_{FM}$[1] estimée	E.-t.$_R$[2] estimée
β_0	-0.151	0.166	0.206
β_1	0.927	0.244	0.257
τ	-0.775	0.409	0.620

[1] D'après l'estimation fondée sur le modèle de la matrice de covariance.

[2] D'après l'estimation robuste de la matrice de covariance.

Puisque l'autocorrélation avec décalage de un est constante, le modèle de Zeger et coll. (1985) est mis en application. Les résultats de l'ajustement de ce modèle sont présentés au tableau 2 et révèlent les mêmes conclusions globales que ceux de la première analyse.

Le tableau 3 donne les résultats de l'ajustement du modèle décrit à la section 3 en fonction des données. On peut conclure de l'observation de ce tableau que la probabilité d'un usage fréquent de drogues, s'il y avait déjà un usage fréquent, ne diffère pas selon que les personnes sont séropositives ou séronégatives. Toutefois, on peut aussi constater que la probabilité globale d'un usage fréquent est plus élevée dans le cas des personnes séropositives.

Tableau 3: Résultats des estimations pour le modèle qui suppose que

$$\text{logit}\ [Pr(Y_{it}=1\,|\,z_i)] = \beta_0 + \beta_1 z_i \text{ et}$$
$$\text{logit}\ [Pr(Y_{it}=1\,|\,Y_{it-1}=1, z_i)] = \gamma_0 + \gamma_1 z_i.$$

Paramètre	Estimation	E.-t.$_{PM}$[1] estimée	E.-t.$_{R}$[2] estimée
β_0	-0.120	0.166	0.169
β_1	0.879	0.247	0.255
γ_0	0.712	0.312	0.326
γ_1	0.458	0.414	0.441

[1] D'après l'estimation fondée sur le modèle de la matrice de covariance.

[2] D'après l'estimation robuste de la matrice de covariance.

Tableau 4: Résultats des estimations pour le modèle qui suppose que

$$\text{logit}\ [Pr(Y_{it}=0\,|\,z_i)] = \beta_0 + \beta_1 z_i \text{ et}$$
$$\text{logit}\ [Pr(Y_{it}=0\,|\,Y_{it-1}=0, z_i)] = \gamma_0 + \gamma_1 z_i.$$

Paramètre	Estimation	E.-t.$_{PM}$[1] estimée	E.-t.$_{R}$[2] estimée
β_0	0.120	0.166	0.169
β_1	-0.879	0.247	0.255
γ_0	0.887	0.237	0.234
γ_1	-0.910	0.377	0.403

[1] D'après l'estimation fondée sur le modèle de la matrice de covariance.

[2] D'après l'estimation robuste de la matrice de covariance.

Ce modèle a également été appliqué à la réponse définie comme un usage non fréquent de drogues, au lieu d'un usage fréquent. Les résultats de l'ajustement de ce modèle en fonction des données sur l'usage non fréquent de drogues sont présentés au tableau 4. Les conclusions, dans ce cas, sont que la probabilité conditionnelle d'un usage non fréquent de drogues, s'il y avait déjà un usage non fréquent, varie selon que la personne est séropositive ou séronégative; en effet, cette probabilité conditionnelle est plus faible pour les personnes séropositives que pour les personnes séronégatives. Par ailleurs, la probabilité globale d'un usage non fréquent de drogues est moins élevée pour les personnes séropositives que pour les personnes séronégatives.

Il semble qu'il soit tout à fait possible, par conséquent, d'introduire une dépendance vis-à-vis de covariables dans la structure de corrélation. L'asymétrie du second modèle est reflétée dans les conclusions différentes qui ressortent des tableaux 3 et 4. Chaque modèle particulier sera sensible aux relations particulières existant entre les covariables et la structure de corrélation générale. Par exemple, seul le modèle du tableau 4 illustre de façon évidente une dépendance vis-à-vis d'une covariable, même si les modèles des tableaux 1, 3 et 4 donnent tous des ajustements identiques. Il importe donc de tenir compte de la forme du modèle adopté et de son lien avec les aspects précis étudiés.

4.2 Étude sur le SIDA de Toronto

Les résultats d'une étude sur l'infection par le VIH parmi une cohorte de contacts sexuels masculins de personnes atteintes du SIDA ou du para-SIDA sont présentés par Calzavara et coll. (1991) et par Calzavara et coll. (1993). Grâce à un sous-ensemble de données tirées de cette étude, nous avons examiné si la fréquence d'un comportement sexuel à haut risque était associée à la consommation de drogues à usage récréatif. Le sous-ensemble de données contenait de l'information obtenue de 176 membres de la cohorte. L'information relative au comportement sexuel à haut risque a été obtenue tous les trois mois pendant des périodes allant jusqu'à cinq

ans. Une cote a été élaborée pour mesurer l'exposition à un comportement sexuel à haut risque (Calzavara et coll. 1993), en tenant compte aussi bien du type que de la fréquence des contacts sexuels. Dans le sous-ensemble de données, une personne était classée comme ayant une activité sexuelle à haut risque (à bas risque) si sa cote était supérieure (inférieure) à la médiane. Des données sur la consommation de drogues à usage récréatif au moment de la visite initiale ont également été incluses.

Les méthodes décrites aux sections 2 et 3 ont été utilisées pour examiner la relation entre le comportement sexuel et la consommation de drogues à usage récréatif. La réponse est donc ainsi définie: $Y_{it} = 1$ si la cote de risque du comportement sexuel pour la personne i au temps t se situait au-dessus de la médiane, $i = 1, ..., 176$, $t = 1, ..., n_i, n_i \leq 20$ (maximum de 4 observations par année pendant 5 ans); sinon, $Y_{it} = 0$. La covariable est un indicateur précisant si la personne i consommait des drogues à usage récréatif ($z_i = 1$) ou non ($z_i = 0$).

Les résultats de l'ajustement du modèle d'autocorrélation de la section 2 sont présentés au tableau 5. Si l'on compare l'estimation de τ_1 à l'estimation robuste de l'erreur-type, on peut conclure que l'autocorrélation avec décalage de un ne dépend pas de la consommation de drogues à usage récréatif. Si l'estimation de l'erreur-type fondée sur le modèle est utilisée, on observe des indications d'une telle dépendance, ce qui met clairement en évidence la possibilité de conclusions erronées résultant de l'utilisation des estimations de l'erreur-type fondées sur le modèle. On peut aussi conclure que la probabilité marginale d'un comportement sexuel à haut risque est liée à la consommation de drogues à usage récréatif, c'est-à-dire que la probabilité d'un comportement sexuel à haut risque est plus élevée pour les personnes ayant des antécédents de consommation de drogues à usage récréatif, comparativement aux personnes n'ayant pas de tels antécédents. Puisque le résultat de l'estimation robuste ne permet pas de conclure à la présence d'une corrélation non constante, le modèle de Zeger et coll. (1985) peut être appliqué (résultats non présentés).

Si la méthode de probabilité conditionnelle de la section 3 est employée (tableau 6), les résultats suivants sont obtenus. Comme prévu, les conclusions relatives à la probabilité marginale d'un comportement sexuel à haut risque sont les mêmes que celles présentées plus haut au sujet du modèle d'autocorrélation. En ce qui a trait à la probabilité conditionnelle, on peut conclure que la probabilité d'un comportement sexuel à haut risque, si la personne avait déjà un tel comportement sexuel, est plus grande pour les personnes ayant des antécédents de consommation de drogues à usage récréatif que pour les personnes n'ayant pas de tels antécédents.

Tableau 5: Résultats des estimations pour le modèle qui suppose que

$$\text{logit}\,[Pr(Y_{it} = 1 \mid z_i)] = \beta_0 + \beta_1 z_i \text{ et}$$
$$\text{logit}\,[\text{corr}(Y_{it}, Y_{it-1} \mid z_i)] = \tau_0 + \tau_1 z_i.$$

Paramètre	Estimation	E.-t.$_{RM}$[1] estimée	E.-t.$_R$[2] estimée
β_0	-1.069	0.188	0.211
β_1	0.534	0.195	0.273
τ_0	1.259	0.207	0.278
τ_1	0.240	0.104	0.151

[1] D'après l'estimation fondée sur le modèle de la matrice de covariance.

[2] D'après l'estimation robuste de la matrice de covariance.

Tableau 6: Résultats des estimations pour le modèle qui suppose que

$$\text{logit}\,[Pr(Y_{it} = 1 \mid z_i)] = \beta_0 + \beta_1 z_i \text{ et}$$
$$\text{logit}\,[Pr(Y_{it} = 1 \mid Y_{it-1} = 1, z_i)] = \gamma_0 + \gamma_1 z_i.$$

Paramètre	Estimation	E.-t.$_{RM}$[1] estimée	E.-t.$_R$[2] estimée
β_0	-0.887	0.165	0.233
β_1	1.444	0.198	0.275
γ_0	0.617	0.182	0.260
γ_1	1.117	0.210	0.298

[1] D'après l'estimation fondée sur le modèle de la matrice de covariance.

[2] D'après l'estimation robuste de la matrice de covariance.

Il est à remarquer que le sous-ensemble de données comprenait aussi un indicateur précisant, pour chaque personne, si elle souffrait ou non d'une infection par le VIH. Les modèles initiaux incluaient cette variable, mais les résultats de leur ajustement ne sont pas inclus, puisque cet indicateur VIH n'apportait pas de contribution significative à aucun des modèles, et que l'inclusion ou l'exclusion de cette variable n'avait pas d'effet sur les estimations des répercussions de la consommation de drogues à usage récréatif.

REMERCIEMENTS

Les auteurs sont reconnaissants envers le Dr Vernon T. Farewell et le Dr Janet M. Raboud. Cette recherche a été appuyée par le Conseil de recherches en sciences naturelles et en génie du Canada, ainsi que par une subvention versée au Societal Institute of the Mathematical Sciences (SIMS) par le National Institute on Drug Abuse (subvention NIDA DA-04722). Les données tirées de l'étude de New York à des fins d'illustration ont été recueillies par les Dr Don Des Jarlais et Michael Marmor, ainsi que des collègues, sous le parrainage du National Institute of Drug Abuse (subvention NIDA DA-03574). Les données tirées de l'étude de Toronto à des fins d'illustration ont été recueillies par les Dr Randall A. Coates et Stanley E. Read, ainsi que des collègues, sous le parrainage du ministère de la Santé de l'Ontario (subvention n° 01184) et du Programme national de recherche et de développement en matière de santé de Santé et Bien-être social Canada (subvention n° 6606-2587-54).

BIBLIOGRAPHIE

Calzavara, L.M., Coates, R.A., Johnson, K., Read, S.E., Farewell, V.T., Fanning, M.M., Shepherd, F.A., et MacFadden, D.K. (1991). Sexual behaviour changes in a cohort of male sexual contacts of men with HIV disease: A three-year overview. *Canadian Journal of Public Health*, 82, 150-156.

Calzavara, L.M., Coates, R.A., Raboud, J.M., Farewell, V.T., Read, S.E., Shepherd, F.A., Fanning, M.M., et MacFadden, D. (1993). Ongoing high risk sexual behaviours in relation to recreational drug use in sexual encounters: Analysis of five years of data from the Toronto sexual contact study. *Annals of Epidemiology* (accepté pour publication).

Cox, D.R. (1970). *Analysis of Binary Data*. London: Methuen.

Des Jarlais, D.C., Friedman, S.R., Marmor, M., Cohen, H., Mildvan, D., Yancovitz, S., Mathur, U., El-Sadr, W., Spira, T.J., Garber, J., Beatrice, S.T., Abdul-Quader, A.S., et Sotheran, J.L. (1987). Development of AIDS, HIV seroconversion, and potential co-factors for T4 cell loss in a cohort of intravenous drug users. *AIDS*, 1, 105-111.

Farewell, V.T. (1982). Alternatives to the proportional hazards model. Dans *Environmental Epidemiology: Risk Assessment*, R.L. Prentice et A.J. Whittemore, Éds., 216-229.

Inagaki, N. (1973). Asymptotic relations between the likelihood estimating function and the maximum likelihood estimator. *Annals. Institute Statistical Mathematics*, 25, 1-26.

Marmor, M., Des Jarlais, D.C., Cohen, H., Friedman, S.R., Beatrice, S.T., Dubin, N., El-Sadr, W., Mildvan, D., Yancovitz, S., Mathur, U., et Holzman, R. (1987). Risk factors for infection with Human Immunodeficiency virus among intravenous drug abusers in New York City. *AIDS*, 1, 39-44.

Zeger, S.L., Liang, K.Y., et Self, S.G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, 72, 31-38.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

ANALYSE STATISTIQUE DE SÉRIES CHRONOLOGIQUES PARALLÈLES: EFFETS DE LA POLLUTION ATMOSPHÉRIQUE SUR LES ADMISSIONS DANS LES HÔPITAUX

R. Burnett, S. Bartlett, D. Krewski, G. Roberts et M. Raad-Young[1]

RÉSUMÉ

Les effets nocifs possibles de la pollution de l'air ambiant sur la santé sont examinés par la relation entre les admissions quotidiennes dans des hôpitaux de soins actifs en Ontario pour des troubles respiratoires et les niveaux d'ozone quotidiens. L'ensemble de données est constitué de 197 séries chronologiques de dénombrement parallèle, correspondant aux 197 hôpitaux de soins actifs visés par l'analyse. Les estimations des niveaux d'ozone au voisinage de chaque hôpital proviennent des mesures des stations de surveillance atmosphérique exploitées par le ministère de l'Environnement de l'Ontario. Des méthodes fondées sur des équations d'estimation sont utilisées pour examiner les effets de l'ozone sur les admissions pour troubles respiratoires, ainsi que la nature stochastique des réponses. Les données sur les admissions montrent peu d'indications de la présence d'une corrélation sériale. Toutefois, les taux d'admission varient énormément entre les hôpitaux. Cette dernière source de variation doit être prise en considération dans l'examen des effets de la pollution atmosphérique.

Mots clés: Équations d'estimation généralisées; surdispersion; pollution atmosphérique; troubles respiratoires; admissions dans les hôpitaux.

1. INTRODUCTION

On a souvent recours à des études épidémiologiques pour examiner les effets nocifs possibles de la pollution de l'air ambiant (Office of Technology Assessment 1984). En raison des niveaux relativement faibles de pollution atmosphérique observés de nos jours dans la plupart des régions de l'Amérique du Nord, tout effet néfaste sur la santé est vraisemblablement subtil, de sorte que des échantillons de taille importante sont nécessaires pour qu'un tel effet puisse être détecté. Bon nombre des protocoles utilisés pour étudier les effets possibles de la pollution atmosphérique sur la santé se fondent sur une forme quelconque d'échantillonnage en grappes et emploient des données longitudinales sur la santé et la qualité de l'air.

Les dossiers administratifs de la santé, tels que les relevés quotidiens des décès ou les dossiers des hôpitaux sur la morbidité, ont été utilisés comme indicateurs possibles des effets de la pollution de l'air ambiant sur la santé humaine. Bates et Sizto (1989) ont examiné le lien entre les admissions quotidiennes dans 79 hôpitaux de soins actifs dans le sud-ouest de l'Ontario et les niveaux quotidiens de pollution atmosphérique. Dans cette étude, le nombre total d'admissions quotidiennes pour des troubles respiratoires dans 79 hôpitaux a été utilisé comme mesure de la réponse globale et mis en relation avec plusieurs polluants atmosphériques, notamment l'ozone, l'anhydryde sulfureux, le dioxyde d'azote et les sulfates. En raison de la nature longitudinale de ces données, une certaine forme de corrélation sériale pourrait être présente. Vu les tailles variables des hôpitaux et la nature de leur rôle dans la prestation de soins de santé, une variation appréciable des taux d'admission entre les hôpitaux est également prévisible. Il en résulte une corrélation positive entre les observations venant du même hôpital. Les études de ce genre comportent généralement un grand nombre (plusieurs centaines ou plus) d'observations par hôpital.

[1] R. Burnett, S. Bartlett et D. Krewski, Direction d'hygiène du milieu, Santé et Bien-être social Canada, Ottawa, (Ontario), Canada, K1A 0L2. G. Roberts, Division des méthodes d'enquêtes-entreprises, et M. Raad-Young, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, (Ontario), Canada, K1A 0T6.

L'élaboration de méthodes statistiques permettant d'analyser de telles données a fait l'objet de nombreux travaux ces dernières années. Stiratelli et coll. (1984) ont utilisé une approche bayesienne empirique exigeant beaucoup de calculs pour analyser des données binaires longitudinales provenant d'une enquête par panel relative aux effets de la pollution par les matières en suspension sur les taux de crises d'asthme. D'autres méthodes, fondées sur des équations d'estimation généralisées, sont un peu moins exigeantes sur le plan des calculs (Liang et Zeger 1986; Prentice et Zhao 1991; et Liang et coll. 1992). Toutefois, ces méthodes d'analyse ont été employées principalement dans des plans de recherche ne comportant que quelques observations par grappe. Or, ce qui nous intéresse ici, ce sont des grappes (hôpitaux) comportant des centaines ou des milliers d'observations. Les méthodes existantes d'inférence statistique ne conviennent pas à des plans comprenant des grappes d'aussi grande taille.

Dans la présente communication, quelques adaptations simplifiées au niveau des calculs de l'approche des équations d'estimation de Liang et Zeger (1986) pour des modèles de régression avec données corrélées sont explorées dans le cadre de plans de recherche impliquant des grappes de grande taille. Nos méthodes sont élaborées de telle façon qu'elles puissent être appliquées à l'aide des logiciels statistiques d'usage courant. Les méthodes proposées sont illustrées en utilisant les données sur les admissions quotidiennes pour troubles respiratoires dans 197 hôpitaux de l'Ontario en relation avec les niveaux quotidiens d'ozone ambiant, au cours de la période 1983-1988.

2. MODÈLES DE RÉGRESSION POUR DES DONNÉES DE DÉNOMBREMENT LONGITUDINALES

Soit y_{kt} le nombre d'admissions en urgence pour des troubles respiratoires le t -ième jour dans le k-ième hôpital $(t = 1,...,T; k = 1,...,K)$. Considérons le modèle suivant pour des données de dénombrement longitudinales

$$E(y_{kt} \mid \varepsilon_k, \eta_t) = \lambda_{kt}\, \varepsilon_k \eta_t,$$

où λ_{kt} est l'espérance inconditionnelle de y_{kt}, qui est fonction d'un vecteur $(p \times 1)$ de covariables x_{kt}, par exemple des niveaux de pollution atmosphérique ou des valeurs climatiques, avec le vecteur de régression inconnu β. Les ε_k sont des variables aléatoires scalaires indépendantes avec espérance égale à l'unité et variance commune τ, qui représentent l'effet aléatoire attribuable à l'hôpital, et les η_t sont des variables aléatoires ayant une espérance égale à l'unité et une matrice de variance-covariance $\phi\Omega$, qui engendrent une corrélation sériale dans les réponses. Ici, $\Omega = ((\omega_{st}))$ est une matrice de corrélation $(T \times T)$ dont les éléments sont déterminés par un vecteur $(s \times 1)$ de paramètres inconnus, dénoté ρ. Par exemple, une structure de corrélation AR[1] est obtenue en établissant

$$\omega_{t,t+l} = \rho^l,$$

$(l = 0,..., T-1; |\rho| < 1)$.

Nous supposons en outre que la variance conditionnelle est donnée par

$$\mathrm{Var}(y_{kt} \mid \varepsilon_k \eta_t) = \theta\, \lambda_{kt}\, \varepsilon_k\, \eta_t,$$

$(\theta > 0)$, avec $\theta = 1, \theta > 1$, et $\theta < 1$ représentant les variations de Poisson, extra-Poisson et intra-Poisson respectivement. Pour des valeurs de ε_k et η_t données, la covariance conditionnelle entre les admissions quotidiennes dans le même hôpital est supposée égale à zéro.

Si l'on fait la moyenne sur les η_t, on obtient

$$E(y_{kt} \mid \varepsilon_k) = \lambda_{kt}\, \varepsilon_k,$$

$$\mathrm{Var}(y_{kt} \mid \varepsilon_k) = \theta\, \lambda_{kt}\, \varepsilon_k + \phi\, (\lambda_{kt}\varepsilon_k)^2 \equiv v_{kt},$$

$$\mathrm{Cov}(y_{kt}, y_{ks} \mid \varepsilon_k) = \phi\, \omega_{|t-s|}\, \lambda_{kt} \lambda_{ks} \varepsilon_k^2$$

et

$$\mathrm{Corr}(y_{kt}, y_{ks} \mid \varepsilon_k) = \phi\, \omega_{|t-s|} \lambda_{kt} \lambda_{ks} \varepsilon_k^2\, (v_{kt} v_{ks})^{-1/2}.$$

Notons que la corrélation sériale est une fonction de la période qui s'écoule entre les réponses et de l'espérance conditionnelle $E(y_{kt} \mid \varepsilon_k) = \lambda_{kt} \varepsilon_k$. La dépendance à l'égard de cette espérance conditionnelle est minimale si le nombre d'admissions quotidiennes est élevé. Dans ce cas, $\theta \lambda_{kt} \varepsilon_k << \phi (\lambda_{kt} \varepsilon_k)^2$, et la corrélation sériale peut être obtenue approximativement par

$$\mathrm{Corr}(y_{kt}, y_{ks} \mid \varepsilon_k) \approx \omega_{|t-s|}.$$

Si ϕ est petit, on a $\theta \lambda_{kt} \varepsilon_k >> \phi (\lambda_{kt} \varepsilon_k)^2$, et la corrélation est donnée approximativement par

$$\mathrm{Corr}(y_{kt}, y_{ks} \mid \varepsilon_k) \approx (\phi / \theta)\, \omega_{|t-s|}\, (\lambda_{kt} \lambda_{ks})^{1/2}\, \varepsilon_k.$$

Dans le cas où ϕ est petit, la corrélation peut aussi être faible même pour des valeurs élevées de $\omega_{|t-s|}$.

Si on fait la moyenne pour les ε_k, on obtient

$$E(y_{kt}) = \lambda_{kt}.$$

La matrice de covariance à l'intérieur d'un hôpital $\mathrm{Cov}(Y_k) \equiv V_k$, $Y_k = (y_{k1}, \ldots, y_{kT})'$, est définie par

$$V_k(\tau, \phi, \rho) = \theta \Lambda_k + \Lambda_k (\tau J_k + \phi (\tau + 1) \Omega) \Lambda_k, \tag{1}$$

où $\Lambda_k = \mathrm{diag}(\lambda_{k1}, \ldots, \lambda_{kT})$ et J_k est une matrice $(T \times T)$ dont tous les éléments sont des 1.

Zeger (1988) a examiné une seule série chronologique de données de dénombrement avec covariance donnée par (1), en posant $\tau = 0$. Thall et Vail (1990) ont également utilisé la fonction de covariance (1) dans leur analyse de données de dénombrement longitudinales sans corrélation sériale (c.-à-d. $\phi = 0$). Burnett et coll. (1992a) ont examiné des structures d'erreur semblables à celles proposées ici, en posant comme hypothèse additionnelle que les effets aléatoires ε_k et η_t suivent une loi normale logarithmique.

Deux autres sources de variation dans les séries relatives aux admissions sont évidentes. Les taux d'admission varient selon le jour de la semaine; les plus hauts taux sont observés les lundis, ils baissent ensuite tout au long de la semaine pour atteindre leur plus bas niveau la fin de semaine. Les variations saisonnières des admissions sont également manifestes; les taux les plus élevés sont observés pendant les mois d'hiver, puis survient une baisse de mars à août, suivie d'une augmentation au cours de l'automne. La composante déterministe λ_{kt} du modèle peut tenir compte de ces variations dans le temps ainsi que de l'effet de la pollution atmosphérique et du climat sur les taux d'admission si l'on pose

$$\lambda_{kt} = D_t\, S_t\, f(x_{kt}; \beta).$$

Ici, $(D_t; t = 1, \ldots, T)$ est une série chronologique formée de sept valeurs uniques représentant le rapport du nombre moyen d'admissions en urgence à chacun des sept jours de la semaine sur le taux d'admission quotidien moyen. Le facteur S_t défini par

$$S_t = \sum_{l=-9}^{9} \Psi_l\, \bar{y}_{t-l},$$

est un filtre linéaire symétrique sur 19 jours conçu pour supprimer les tendances temporelles et la corrélation sériale à progression lente. Les poids $(\Psi_0, ..., \Psi_9)$ sont donnés par (0.087, 0.086, 0.081, 0.073, 0.063, 0.052, 0.040, 0.030, 0.020, 0.012), et $\bar{y}_t$ est le nombre moyen d'admissions le t -ième jour dans l'ensemble des hôpitaux. La fonction f relie les variables environnementales x_{kt} à la série des admissions quotidiennes y_{kt} après un ajustement tenant compte des tendances selon les saisons et les jours de la semaine.

Le but de la présente analyse est de nous permettre de faire des inférences quant au vecteur inconnu de paramètres de régression β et d'obtenir des estimations du vecteur du paramètre de corrélation $\alpha = (\tau, \phi, \rho, \theta)'$ qui définit la structure de variance-covariance des données.

3. ESTIMATION

Puisque la distribution des données n'a pas été spécifiée, les méthodes de vraisemblance d'inférence statistique ne peuvent être utilisées. Liang et Zeger (1986) proposent d'utiliser des équations d'estimation pour faire des inférences sur les paramètres de régression si la distribution conjointe des T observations de la k-ième grappe n'est pas spécifiée.

Selon la méthode de Liang et Zeger (1986), le vecteur de paramètres de régression β serait estimé de la façon suivante. Pour un estimateur convergent $\hat{\alpha}$ de α, la nouvelle estimation $\hat{\beta}$ de β est donnée par

$$\hat{\beta} = \beta + \left(\sum_{k=1}^{K} X_k' V_k^{-1} X_k \right)^{-1} \sum_{k=1}^{K} X_k' V_k^{-1} (Y_k - \lambda_k), \tag{2}$$

où X_k est la matrice $(T \times p)$ des dérivées de λ_{kt} par rapport à β et $\lambda_k = (\lambda_{k1}, ..., \lambda_{kT})'$. L'expression (2) est évaluée selon les estimations courantes des paramètres β et α.

La mise en oeuvre de ce processus d'estimation itératif exige la conception d'un logiciel informatique particulier; l'inversion de V_k, par ailleurs, peut poser des problèmes de calcul si les tailles des grappes sont élevées.

La procédure NLIN du SAS (SAS 1988) peut permettre d'obtenir des estimations des paramètres pour des modèles de régression non linéaires, par la méthode des moindres carrés repondérés itérativement. L'estimation $\tilde{\beta}$ de β est équivalente à l'estimation $\hat{\beta}$ à partir de (2) si V_k est remplacée par la matrice de covariance $\tilde{V}_k = \text{diag}(v_{k1}, ..., v_{kT})$, le poids NLIN pour y_{kt} étant donné par v_{kt}^{-1}. Dans la présente communication, nous utilisons les poids λ_{kt}^{-1}.

Pour illustrer nos méthodes, considérons une structure d'erreur AR[1] pour les η_t avec le paramètre scalaire ρ. Si l'on a $\tilde{\beta}$ comme estimateur convergent de β, des estimateurs convergents de ϕ, ρ et τ sont donnés par

$$\hat{\rho} = \frac{\hat{C}_3 - \hat{C}_2}{\hat{C}_2 - \hat{C}_1},$$

$$\hat{\tau} = \hat{C}_1 - \frac{\hat{C}_1 - \hat{C}_2}{1 - \hat{\rho}} \tag{3}$$

et

$$\hat{\phi} = \frac{\hat{C}_1 - \hat{C}_2}{\hat{\rho}(1 - \hat{\rho})(\hat{\tau} + 1)}, \tag{4}$$

où

$$\hat{C}_l = \frac{\sum_{k=1}^{K} \sum_{t=l+1}^{T} \hat{r}_{kt} \hat{r}_{k,t-l}}{\sum_{k=1}^{K} \sum_{t=l+1}^{T} \hat{\lambda}_{kt} \hat{\lambda}_{k,t-l}},$$

($l = 1,2,3$), avec $\hat{r}_{kt} = (y_{kt} - \hat{\lambda}_{kt})$ et $\hat{\lambda}_{kt} = \exp(x_{kt}'\tilde{\beta})$. Un estimateur convergent $\hat{\theta}$ de θ est donné par

$$\hat{\theta} = \frac{\sum_{k=1}^{K} \sum_{t=l+1}^{T} \hat{r}_{kt}^2 - \hat{\lambda}_{kt}^2 (\hat{\tau} + \hat{\phi}(\hat{\tau}+1))}{\sum_{k=1}^{K} \sum_{t=l+1}^{T} \hat{\lambda}_{kt}}.$$

S'il n'y a pas de corrélation sériale ($\rho = 0$), τ peut être estimé par

$$\hat{\tau} = \frac{\sum_{k=1}^{K} \left(\hat{S}_k^2 - \sum_{t=1}^{T} \hat{r}_{kt}^2\right)}{\sum_{k=1}^{K} \left(\hat{G}_k^2 - \sum_{t=1}^{T} \hat{\lambda}_{kt}^2\right)}, \tag{5}$$

où $\hat{S}_k = \sum_{t=1}^{T} \hat{r}_{kt}$ et $\hat{G}_k = \sum_{t=1}^{T} \hat{\lambda}_{kt}$. (Notons que l'estimateur de $\hat{\tau}$ en (5) est basé sur tous les produits croisés des résidus à l'intérieur de chaque grappe, tandis que l'estimateur donné en (3) se fonde seulement sur les premiers et deuxièmes éléments non diagonaux de la matrice des produits croisés des résidus.)

Si $\rho = 0$, une estimation de ϕ ne peut être obtenue à l'aide de (4). Dans ce cas, on peut obtenir les estimations aussi bien de ϕ que de θ en notant que si l'on a $\tau = \hat{\tau}$ et $\lambda_{kt} = \hat{\lambda}_{kt}$,

$$\mathrm{Var}(y_{kt}) = \theta \hat{\lambda}_{kt} + (\hat{\tau} + \phi(\hat{\tau}+1)) \hat{\lambda}_{kt}^2. \tag{6}$$

On estime les paramètres θ et ϕ par régression linéaire simple qui utilise la partie à droite de l'égalité en (6) pour prédire les observations $(y_{kt} - \hat{\lambda}_{kt})^2$.

La matrice de covariance de $\tilde{\beta}$ est donnée par

$$\mathrm{Cov}(\tilde{\beta}) = \left[\sum_{k=1}^{K} X_k' \tilde{V}_k^{-1} X_k\right]^{-1} \left[\sum_{k=1}^{K} X_k' \tilde{V}_k^{-1} V_k \tilde{V}_k^{-1} X_k\right] \left[\sum_{k=1}^{K} X_k' \tilde{V}_k^{-1} X_k\right]^{-1} \tag{7}$$

(Liang et Zeger 1986). Une estimation de la covariance de $\tilde{\beta}$ est obtenue par l'évaluation de $\mathrm{Cov}(\tilde{\beta})$ à $(\tilde{\beta}, \tilde{\alpha})$.

La multiplication de matrices de très grande taille comme $X_k' \tilde{V}_k^{-1} V_k \tilde{V}_k^{-1} X_k$ peut poser des difficultés au langage matriciel de la procédure PROC IML du SAS, en raison des limites de mémoire vive des micro-ordinateurs. Toutefois, les multiplications de matrices en (7) peuvent être écrites sous forme de sommes de carrés. De cette façon, la multiplication de matrices requise peut être exécutée par la procédure PROC CORR du SAS, qui calcule la matrice des sommes de carrés pour plusieurs variables.

4. EFFETS DE LA POLLUTION ATMOSPHÉRIQUE SUR LES ADMISSIONS DANS LES HÔPITAUX

Bates et Sizto (1989) ont examiné la relation entre les hospitalisations quotidiennes pour des problèmes respiratoires dans 79 hôpitaux de soins actifs du sud-ouest de l'Ontario et la pollution de l'air ambiant au cours des mois de janvier, février, juillet et août, pour les années 1976 à 1983. Une étude subséquente (Burnett et coll. 1992b) se fonde sur le nombre d'admissions quotidiennes en urgence pour des troubles respiratoires dans des hôpitaux de soins actifs de l'Ontario au cours de la période du 1er janvier 1983 au 31 décembre 1988, représentant

un total de 400 000 jours d'hôpital. Nous avons pu obtenir des estimations raisonnables de l'exposition à l'ozone au voisinage de 197 hôpitaux. Puisque les observations à l'intérieur des hôpitaux sont ordonnées chronologiquement, la possibilité d'une corrélation sériale existe.

Examinons, à des fins d'illustration, la relation entre les taux des admissions en urgence pour des troubles respiratoires (codes 466, 480-486, 490-496, 786 de la *Classification internationale des maladies*) au cours des mois de mai à août et le niveau maximum quotidien d'ozone moyen sur une heure x_{kt}, en utilisant le modèle

$$f(x_{kt};\beta) = \exp(\beta_0 + \beta_1 x_{kt}). \tag{8}$$

Puisque $\hat{\phi} \approx 0$, il y a peu d'indications de la présence d'une corrélation sériale. Les estimations de θ, de τ, de β_1 et de l'erreur-type de β_1 pour les modèles de *la moyenne de la population* et les modèles *propres à l'hôpital* sont présentées au tableau 1 utilisant les données de l'Ontario globalement ou de Toronto seulement. (Les modèles de *la moyenne de la population* décrivent les variations des taux d'admission de l'ensemble de l'Ontario associées aux variations des niveaux d'ozone, tandis que les modèles *propres à l'hôpital* décrivent ces variations dans chaque hôpital.) Puisque $\hat{\theta} \approx 1$, il y a peu d'indications d'une surdispersion ou d'une sous-dispersion par rapport à la variation de Poisson. Cela est attribuable au fait que les admissions dans les hôpitaux pour des troubles respiratoires sont des événements rares, avec seulement 154 admissions quotidiennes au cours de l'été en Ontario, comparativement à environ neuf millions de personnes susceptibles d'être admises à l'hôpital chaque jour.

Tableau 1: Estimations des paramètres et erreurs-types pour les modèles de régression de la moyenne de la population et propres à l'hôpital.

Paramètre (unités)	Ontario (197 hôpitaux)		Toronto (24 hôpitaux)	
	Moyenne de la population	Propre à l'hôpital	Moyenne de la population	Propre à l'hôpital
θ (adms^2)	1.04	1.04	1.08	1.08
τ (adms^2)	0.75	≈ 0	0.31	≈ 0
$\beta_1 \times 10^{-4}$ (adms/ppb)	24.4	1.78	1.75	1.73
$e.\text{-}t.(\hat{\beta}_1)_{Ind} \times 10^{-4}$ (adms/ppb)	1.74	1.42	3.52	2.97
$e.\text{-}t.(\hat{\beta}_1)_{Dép} \times 10^{-4}$ (adms/ppb)	10.4	1.42	2.97	2.97

Selon le modèle de *la moyenne de la population* pour l'Ontario, la variance entre les hôpitaux est $\hat{\tau} = 0.75$, et la corrélation intra-hôpital est $\hat{\tau}\hat{\lambda}(1 + \hat{\tau}\hat{\lambda})^{-1} = 0.36$, évaluée au taux moyen de $\hat{\lambda} = 0.78$ admissions par jour. L'estimation $\hat{\beta} = 24.4 \times 10^{-4}$ du coefficient de régression pour l'ozone signifie qu'il y a 24.8 admissions quotidiennes de moins dans l'ensemble des 197 hôpitaux lorsque le niveau d'ozone tombe à 0 ppb, par rapport à sa valeur moyenne de 51.6 ppb. Ce résultat veut dire que l'ozone est associé à 24.8/154 = 16% de toutes les admissions en urgence pour des troubles respiratoires au cours de l'été en Ontario.

Si les observations sont supposées indépendantes ($\tau = 0$), l'erreur-type de $\hat{\beta}_1$ est 1.74×10^{-4} (voir $e.\text{-}t.(\hat{\beta}_1)_{Ind}$ au tableau 1), ce qui constitue une forte indication de la présence d'un effet positif de l'ozone. Toutefois, si τ est supposé positif, l'erreur-type grimpe à 10.4×10^{-4} ($e.\text{-}t.(\hat{\beta}_1)_{Dép}$ au tableau 1). Cette multiplication par 6 de l'erreur est due à un effet de l'ozone lié aux écarts des taux d'admission entre les hôpitaux. Des niveaux d'ozone plus élevés sont observés dans la partie sud-ouest de la province, qui est plus densément peuplée que les régions

du nord et de l'est. Ces différences de densité démographique engendrent des écarts des taux d'admission qui affichent aussi une corrélation positive avec les niveaux d'ozone.

Le but de la présente analyse est d'examiner les effets de courts épisodes de pollution atmosphérique sur les taux d'admission. Afin d'étudier adéquatement les fluctuations quotidiennes des admissions en relation avec la pollution atmosphérique pour chaque hôpital, les écarts de taux entre les hôpitaux devraient être supprimés. On peut y arriver en examinant l'espérance conditionnelle

$$E(y_{kt} \mid \epsilon_k) = D_t S_t \exp(\beta_0 + \beta_1 x_{kt})\ \epsilon_k.$$

Selon ce modèle, un effet courant de la pollution atmosphérique, β_1,, fait l'objet d'une régression par rapport au ratio entre les admissions quotidiennes dans un hôpital donné, y_{kt}, et l'effet correspondant attribuable à l'hôpital, y_{kt}. Une estimation de ϵ_k est donnée par le rapport du nombre moyen d'admissions quotidiennes pour le k-ième hôpital sur le taux d'admission quotidien moyen pour l'ensemble des données.

Bien que l'estimation de θ selon ce modèle *propre à l'hôpital* soit la même que celle selon le modèle de *la moyenne de la population*, on a $\hat{\tau} \approx 0$, ce qui indique que toute la variation des taux entre les hôpitaux a été supprimée. Cette correction supprime tout effet transversal de la pollution atmosphérique sur les admissions, et ne laisse que l'effet longitudinal. L'estimation $\hat{\beta}_1 = 1.78 \times 10^{-4}$ selon le modèle *propre à l'hôpital* est de beaucoup inférieure à celle obtenue selon le modèle de *la moyenne de la population*, ce qui indique qu'une grande partie de la relation entre la pollution atmosphérique et les admissions pour troubles respiratoires provient de la variation des taux entre les hôpitaux.

Selon le modèle *propre à l'hôpital*, $\hat{\tau} \approx 0$ et, par conséquent, τ se voit attribuer la valeur zéro. En vertu de cette structure d'erreur, les deux formulations de l'erreur-type (*Ind* et *Dép*) sont équivalentes, puisque les données sont indépendantes. Cette approche est avantageuse par rapport au modèle de *la moyenne de la population* du fait que des logiciels informatiques courants peuvent être utilisés tant pour l'estimation que pour l'inférence. Les estimations des erreurs-types fournies par la procédure PROC NLIN, par exemple, sont convergentes en vertu de la définition du modèle *propre à l'hôpital*.

Il est bien connu qu'une analyse faisant appel au groupement offrirait un meilleur pouvoir de détection des effets qui varient seulement à l'intérieur d'un groupe et équilibrés d'un groupe à l'autre, comparativement à une analyse supposant que les données sont indépendantes. En guise d'illustration, considérons les données des 24 hôpitaux de soins actifs de Toronto. Dans ce cas, on suppose que les patients admis dans tous les hôpitaux de Toronto ont subi le même niveau d'exposition à l'ozone troposphérique. Ainsi, cette covariable est commune aux hôpitaux et varie dans le temps à l'intérieur de chaque hôpital selon les modèles de *la moyenne de la population* et *propres à l'hôpital* pour Toronto. Les estimations de β_1 sont semblables et, fait intéressant, comparables aux résultats pour l'ensemble de la province selon le modèle *propre à l'hôpital*. L'erreur-type de $\hat{\beta}_1$, en supposant des réponses indépendantes, est 3.52×10^{-4}. On obtient une réduction de 15% de l'erreur ($e.-t.(\hat{\beta}_1) = 2.97 \times 10^{-4}$) en reconnaissant la dépendance des observations à l'intérieur d'un hôpital. Il ne s'agit que d'un gain modeste de pouvoir de détection, car la variation des taux entre les hôpitaux de Toronto ($\hat{\tau} = 0.31$) est inférieure à celle observée entre l'ensemble des hôpitaux de l'Ontario ($\hat{\tau} = 0.75$). Cela est attribuable à la gamme limitée des tailles des hôpitaux de Toronto. L'erreur-type de $\hat{\beta}_1$ selon le modèle *propre à l'hôpital* est semblable à celle obtenue selon le modèle de *la moyenne de la population*.

5. ANALYSE

Des méthodes d'estimation et d'inférence, simples du point de vue des calculs, sont présentées pour des données de dénombrement longitudinales comprenant des grappes de grande taille. La forme de la structure de covariance tient compte de la variation dans les admissions quotidiennes entre les hôpitaux et de l'autocorrélation des données dans le temps. Le biais de l'estimation des effets de la pollution atmosphérique sur les admissions pour des troubles respiratoires, dans les hôpitaux attribuable à la relation positive entre la densité démographique et les niveaux de pollution atmosphérique, peut être supprimé si l'on considère les modèles de régression *propres à l'hôpital*. Cette définition permet l'utilisation de logiciels statistiques courants pour l'estimation et l'inférence des paramètres de régression.

La méthode d'estimation des paramètres de surdispersion présentée à la section 3 est choisie en fonction des exigences de calcul. Bien que l'estimation de β et de α avec des équations distinctes donne des estimations très efficaces de β, une perte d'efficacité considérable est prévisible pour α. Liang et coll. (1992) montrent que pour des données binaires corrélées, le fait d'estimer β et α par des équations distinctes peut engendrer une perte d'efficacité aussi élevée que 50% pour les estimations de α, comparativement à l'estimation conjointe de α et β. Des équations d'estimation conjointes pourraient être utilisées pour éviter cette perte d'efficacité, bien que leur mise en oeuvre soit peu pratique dans le cas des grappes de grande taille.

REMERCIEMENTS

Les auteurs remercient le ministère de la Santé de l'Ontario, qui a fourni les données sur les admissions dans les hôpitaux, et le ministère de l'Environnement de l'Ontario, qui a fourni les données sur la pollution atmosphérique.

BIBLIOGRAPHIE

Bates, D., et Sizto, R. (1989). The Ontario air pollution study: Identification of the causative agent. *Environmental Health Perspectives*, 79, 69-72.

Burnett, R.T., Shedden, J., et Krewski, D. (1992a). Nonlinear regression models for correlated count data. *Environmetrics*, 3, 211-222.

Burnett, R.T., Dales, R.E., Raizenne, M.E., Krewski, D., Summers, P.W., Roberts, G.R., et Dann, T.F. (1992b). The relationship between hospital admissions and ambient air pollution in Ontario, Canada: a preliminary report. Dans *Proceedings of the Air and Waste Management Association* 85th Annual Meeting and Exhibition, Kansas City, Missouri, 21-26 juin 1992. Réimpression 92-146.05.

Liang, K.Y., et Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Liang, K.Y., Zeger, S.L., et Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society Series* A, 54, 3-40.

Office of Technology Assessment (1984). Acid rain and transported air pollutiants - implications of public policy. U.S. Government Printing Office, Washington, DC.

Prentice, R.L., et Zhao, L.P. (1991). Estimating equations for parameters in means and covariates of multivariate discrete and continuous responses. *Biometrics*, 47, 825-839.

SAS Institute Inc. (1988). *SAS/STAT User's Guide, Release 6.03 Edition*. Cary, NC: SAS Institute Inc., 1028.

Stiratelli, R., Laird, N., et Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40, 961-971.

Thall, P.F., et Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46, 657-671.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika*, 75, 621-629.

SESSION 6

Applications générales I

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

L'ÉCHANTILLON DÉMOGRAPHIQUE PERMANENT: UNE EXPÉRIENCE FRANÇAISE DE SUIVI DE PERSONNES

M. Isnard[1]

RÉSUMÉ

Dès 1968, l'INSEE a mis en place un suivi longitudinal de personnes: il s'agissait d'apparier à un niveau individuel les données issues des recensements successifs et des bulletins statistiques de l'état civil. Le fichier, qui contient plus de 700 000 individus, vivants ou décédés, peut être utilisé pour des études différentielles (mobilité, mortalité), mais aussi des études méthodologiques, notamment sur certaines variables du recensement. En 1995, une nouvelle gestion informatisée remplacera la gestion manuelle actuelle, source de lourdeurs et d'erreurs.

MOTS CLÉS: Recensement; bulletin de l'état civil.

1. INTRODUCTION

La plupart des statistiques démographiques utilisent les recensements et l'état civil. La première source a l'avantage de fournir, sur la totalité de la population, un certain nombre d'informations à une date donnée (par exemple, le lieu de résidence ou la catégorie sociale). On pourrait la comparer à une photographie ou à l'examen d'un stock. L'état civil, quant à lui, s'intéresse plus aux flux: il permet de mieux connaître et comprendre les différents événements démographiques qui arrivent à l'intérieur d'une population.

Ces deux sources complémentaires sont utilisées par le démographe en fonction du phénomène étudié. Toutefois, ces données se révèlent insuffisantes. Par exemple, pour mesurer la fécondité par milieu social, la méthode classique, utilisant l'état civil et les recensements, est inadaptée. En effet, la catégorie sociale est une caractéristique difficile à mesurer et, les conditions de relevé de l'information, un même individu peut être classé dans des catégories différentes.

Par ailleurs, l'état civil et les recensements ne détaillent pas l'histoire passée des individus, même sur le seul plan démographique. Par exemple, une étude de la mobilité géographique ou professionnelle sur une vingtaine d'années est impossible à l'aide des recensements. Pour aller plus loin, il est nécessaire d'utiliser des outils plus adaptés, tels que les enquêtes rétrospectives, où l'on interroge les personnes sur leur passé, ou les «panels», c'est-à-dire des systèmes d'information qui accumulent l'information par un suivi individuel.

Dans les enquêtes rétrospectives, les réponses peuvent souffrir d'un défaut de mémoire. Certains événements sont purement oubliés ou passés sous silence, qu'il s'agisse de courtes d'interruptions d'activité ou même de naissances d'enfants décédés en bas âge.

Pour ces diverses raisons, ainsi que pour mieux cerner la qualité des recensements, l'INSEE a mis en place un fichier longitudinal appelé *Échantillon Démographique Permanent* (EDP). Il s'agit d'apparier à un niveau individuel les bulletins individuels des différents recensements depuis 1968 et les bulletins de l'état civil concernant le même individu (Voir annexe 1: le contenu de l'EDP). Afin de permettre une identification plus facile des individus, le Répertoire National des Personnes Physiques (RNIPP) a été utilisé: un numéro d'inscription au répertoire permet de suivre plus facilement chaque individu.

[1] M. Isnard, INSEE, 18 Bd Adolphe Pinard, 75675, Paris Cedex 14.

2. L'ÉCHANTILLON DÉMOGRAPHIQUE PERMANENT DE L'INSEE

Un échantillon ...

Il était difficile de concevoir un appariement entre les données de l'état civil et du recensement pour la totalité de la population française, aussi a-t-il été décidé de ne conserver que pour un sous-ensemble de la population. Afin d'obtenir un taux de sondage voisin de 1%, 4 jours de naissance ont été choisis et toute personne née un de ces quatre jours fait partie de l'EDP dès qu'un bulletin décrit en annexe et concernant cette personne a été collecté.

... Démographique ...

Le but de cet échantillon est de pouvoir mener des études démographiques. Le contenu du recensement et des bulletins de l'état civil diffère de celui d'autres pays: aucune notion de revenu n'est, par exemple, présente dans le recensement et il est légalement interdit à l'INSEE de mener à bien des études prenant en compte des données médicales (causes de décès par exemple). Le but de ce fichier est plutôt de mener à bien des études démographiques ou sociologiques, qui nécessitent un suivi individuel.

... Permanent

L'EDP a débuté en 1968 et son enrichissement se poursuit de nos jours. Actuellement, sont ajoutées les données provenant du Recensement de la Population 1990 et des bulletins de l'état civil des années 1982 à 1989. Un projet d'enrichissement continu est en cours d'élaboration: il devrait se terminer au cours du premier trimestre 1994.

3. LA GESTION DE L'EDP

Créé en 1968, l'EDP se caractérise par une gestion manuelle lourde et délicate à gérer. À chaque individu correspond une chemise en papier regroupant les différents bulletins le concernant. Ce dossier est stocké dans l'établissement régional de l'INSEE correspondant à son lieu de naissance. Il y a, à l'heure actuelle, 711 038 dossiers regroupant environ 3 millions de bulletins de tout type. L'insertion d'un bulletin dans un dossier est entièrement manuelle et est rendue possible grâce à un classement ad hoc des dossiers. Le RNIPP n'est utilisé qu'en cas de problèmes d'identification. L'insertion des bulletins du recensement de 1990 et de l'état civil 1982-1989 est en cours. Elle se terminera à la fin de l'année 1992 et sera suivie d'une phase de chiffrement, puis de saisie de l'information. Le fichier complet comprenant les bulletins des années 1968 à 1990 sera disponible vers la fin de l'année 1994.

Toutefois, l'information contenue dans les bulletins individuels du recensement est insuffisante. Aussi, a-t-il été décidé d'enrichir partiellement le fichier magnétique de l'EDP des informations concernant le logement et la famille de l'individu aux recensements de 1975 et de 1982. De même, les informations concernant le décès de l'individu ont été récupérées au RNIPP.

Bien entendu, l'information n'est complète que si l'individu se trouvait en France métropolitaine aux différents recensements, y a été recensé et a précisé sa date de naissance.

4. LES ÉTUDES MENÉES À PARTIR DE L'EDP

Malgré ces limitations, l'échantillon offre de multiples possibilités, surtout si l'on tient compte de l'information fournie par l'état civil. À titre d'exemple, le taux de sondage de 1% rend possible l'étude de la mobilité résidentielle sur 4 recensements (1962, 1968, 1975, 1982) avec environ 315 000 individus, la présence du bulletin de mariage permet de mettre en évidence l'influence du changement d'état matrimonial sur cette mobilité.

De plus, la méthode de constitution de l'EDP évite tout effet du type «effet de mémoire» et permet même de le mesurer dans certains cas. Le champ des études réalisables est vaste et les indications données ci-après ne sont nullement exhaustives. Le lecteur trouvera, en annexe 2, une liste des études publiées utilisant le fichier.

Les études démographiques différentielles

Elles ont pour objet, par rapprochement des informations tirées de l'état civil et des recensements, de calculer des taux démographiques relatifs à la fécondité, la nuptialité ou la mortalité. En particulier, l'EDP permet de mieux mettre en évidence les liens entre la mobilité professionnelle (changement de profession) et la mortalité. Les échantillons classiques de mortalité ne permettent pas de calculer ce type d'indicateurs.

Cycle de vie, mobilité, migrations

L'EDP permet de reconstituer une image du cycle de vie familiale et professionnelle des individus, notamment en relation avec leur mobilité spatiale. L'analyse des changements de situation des individus entre deux bulletins de recensement permet de donner une mesure de la mobilité professionnelle ou géographique. L'utilisation des bulletins de l'état permet de mieux étudier le lien entre les deux formes de mobilité et la nuptialité ou la fécondité.

Études méthodologiques

L'EDP permet des études méthodologiques relatives à la qualité des recensements. Bien que ce champ reste grand ouvert, on peut citer des études d'omissions au recensement pour les personnes décédées peu avant ou peu après le recensement, des études sur les doubles comptes, c'est-à-dire les personnes qui ont été recensées à tort plusieurs fois (on possède alors plusieurs bulletins de recensement dans le dossier). La qualité des réponses à certaines questions (diplôme, nationalité) pour deux recensements successifs a pu être ainsi étudiée (voir annexe 3: étude méthodologique sur les diplômes).

Par ailleurs, le rapprochement de bulletins de recensement et de l'état civil permet d'étudier la cohérence, entre ces deux sources, de variables «délicates» telles que la catégorie socioprofessionnelle ou la nationalité.

L'EDP comme base de sondage

L'EDP a aussi été utilisé comme base de sondage pour un certain nombre d'enquêtes organisées par l'INSEE:

- L'enquête socio-démographique de Lorraine. Il s'agit d'une opération régionale mise en place par la direction régionale de Lorraine avec la collaboration de l'Université de Nancy qui a pour objectif la description sociale et économique des ménages en Lorraine, ainsi que l'analyse de l'impact des politiques sociales sur les individus. L'échantillon a été constitué d'individus tirés au hasard dans l'EDP: l'intérêt de cette base de sondage est de ne pas avoir à collecter les informations concernant la période 68-82.

- Les études sur l'inscription sur les listes électorales et la participation aux élections. Ces études ont été menées sur des sous-échantillons de l'EDP, pour lesquels des caractéristiques du recensement de 1975 ont été conservées.

- L'enquête Patrimoine au Décès. Il s'agissait ici de mieux comprendre les comportements des individus en ce qui concerne la transmission du patrimoine au moment de leur décès. Un échantillon d'individus décédés en 1988-1989 a été choisi dans l'EDP et une enquête a été menée dans les trésoreries générales. L'avantage de l'EDP pour cette enquête était de pouvoir bénéficier sans interrogation supplémentaire de la situation de l'individu en 1968, c'est-à-dire en période d'activité.

- L'enquête «Mobilité Sociale et Géographique». Le but de cette enquête, actuellement sur le terrain, est de fournir une mesure de l'intégration des immigrants à la société française. Afin de mieux mesurer l'effet «deuxième génération», deux sous-échantillons ont été tirés: le premier est constitué d'individus tiré du recensement de 1990 et nés à l'étranger aux alentours des années 1960 et ayant immigré en France entre 1982 et 1990. Le deuxième était constitué d'individus appartenant aux mêmes générations, mais nés en

France métropolitaine de père naturalisé ou de nationalité étrangère: seule une source comme l'EDP permet de les repérer.

5. LA NOUVELLE GESTION DE L'EDP

La gestion manuelle des 711 038 dossiers était très lourde et pouvait générer des erreurs. À titre d'exemple, le chiffrement des bulletins de la période 1968-1982 a demandé 700 000 heures, soit environ 460 années-personnes. Le chiffrement des bulletins, qui est en train de se dérouler, devrait consommer environ 100 000 heures. De plus, il était difficile de concilier une alimentation continue et une telle manipulation des dossiers papier. Enfin, de plus en plus de bulletins de l'état civil parviennent à l'INSEE sous une forme magnétique et le stockage papier ne semblait pas être une solution de long terme.

Aussi, l'INSEE a-t-il décidé de concevoir une nouvelle gestion beaucoup plus automatique de l'EDP. Cette gestion, qui sera mise en place à la fin du premier trimestre 1994, doit permettre un enrichissement continu à l'aide des bulletins d'état civil et une alimentation automatique dans le cas des bulletins informatisés. Le principe est le suivant: il s'agit d'utiliser le RNIPP afin d'identifier un individu grâce à l'état civil figurant sur le bulletin. Une fois l'identification faite, les informations seront directement récupérées dans les fichiers statistiques de l'état civil. Seuls les litiges d'identification ne pourront pas être insérés immédiatement et devront faire l'objet d'enquêtes.

Il a été estimé que le temps nécessaire au traitement complet (litiges compris) d'une année de l'état civil serait, en régime courant, d'environ 15 000 h, chiffrement, saisie et traitement des litiges compris. Les bulletins d'une année ne devraient être intégrés avant la fin de l'année n+2.

ANNEXE 1

LE CONTENU DE L'EDP

Les bulletins collectés

Les bulletins individuels des recensements de 1968, 1975 et 1982 ont été collectés et insérés dans les dossiers, qu'il s'agisse de bulletins ordinaires ou de bulletins de personnes vivant en collectivité.

Les bulletins de l'état civil suivants ont été insérés:

- bulletin de mention en marge concernant un individu EDP;
- bulletin de reconnaissance concernant soit un enfant EDP, soit un parent EDP;
- bulletin de mort-né concernant un parent EDP;
- bulletin de décès d'une personne EDP (avant 1974);
- bulletin de naissance si l'enfant, le père ou la mère est EDP;
- bulletin de mariage pour lequel l'époux, l'épouse ou l'enfant légitimé est EDP.

Ces mêmes bulletins sont en cours d'insertion pour la période 1982-1990, de même les bulletins individuels du recensement de 1990.

Le fichier informatique

Le fichier informatique contient, en plus des bulletins mentionnés ci-dessus, des informations concernant le logement de l'individu et sa famille pour les recensements de 1975 et 1982. Un appariement avec le RNIPP a permis de connaître les dates de décès des personnes nées en métropole et décédées en 1990 ou avant.

Il est stocké sous la forme d'une base SAS contenant 1 295 variables.

ANNEXE 2

LISTE DES PRINCIPALES PUBLICATIONS CONCERNANT L'EDP

Janvier 1987: L'échantillon démographique permanent de l'INSEE. Courrier des Statistiques (O. SAUTORY, INSEE).

1988: Participation électorale. Économie et statistique n° 152, 165 et 178 (J. MORIN, INSEE).

Février 1988: note interne. Étude sur les déclarations des diplômes aux recensements de la population de 1975 et 1982 (O. SAUTORY).

Avril 1988: Économie et Statistique n° 209. Plus de la moitié de la population a changé au moins une fois de commune entre 1962 et 1982 (O. SAUTORY).

Octobre 1989: 5ème réunion du réseau CICRED (Paris): Mortalité différentielle (M. ISNARD).

Septembre 1990: Congrès de l'ISI (Le Caire): La mobilité en France entre 1962 et 1982 (M. ISNARD).

Octobre 1991: Congrès européen de Démographie: Mobilité géographique d'après l'EDP (G. DESPLANQUES et M. ISNARD).

En cours de publication: Mobilité Sociale (A. CHENU - CNRS/INSEE).

En cours de publication: Les migrations des personnes en Île de France (F. CRIBIER et A. KYCH, CNRS).

ANNEXE 3

ÉTUDE MÉTHODOLOGIQUE SUR LES DIPLÔMES

Lors de la réalisation d'une étude différentielle, il est important que les critères qui définissent les différentes sous-populations soient les plus fiables possible (et bien sûr pertinents quant aux phénomènes étudiés).

Un critère explicatif classique des différences, en démographie et en sociologie, est le niveau de diplôme. C'est une variable quasi fixe pour les personnes âgées d'au moins de 25 ans. On a comparé, grâce à l'EDP, les déclaration de diplômes de 56 000 individus aux recensements de 1975 et de 1982.

Le tableau ci-dessus résume le niveau de diplôme aux deux recensements pour les individus ayant répondu 2 fois. Seuls 85 % des déclarations sont cohérentes.

Catégorie de population	dip75 = dip82	dip75 > dip82	dip75 < dip82
Ensemble	84,8 %	7,5 %	7,7 %
Hommes	82,7	8,5	8,7
Femmes	86,7	6,6	6,7
Age en 1975			
25-34 ans	77,9	11,7	10,4
35-44 ans	83,0	8,8	8,2
45-54 ans	86,0	6,8	7,2
55-64 ans	87,9	5,0	7,0
65 ans et +	93,0	2,7	4,3
Nationalité en 1982			
Française	84,6	7,6	7,8
Étrangère	92,2	3,9	3,9
CS en 1982			
Agriculteurs	90,8	5,8	3,4
Artisans, commerçants	78,4	10,2	11,4
Cadres supérieurs	70,8	13,2	16,0
Prof. intermédiaires	69,0	17,4	13,7
Employés	79,5	10,5	9,9
Ouvriers	88,4	5,8	5,8
Retraités	90,2	4,3	5,5
Autres inactifs	88,2	5,4	6,4

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

EXPÉRIENCES MÉTHODOLOGIQUES RELATIVES À L'ENQUÊTE «SURVEY OF INCOME AND PROGRAM PARTICIPATION»

R.P. Singh[1]

RÉSUMÉ

L'enquête «Survey of Income and Program Participation (SIPP)» est une enquête longitudinale menée auprès des ménages par le Bureau of the Census des É.-U. Elle produit des données transversales et longitudinales sur la population active, le revenu, les programmes gouvernementaux et d'autres caractéristiques des personnes et des ménages qui peuvent influer sur le bien-être économique des individus. Le Bureau of the Census des É.-U. a créé un vaste programme de recherche et d'évaluation pour l'enquête SIPP. Dans le cadre de ce programme, nous avons réalisé environ dix expériences méthodologiques touchant divers aspects d'une enquête: mode d'interview, collecte des données manquantes d'une vague, avantages offerts par l'employeur et cadeau offert aux répondants. Cette communication présente en bref ces expériences méthodologiques et les résultats observés.

MOTS-CLÉS: Taux de réponse; qualité des données; effet de lisière.

1. INTRODUCTION

Le Bureau of the Census réalise des interviews pour l'enquête «Survey of Income and Program Participation» (SIPP) depuis octobre 1983. L'enquête SIPP est une enquête nationale visant à fournir de meilleurs renseignements sur le revenu et sur la participation aux programmes gouvernementaux de la population des États-Unis excluant les pensionnaires d'établissements institutionnels. Avec l'enquête SIPP on recueille aussi des renseignements sur les caractéristiques des personnes et des ménages qui peuvent avoir une incidence sur le revenu et sur la participation aux programmes. Ces informations sont essentielles pour améliorer la capacité des organismes fédéraux à formuler et à évaluer leurs politiques et leurs programmes dans les domaines du revenu et du bien-être social.

L'enquête SIPP est la première enquête longitudinale à grande échelle du Bureau menée auprès des ménages et c'est une enquête très complexe. Le Census Bureau n'avait aucune expérience avec une enquête de ce genre et un certain nombre de questions et de problèmes n'étaient pas résolus. Par conséquent, en 1984, le Bureau a entrepris un important programme de recherche et d'évaluation afin de trouver les points forts et les points faibles de l'enquête SIPP. Ces travaux comprenaient également de la recherche sur les façons de réduire ou d'éliminer les faiblesses dans les données et de rendre l'enquête SIPP plus efficiente. Dans le cadre de cette recherche, nous avons effectué des expériences méthodologiques portant sur l'enquête SIPP. Cette communication présente un bref résumé de ces expériences et de leurs résultats.

Dans la section 2 de cette communication on présente le plan d'échantillonnage de l'enquête SIPP. Dans la section 3 on traite brièvement des expériences méthodologiques relatives à l'enquête SIPP. La section 4 renferme un résumé et des conclusions.

---

[1] R. Singh, Demographic Statistical Methods Division, U.S. Bureau of the Census, Washington (DC) É.-U. 20233. Cette communication présente les résultats généraux de recherches entreprises par les employés du U.S. Census Bureau. Les opinions exprimées sont celles de l'auteur et ne représentent pas nécessairement celles du Census Bureau.

2. PLAN D'ÉCHANTILLONNAGE DE L'ENQUÊTE SIPP

L'enquête SIPP est une enquête pour laquelle on utilise un échantillon systématique stratifié à plusieurs degrés de la population des États-Unis excluant les pensionnaires d'établissements institutionnels. Seules les personnes qui ont au moins 15 ans sont admissibles pour être interviewées, bien que certaines données sur les enfants soient aussi recueillies à l'aide d'interviews par personne interposée.

Au début, un échantillon d'unités d'habitation est tiré dans des unités primaires d'échantillonnage (UPÉ) choisies. L'échantillon de l'enquête SIPP est divisé en quatre groupes de même taille appelés groupes de renouvellement. Un groupe de renouvellement est interviewé chaque mois. En général, un ensemble de quatre groupes de renouvellement interviewés l'un après l'autre pendant quatre mois est appelé un cycle. Les personnes dans l'échantillon sont interviewées une fois par période de quatre mois pendant 32 mois. La période de référence pour l'interview est la période de quatre mois qui précède le mois de l'interview. Les personnes de 15 ans et plus présentes, comme membres du ménage, au moment de la première interview font partie de l'enquête pour toute la période de 32 mois. Avec certaines restrictions, nous suivons ces membres de l'échantillon initial s'ils déménagent à une nouvelle adresse. Les «nouvelles» personnes habitant avec des membres de l'échantillon initial font aussi partie de l'échantillon mais seulement pendant qu'elles habitent avec ces membres de l'échantillon initial. Pour plus de détails sur le plan d'échantillonnage de l'enquête SIPP, voir Nelson, McMillen et Kasprzyk (1985).

Le questionnaire de l'enquête SIPP est long et complexe. On pose des questions à propos de l'actif et de l'activité sur le marché du travail et sur des genres particuliers de revenus monétaires et non monétaires selon les mois où ces revenus sont reçus et les montants mensuels. Pour de nombreux genres d'éléments d'actif selon l'activité sur le marché du travail ainsi que selon les revenus, des questions additionnelles sont posées aux répondants. Dans la majorité des interviews, les questionnaires comprennent aussi des questions additionnelles (modules spécialisés) sur divers sujets.

3. EXPÉRIENCES EN MÉTHODOLOGIE

À cause de la nature longitudinale de l'enquête SIPP, les procédures qu'on emploie pour cette enquête diffèrent considérablement de celles utilisées pour les autres enquêtes du Census Bureau. Les résultats des expériences réalisées pour d'autres enquêtes ne pouvaient donc être appliqués directement à l'enquête SIPP. C'est pourquoi nous avons conçu et réalisé, pour cette enquête, les expériences d'ordre méthodologique suivantes:

1. Rétroaction à propos de l'actif et du passif
2. Avantages fournis par l'employeur
3. Cadeau au répondant
4. Maximisation des interviews téléphoniques
5. Données manquantes pour un cycle
6. Huitième interview
7. Utilisation d'un agenda pour aider les répondants
8. Recherche cognitive
9. Participation de répondants à une séance de compte rendu
10. Vérification des dossiers - sommaire des sources de revenu (Income Source Summary (ISS))

Ci-dessous, je traite brièvement de ces expériences ainsi que de leurs résultats. Certains de ces derniers ont déjà été présentés lors de diverses réunions de statisticiens.

1. Expérience de rétroaction à propos de l'actif et du passif

L'enquête SIPP recueille des données sur l'actif et le passif à cause de l'importance de ces éléments pour la détermination de l'admissibilité à des programmes et l'évaluation de la situation financière des familles. Les données sur l'actif et le passif ont été recueillies à l'aide des modules spécialisés des 4^{e} et 7^{e} cycles du panel de 1984. Par conséquent, cette expérience était conçue pour déterminer si un répondant déclarait des données de meilleure qualité lorsqu'on lui fournissait les renseignements qu'il avait déclarés l'année précédente sur son actif et son passif. La justification de l'expérience était le fait que l'on pensait que les répondants fourniraient des estimations plus fiables lors du 7^{e} cycle si on leur rappelait tout d'abord les montants qu'ils avaient déclarés lors du 4^{e} cycle pour l'année précédente. Si un répondant connaissait le montant de la variation dans la valeur de

son actif, il y aurait alors cohérence entre la réponse lors du 7e cycle et le montant de la variation pendant l'année.

Pour l'expérience, nous avons divisé l'échantillon en deux moitiés. Pour une moitié nous avons utilisé les procédures d'interview courantes. Pour l'autre moitié, l'intervieweur remettait à chaque répondant, au début de la partie de l'interview portant sur l'actif et le passif, une formule de rétroaction, produite par ordinateur, sur son actif. L'intervieweur lisait aussi une brève introduction dans laquelle on expliquait que le document produit par ordinateur renfermait des renseignements recueillis à propos du répondant une année auparavant et que ce dernier devrait s'y reporter quand il répondrait à des questions semblables pendant l'interview. L'intervieweur aidait le répondant, au besoin, en se reportant à un article sur la formule de rétroaction dans lequel on trouvait les renseignements correspondants fournis lors du 4e cycle. Pour des raisons de confidentialité, les formules de rétroaction n'étaient utilisées que lorsque la même personne répondait lors des deux cycles.

Pour évaluer l'expérience nous avons fait un test afin de vérifier si les variations annuelles dans les valeurs nettes moyenne et médiane et la corrélation entre deux années différaient pour les deux groupes. Les analyses portaient sur des sous-groupes de la population. Les évaluations des variations pour ces estimations n'ont pas fourni de preuve statistique de l'existence d'écarts cohérents entre les deux groupes. Par exemple, la diminution estimée de la valeur nette médiane pour le groupe avec rétroaction et pour le groupe sans rétroaction était de $590 et de $860 respectivement. Ces variations n'étaient pas statistiquement significatives. De plus, l'analyse n'a pas fait ressortir d'accroissement statistique important dans les corrélations pour les groupes de population étudiés (Lamas et McNeil 1987; Weidman et coll. 1988).

Bien que la procédure de rétroaction semblait réduire les estimations des variations et augmenter les corrélations entre les estimations pour deux années, les résultats ont des implications analytiques négligeables. Cela peut être dû aux très petites variations qui se produisent à court terme dans l'actif et le passif. À cause du fardeau additionnel imposé au répondant et du fait que l'enquête SIPP ne permet pas de détecter des changements à court terme dans l'actif et le passif, la collecte de telles données, deux fois pendant la durée d'un panel, a été abandonnée.

2. Expériences portant sur les avantages fournis par l'employeur

Même si, dans le cadre de l'enquête SIPP, on recueille une quantité considérable de données sur la population active, sur la source des revenus et sur le montant du revenu selon sa source, on n'obtient pas de données sur les contributions patronales pour les assurances, pour la retraite, etc. En 1986, le Bureau a étudié la possibilité de recueillir des données sur les contributions patronales aux régimes d'assurance-vie, d'assurance-maladie et de pension auprès de l'employeur du répondant. Des données de haute qualité tirées des dossiers de l'employeur compléteraient les données de l'enquête SIPP et seraient extrêmement utiles pour les chercheurs et les décideurs.

Puisque l'enquête SIPP est très complexe et qu'elle recueille des données de nature très délicate, l'ajout d'une autre demande de données qui pourrait être de nature délicate mettait les répondants et les intervieweurs dans une situation plus difficile, ce qui aurait pu entraîner une augmentation du roulement des intervieweurs. Les intervieweurs s'attendent à des réactions négatives de la part des répondants quand les questionnaires sont plus longs et qu'ils comportent des parties de nature délicate. Il se peut que la réaction négative ne soit pas limitée seulement aux questions additionnelles, elle peut avoir une incidence sur le rendement global des intervieweurs. À cause de ces inquiétudes, l'expérience a été conçue dans le but d'estimer le taux de réponse pour obtenir, de l'employeur du répondant, des données sur les avantages fournis par l'employeur plutôt qu'un taux de réponse global.

L'expérience a été réalisée en août 1987 et portait sur la moitié de l'échantillon du 4e groupe de renouvellement du 8e cycle du panel de 1985. L'expérience ne visait que les personnes occupées, de 18 ans ou plus, qui avaient répondu au questionnaire de l'interview du 8e cycle. En tout, 1,352 personnes pouvaient participer à l'essai lors du 8e cycle.

Un fois l'interview pour le 8e cycle terminée, l'intervieweur déterminait si le répondant pouvait participer à l'expérience. Si c'était le cas, l'intervieweur demandait au répondant de signer une «Employer Questionnaire

and Authorization Form» (questionnaire et formule d'autorisation pour l'employeur) afin d'autoriser le Bureau à demander à l'employeur de fournir des renseignements portant sur les contributions aux régimes d'assurance et de pension. Au besoin, la formule était laissée à une personne interposée ou postée au répondant dans le cas d'une interview téléphonique. Nous n'avons pas effectué de suivi dans le cas des personnes qui n'ont pas retourné de formule. Lorsque nous avons obtenu une autorisation, les questionnaires ont été postés aux employeurs. Nous avons effectué un suivi auprès des employeurs qui n'ont pas retourné tous leurs questionnaires.

Seulement 40% des répondants admissibles de l'enquête SIPP ont signé les formules d'autorisation. Un suivi auprès des répondants aurait pu permettre d'accroître ce taux. Les employeurs ont retourné 96% des questionnaires. Le taux de non-réponse partielle était généralement faible. La réaction des intervieweurs a aussi été très positive. Pour plus de détails, voir Adams (1988).

3. Expérience avec remise de cadeau

Comme l'enquête SIPP est une enquête longitudinale, les taux de non-réponse pour cette enquête sont plus élevés que pour les autres enquêtes démographiques du Bureau. Dans une tentative visant à réduire le taux de non-réponse, nous avons donné un cadeau symbolique (une petite calculatrice à énergie solaire) immédiatement après que l'intervieweur ait présenté l'enquête aux ménages du 1er cycle du 4^{e} groupe de renouvellement du panel de 1987. Les membres des trois autres groupes de renouvellement n'ont pas reçu de cadeau.

Les taux de non-réponse pour le groupe de renouvellement dont les membres ont reçu un cadeau ont été comparés avec ceux des trois autres groupes de renouvellement et avec ceux des panels de 1984, 1985 et 1986. Les résultats des analyses laissaient supposer qu'il se peut que le cadeau aide à réduire le taux de non-réponse au niveau national. Toutefois, les taux d'accroissement d'un cycle à l'autre pendant la durée d'utilisation du panel pour le groupe de renouvellement dont les membres ont reçu un cadeau et pour ceux dont les membres n'ont pas reçu de cadeau ne différaient pas significativement (Butler 1991).

Les résultats obtenus pour les panels de 1988 et de 1989 ont laissé supposer que l'intervention de la Field Division (division des travaux sur le terrain) réduisait autant le taux de non-réponse que lorsqu'on donnait un cadeau. Par exemple, pour le panel de 1988, le taux de non-réponse moyen pour les trois premiers groupes de renouvellement était de 7.71%. Avant la réalisation de l'interview pour le 4^{e} groupe de renouvellement, la Field Division, préoccupée par les taux de non-réponse élevés, a demandé aux bureaux régionaux les raisons qui expliquaient l'accroissement des taux de non-réponse pour les panels déjà interviewés. Suite à cette demande, le taux de non-réponse pour le 4^{e} groupe de renouvellement a diminué pour atteindre 6.73%. Une situation semblable s'est produite dans le cas du panel de 1989.

Le cadeau semblait avoir eu des effets positifs sur le taux de réponse au début de l'utilisation du panel. Toutefois, l'intervention de la Field Division semblait être tout aussi efficace pour ce qui est de diminuer les taux de non-réponse que le fait de donner un cadeau. Il n'a pas été possible de déterminer si l'intervention de la Field Division combinée avec la remise d'un cadeau réduirait davantage les taux de non-réponse. Rien ne prouve que le cadeau ait eu un effet après la première interview. Il est possible que le fait de donner un cadeau plus d'une fois (de préférence au milieu de la durée d'utilisation du panel) pourrait réduire le taux de non-réponse.

4. Maximisation des interviews téléphoniques

À cause de la complexité et de la nature délicate des données recueillies dans le cadre de l'enquête SIPP ainsi que de la longueur de l'interview, on croyait généralement que l'interview sur place était le seul mode efficace de collecte des données. Mais, à cause de diverses compressions dans les dépenses engagées pour l'enquête SIPP ainsi que des efforts continus du Bureau pour rendre cette enquête plus efficiente, nous avons étudié l'emploi de l'interview téléphonique pour l'enquête SIPP. L'objectif de l'expérience d'interview téléphonique était de déterminer si les ménages visés par l'enquête SIPP pouvaient être interviewés par téléphone avec une perte négligeable sinon nulle dans la qualité des données.

Pour atteindre notre but, nous avons fait l'essai du mode d'interview téléphonique en deux étapes: une étude de faisabilité et un essai national. Tout cas faisant partie de l'ensemble des cas à interviewer par téléphone était

considéré être en mode d'interview téléphonique même s'il avait fallu effectuer une visite sur place pour terminer l'interview. Nous avons appelé ce mode la maximisation des interviews téléphoniques («Maximum Telephone Interviewing» (MTI)).

Le Bureau a réalisé l'essai de faisabilité en juin 1985 pour voir si les répondants à l'enquête SIPP pouvaient être interviewés au téléphone sans accroître le taux de non-réponse. Le taux de réponse pour le mode MTI était aussi élevé que pour le mode de maximisation des visites sur place («Maximum Personal Visit» (MPV)) (Durant et Gbur 1988). Globalement, l'essai était très encourageant.

Nous avons effectué l'essai national pendant la durée du panel de 1986. Nous avons réparti approximativement la moitié des ménages faisant partie de l'échantillon du panel de 1986 de l'enquête SIPP entre chacun des deux modes MTI et MPV par tout le pays. Une lettre de présentation informait les répondants que leur prochaine interview serait effectuée par téléphone. Les intervieweurs ont réalisé les interviews téléphoniques à partir de leur domicile.

Les intervieweurs n'ont pas reçu de formation spéciale en classe pour utiliser le mode MTI mais ils ont terminé un programme d'autoformation avant de commencer une tâche. Les intervieweurs ont effectué une première série d'interviews en mode MTI d'août à novembre 1986 (période définie comme la phase I) et une seconde série d'interviews de février à avril 1987 (période définie comme la phase II). Ces interviews ont été réalisées pendant les 2e à 4e cycles inclusivement. Ce plan de sondage permettait de faire des estimations pour les cycles et des estimations des transitions entre deux cycles consécutifs.

Globalement, Gbur et coll. (1989, 1990) ont trouvé des effets minimaux pour les estimations transversales et des biais dans certaines estimations longitudinales. Voici un résumé des résultats de leurs analyses:

- À l'échelle nationale, les taux de non-réponse des ménages étaient identiques pour les deux modes d'interview.
- Le revenu familial médian pour 1986 et 1987, selon les groupes démographiques et géographiques, était identique pour les deux modes.
- Le rapport entre le revenu moyen et le seuil de la pauvreté était différent pour les hispaniques mais pas pour les autres groupes.
- La taille moyenne d'un ménage pour l'échantillon à interviewer par téléphone (2.8 personnes) est inférieure à la taille de l'échantillon désigné pour visite sur place (2.9 personnes).
- Le pourcentage d'autodéclaration pour le mode MTI (62.2%) est inférieur au taux de 64.7% pour les cas interviewés en mode MPV.
- Les taux de participation aux programmes gouvernementaux et les montants reçus en vertu de ces programmes sont plus élevés, mais pas statistiquement différents, pour les cas désignés pour interview en mode MPV.
- Le nombre de transitions pour ce qui est de la réception des sommes versées dans le cadre des programmes (passage de bénéficiaire à non-bénéficiaire) différait pour seulement 3 des 26 sources de revenus étudiées.
- Le pourcentage des personnes qui deviennent pauvres est généralement plus faible parmi les cas interviewés en mode MTI que pour les cas interviewés en mode MPV.
- Un plus fort pourcentage du nombre total de personnes et de sous-groupes choisis (race et sexe) ont terminé leur période de chômage dans le cas des personnes interviewées en mode MTI que parmi celles interviewées en mode MPV.
- Les taux de salaire horaire pour l'ensemble des répondants et pour les Noirs interviewés en mode MTI étaient plus élevés que pour les personnes interviewées en mode MPV.
- Il est arrivé souvent que les répondants en mode MTI n'utilisaient pas les cartes-questionnaires fournies comme aide à l'interview.
- La distribution du nombre d'enfants mis au monde par une femme différait selon le mode d'interview.

- Un pourcentage significativement plus élevé de répondants en mode MPV ont déclaré que le deuxième et le troisième de leurs plus jeunes enfants fréquentaient une garderie.

Les deux derniers résultats sont tirés des données recueillies à l'aide du module spécialisé. La majorité des autres estimations étudiées selon le mode n'étaient pas statistiquement différentes.

McNeil (1989) a comparé le revenu trimestriel, le fait de recevoir des sommes versées dans le cadre d'un programme ainsi que les estimations relatives à la population active pour tout le panel de 1986 avec des estimations semblables tirées du panel de 1985. Il n'a pas trouvé de différence qui soit propre à la période pendant laquelle l'expérience d'interview téléphonique a été réalisée.

Les estimations basées sur les interviews téléphoniques et sur les interviews sur place pour les années civiles 1986 et 1987 ont aussi été comparées aux estimations provenant de la CPS. Les rapports entre les estimations de la variation de 1986 à 1987 ont aussi été comparés. Cette comparaison n'a pas fourni d'indication précise que les estimations basées sur des renseignements obtenus par téléphone étaient meilleures ou pires que celles fondées sur des renseignements recueillis lors d'interviews sur place.

La comparaison des deux modes d'interview laissait supposer qu'il existe des effets minimaux sur les estimations pour la population totale. La majorité des écarts significatifs étaient concentrés dans les groupes à faible revenu. Toutefois, ces écarts étaient peu élevés et ne devraient pas avoir d'effet significatif sur la réalisation des objectifs de l'enquête SIPP. Par conséquent, le Bureau a décidé d'utiliser le mode de maximisation des interviews téléphoniques pour toutes les interviews sauf celles des 1er, 2e et 6e cycles. Les interviews sur place lors des 1er, 2e et 6e cycles aideront les intervieweurs à établir et à entretenir un lien avec les répondants et, par conséquent, à obtenir des taux de réponse élevés.

5. Expérience relative aux données sur un cycle manquant

On considère qu'un ménage répond lors d'un cycle si au moins un membre du ménage répond. Les ménages peuvent être classés en trois catégories selon leur comportement en matière de réponse:

a. ménages qui ne répondent à aucun cycle

b. ménages qui ne répondent qu'à certains cycles et

c. ménages qui répondent à tous les cycles.

Nous avons recours à la pondération pour corriger la deuxième catégorie dans une estimation transversale. Toutefois, les personnes dans cette catégorie peuvent être traitées par imputation ou comme des cas de non-interview pour une pondération longitudinale (Singh et coll. 1990). Cependant, compte tenu de l'importance analytique des transitions et de la durée des périodes de chômage, nous ne savons pas si, dans le cas des questions pour lesquelles des réponses manquent, il faudrait faire une imputation pour des ménages complets.

L'enquête SIPP permet d'obtenir des données rétrospectives pour un cycle manquant si une interview est réalisée au cours d'un cycle ultérieur. L'expérience sur les données pour un cycle manquant a donc été réalisée afin de voir si l'utilisation de données rétrospectives sur les transitions pourrait améliorer considérablement l'imputation du revenu et les procédures d'estimation longitudinale de l'enquête SIPP.

Le Bureau a recueilli des données sur le cycle manquant à la fin de l'interview ordinaire à l'aide d'un questionnaire abrégé intitulé «Missing Wave Section» (section sur le cycle manquant). Cette section renfermait un ensemble réduit de questions de base de l'enquête SIPP portant sur l'activité sur le marché du travail touché, sur l'actif et sur la participation aux programmes. Comme la période dont l'enquêté doit se rappeler est plus longue et que cela pourrait avoir un effet négatif sur la qualité des données, on n'a utilisé le questionnaire pour le cycle manquant que dans le cas des personnes dont le cycle de réponses pendant trois interviews consécutives était: réponse-non-réponse-réponse. Cela limitait la durée de la période dont l'enquêté devait se rappeler à un maximum de 8 mois. Nous avons utilisé la section sur la cycle manquant pour la première fois lors du 4e cycle du panel de 1984. Les données tirées de la dernière interview réalisée pour chaque groupe de renouvellement du panel de 1984 ont été analysées afin d'évaluer l'utilité des données sur le cycle manquant. Les analyses

portaient sur les transitions relatives au fait de toucher un revenu, à l'actif et à l'aide gouvernementale. Un bref résumé figure ci-dessous, pour les détails, voir Huggins (1987).

Les questions de la section sur le cycle manquant n'ont permis de détecter qu'un petit nombre de changements liés au fait de toucher un revenu et dans l'actif. Cinq cent douze personnes étaient admissibles à répondre aux questions pour le cycle manquant. Trente-huit de ces personnes ont déclaré une transition pour ce qui est de la réception d'un des genres de revenu et une personne a déclaré un changement pour deux genres de revenu reçus. Seulement deux et quatre changements relatifs aux sommes reçues dans le cadre de la AFDC et de la sécurité sociale, respectivement, ont été déclarés. Une personne a déclaré un changement dans les pensions alimentaires, 68 personnes ont déclaré une transition dans un élément d'actif et une seule personne a déclaré des transitions dans deux éléments d'actif. Les transitions tirées des données inscrites sur la formule pour le cycle manquant étaient beaucoup plus faibles que les données repères estimées.

Puisqu'elle a permis de trouver un nombre proportionnellement faible de transitions relatives au fait de toucher un revenu et à l'actif, la formule utilisée pour le cycle manquant n'améliorera pas considérablement notre imputation pour la pondération longitudinale. De plus, l'effet du nombre de transitions perdues parce qu'on n'a pas recueilli de données sur un cycle manquant pour l'imputation et pour la pondération devrait être négligeable. Le Bureau a donc cessé, en 1988, de recueillir des données sur un cycle manquant.

6. Huitième interview

Nous avons effectué l'étude portant sur la huitième interview (aussi appelée le 8e cycle) en août 1988 afin de déterminer si nous pouvions réduire le problème de lisière dans l'enquête SIPP à l'aide d'autres procédures d'interview. Le problème de lisière consiste en une surdéclaration des changements dans les sources de revenu et dans les montants des revenus entre des mois consécutifs visés par deux interviews consécutives et en une sous-déclaration des changements entre d'autres mois consécutifs.

Pour l'étude portant sur la huitième interview (Eighth Interview Study (EIS)) on a utilisé un sous-échantillon du 4e groupe de renouvellement du panel de 1986 des ménages interviewés lors du 7e cycle. Les données pour l'étude devaient être recueillies quand ces ménages étaient interviewés pour une interview du 8e cycle à l'aide de la partie de base du questionnaire du 7e cycle. Les ménages dans la EIS ont été interviewés à l'aide d'une des trois procédures d'interview: R, B, ou W. Les ménages visés par la procédure R ont été interviewés à l'aide de procédures d'interview courantes et ont été traités comme un groupe témoin. Pour la procédure B, le répondant recevait une formule de rétroaction sur laquelle figuraient les réponses, portant sur le montant de son revenu mensuel, fournies lors de l'interview du 7e cycle. Pour la procédure W, l'intervieweur encourageait fortement le répondant à utiliser des dossiers pendant toute l'interview. Nous croyions que les procédures B et W pourraient améliorer l'exactitude de la déclaration du moment où des transitions se sont produites.

L'analyse des données de la EIS n'a pas laissé supposer que les procédures B ou W permettaient de réduire le problème de transition ou que ces dernières devraient être appliquées. Nous nous attendions à ce que la procédure B permette de diminuer l'écart entre les cycles, mais elle a entraîné un taux observé plus élevé que la procédure d'interview courante. Nous prévoyions que la procédure W augmenterait la variation à l'intérieur d'un cycle. Toutefois, le pourcentage de variation à l'intérieur d'un cycle obtenu à l'aide de la procédure W n'était pas significativement différent des pourcentages correspondants obtenus avec les procédures R et B.

Pour les procédures B et W on a obtenu des taux d'interview plus faibles que pour la procédure R (Gbur 1990). Toutefois, les taux d'interview auraient pu être différents si les intervieweurs et les répondants n'avaient pas connu les procédures courantes. Nous avons tenu des séances de compte rendu avec les intervieweurs afin de discuter de l'expérience portant sur le 8e cycle. Selon les intervieweurs, la procédure B (la procédure avec rétroaction) semblait améliorer la qualité des données (Singh 1988). Dans certains cas, cette procédure encourageait les répondants à vérifier leurs dossiers et les aidait à fournir de meilleures réponses.

La réaction à la procédure W était un peu désappointante parce que les intervieweurs considéraient que cette procédure ressemblait à ce qu'ils faisaient déjà pour l'enquête SIPP courante. Il semblait que l'expérience n'avait pas été réalisée de la façon dont nous le prévoyions. Il se peut que nous aurions dû avoir recours à la formation

en classe plutôt qu'à l'autoformation pour insister davantage sur l'utilisation des dossiers par les répondants et pour faire ressortir les différences entre les procédures.

7. Utilisation d'un agenda pour aider les répondants

Cette expérience est un prolongement de la procédure B de l'expérience portant sur le 8e cycle. Nous avons utilisé l'agenda (calendrier des événements) pour présenter les réponses, portant sur certains événements relatifs aux programmes gouvernementaux, fournies lors d'interviews antérieures de l'enquête SIPP. Cette présentation aidera les répondants à se souvenir d'événements et à les situer avec exactitude. Cela permettra alors de réduire le problème de lisière.

Nous avons réalisé l'expérience au bureau régional (BR) de Chicago pendant les interviews effectuées auprès du panel de 1989. Pour l'expérience, l'intervieweur donnait au répondant l'agenda de la personne pour laquelle ce dernier fournissait les répondes avant le début de l'interview. L'intervieweur expliquait au répondant l'objet de l'agenda. Après chaque interview, l'intervieweur mettait l'agenda à jour afin que les renseignements qui y figurent quant au statut pour ce qui est de la réception de prestations et du montant du revenu selon la source correspondent à la situation pour l'interview qui venait d'être réalisée.

L'analyse préliminaire (Kominski 1990) laisse entendre que l'agenda peut permettre de réduire le problème de lisière. Kominski fait aussi remarquer que l'agenda facilitait la vérification longitudinale et la correction des données soit pour le cycle courant, soit pour un cycle antérieur. Au début de 1993, la recherche effectuée par Kominski fournira plus de résultats. Ces derniers nous aideront à déterminer si l'on doit ou non utiliser cette méthode.

8. Recherche cognitive

Le problème de lisière a une incidence sur la majorité des estimations de la transition et de la durée des périodes de chômage obtenues dans le cadre de l'enquête SIPP. Le Bureau a entrepris une recherche cognitive afin de réduire l'effet de lisière. Dans le cadre de cette initiative, Cantor et coll. (1991, 1992) ont utilisé la méthode de «pensée à haute voix» pour accueillir des renseignements à propos de la déclaration des prestations ou de la rémunération touchées et des montants correspondants, ils ont demandé aux répondants de paraphraser les questions et de leur faire un compte rendu afin d'évaluer leur compréhension des termes techniques et des sigles (tels que AFDC, SSI) utilisés dans le cadre de l'enquête SIPP et pour obtenir des renseignements plus détaillés par évocation.

La recherche effectuée par Cantor a fourni une base pour la recherche cognitive du Bureau. Le Bureau a planifié sa recherche en trois phases. Moore et coll. (1992) ont présenté la recherche effectuée par le Bureau et je n'en parlerai pas dans la présente communication. La recherche prévue par le Bureau se poursuivra au cours des prochaines années.

9. Participation de répondants à une séance de compte rendu

Nous croyons qu'une collaboration accrue de la part des répondants et qu'une plus grande utilisation de dossiers pendant l'interview donneront des données exactes et, par conséquent, que cela augmentera la qualité des données pour les estimations de petits domaines. Par conséquent, en 1986, le Bureau a effectué une expérience pour que les répondants lui disent: 1) pourquoi ils n'utilisent pas de dossier pendant les interviews de l'enquête SIPP, 2) les raisons qui expliquent leur collaboration et leur participation à l'enquête SIPP, 3) comment les connaissances acquises en participant aux interviews antérieures de l'enquête SIPP ont une incidence sur le comportement des répondants au cours des interviews ultérieures.

Dans le cadre du programme de contrôle de la qualité de l'enquête SIPP, nous réinterviewons certains des répondants après chaque interview mensuelle. Pour cette expérience, toutefois, nous avons eu recours à l'échantillon de réinterview de l'enquête SIPP prévu pour les trois derniers mois du panel de 1985 afin de réduire les coûts et d'obtenir les résultats plus rapidement. Un échantillon de 516 ménages était admissible pour répondre au questionnaire d'évaluation. Seulement une personne par ménage échantillonné pouvait participer

à la réinterview. Il se peut que la personne qui répondait à la réinterview ne soit pas celle qui avait pris part à la dernière interview de l'enquête SIPP.

En général, le taux de réponse était élevé. Globalement, 89.5% des personnes ont répondu au questionnaire d'évaluation. Les répondants semblaient aimer participer au compte rendu. Approximativement 60% des répondants ont déclaré utiliser des relevés bancaires et des talons de chèque de paie pendant les interviews. Environ le même pourcentage de répondants ont déclaré utiliser les formules d'impôt W-2 et de 1986 pendant l'interview à laquelle ils participaient. Ces taux d'utilisation de dossiers, déterminés d'après les résultats des réponses aux questionnaires d'évaluation étaient beaucoup plus élevés que ceux calculés d'après les interviews courantes (environ 30%) et nous ont porté à nous interroger à propos de l'exactitude des réponses aux questions d'évaluation. Cet écart considérable entre les taux d'utilisation de dossiers déclarés lors du compte rendu et pour les interviews courantes pourrait être dû aux procédures différentes utilisées ou peut-être à une interprétation des questions différente de celle qui était prévue.

Environ 2.2% des répondants (12% des bénéficiaires des programmes gouvernementaux) ont déclaré qu'ils ont commencé à participer aux programmes gouvernementaux après avoir appris l'existence d'un programme en prenant part à l'enquête SIPP. Le fait d'avoir appris que de tels programmes existaient a entraîné une modification du comportement des répondants et un biais positif dans nos estimations.

Les raisons les plus fréquemment citées par les répondants pour justifier leur participation à l'enquête étaient «j'aime l'intervieweur (les intervieweurs)» et «par devoir patriotique».

Environ 80% des répondants qui n'ont pas utilisé de dossiers ont déclaré qu'une des raisons principales pour lesquelles ils n'utilisaient pas de dossiers était soit «cela demande trop d'efforts», soit «les dossiers n'étaient pas disponibles», ou «je connaissais les renseignements sans avoir à consulter de dossiers». Quatre-vingt pour cent de ceux qui n'ont pas utilisé de dossiers ont déclaré que rien ne pouvait les inciter à en utiliser. Pour plus de détails, voir Petroni et coll., (1989).

Les conclusions de l'expérience ont mené à l'inclusion, dans la lettre de présentation qui est envoyée aux répondants avant chaque interview, d'une déclaration qui invite ces derniers à consulter leurs dossiers pour répondre aux questions. Ces conclusions ont aussi mené les intervieweurs à téléphoner aux répondants avant l'interview pour leur demander de rassembler leurs dossiers et à la mise en oeuvre de la recherche cognitive.

10. Vérification des dossiers - sommaire des sources de revenu (Income Source Summary (ISS))

Les employés du Census Bureau qui travaillent sur le terrain croient que le fait d'inciter un répondant à utiliser ses dossiers pendant la première interview augmentera l'utilisation que cette personne fera des dossiers dans le cadre de l'enquête SIPP. Nous avons donc réalisé une expérience dans le cadre de laquelle nous incitions les répondants à consulter davantage leurs dossiers pendant l'interview de l'enquête SIPP.

Nous avons insisté, auprès des intervieweurs, sur l'importance de l'utilisation de dossiers pendant une formation de recyclage pour le 1er cycle du panel de 1990 et nous leur avons demandé d'inciter les répondants à faire une plus grande utilisation de dossiers. Pendant le 1er cycle, nous avons aussi demandé aux observateurs sur le terrain (généralement des intervieweurs principaux) de noter l'utilisation de dossiers par les répondants pour certaines questions portant sur les programmes gouvernementaux et d'inscrire leurs observations sur la «SIPP Record Use Study Form» (formule pour l'étude de l'utilisation de dossiers dans le cadre de l'enquête SIPP). Au cours d'interviews ultérieures, les intervieweurs ont relevé l'utilisation de dossiers par source de revenu sur la page du questionnaire utilisée pour le ISS. Nous avons continué d'insister, auprès des intervieweurs, sur l'importance de l'utilisation de dossiers et nous croyions que l'inscription de tels renseignements sur le questionnaire, en plus de fournir une meilleure estimation de l'utilisation de dossiers, constituerait un message encore plus clair aux intervieweurs que l'utilisation de dossiers est un aspect important de l'enquête SIPP. Cela encouragerait aussi les intervieweurs à faire un effort additionnel pour inciter, pendant l'interview du 1er cycle, les répondants à utiliser des dossiers.

L'évaluation des données a révélé que l'emploi de dossiers pendant le 1er cycle du panel de 1990 était faible (Kominski 1991). Les données recueillies lors des cycles ultérieurs ont montré que le taux d'utilisation de

dossiers était à peu près au même niveau. Une évaluation plus récente (Lessard 1992), basée sur le 5e cycle du panel de 1990, a montré que le taux national moyen d'utilisation de dossiers par source de revenu était de 21.44% et que le pourcentage des personnes qui utilisaient au moins un dossier était de 22.53% à l'échelle nationale. Les taux pour les autres cycles et panels étaient semblables.

Dans la recherche cognitive, environ 70% des répondants ont utilisé des dossiers par source de revenu. Ce taux est beaucoup plus élevé que celui observé pour l'interview de l'enquête SIPP courante. Puisqu'il est essentiel d'accroître l'utilisation de dossiers dans le cadre de l'enquête SIPP, pendant les séances de formation et à l'aide de notes de service, nous rappelons continuellement aux intervieweurs d'encourager les répondants à utiliser des dossiers. Dans l'avenir, nous prévoyons analyser les taux au niveau des intervieweurs afin de fournir à ces derniers une rétroaction sur le taux d'utilisation de dossiers par les répondants auprès desquels ils travaillent. Il se peut qu'une telle rétroaction favorise l'utilisation des dossiers.

4. RÉSUMÉ ET CONCLUSIONS

Les expériences relatives à l'enquête SIPP ont contribué de façon importante à l'étude des problèmes éventuels, fournissant des données à propos de questions particulières portant sur l'enquête SIPP et nous aidant à mieux cibler notre recherche. Nous continuons d'analyser certaines des données expérimentales recueillies lors des recherches relatives à l'enquête SIPP. Les recherches réalisées par le Bureau constituent un apport considérable à la méthodologie d'enquête, particulièrement pour les enquêtes longitudinales. Ces recherches nous ont permis de bien comprendre diverses questions liées aux études longitudinales et nous ont aidés à améliorer les procédures et les questionnaires de l'enquête SIPP. Certaines de ces améliorations et modifications ont été mentionnées plus haut.

L'expérience de rétroaction à propos de l'actif a montré que le rappel de renseignements touchant les données recueillies sur l'actif et le passif l'année précédente ne fournissait pas de résultats statistiquement différents de ceux tirés des données recueillies sans rétroaction. De plus, cette expérience a laissé supposer que les variations dans l'actif et le passif étaient peu importantes à court terme. Il faudrait effectuer des recherches afin de déterminer 1) s'il vaut la peine de recueillir des données sur l'actif et le passif plus d'une fois pour un panel plus long, 2) si le fait de rappeler des renseignements améliorera la qualité des données pour les estimations du changement sur une période plus longue.

Pour l'enquête SIPP, il est très important de réduire l'effet de lisière. L'utilisation d'un agenda (Kominski 1990) et la recherche cognitive, Moore et coll. (1992), laissent supposer que l'on peut réduire le problème de lisière. La recherche cognitive a permis de déterminer les concepts et les questions que les répondants ont de la difficulté à saisir. Cette recherche a aussi démontré que l'on peut faire augmenter radicalement l'utilisation de dossiers par les répondants. Toutefois, les taux de réponse actuels pour la recherche cognitive sont de 80% ou moins pour le 1er cycle, taux qui diminueront à mesure que la durée de la participation des membres du panel s'accroîtra. Cela se compare au taux actuel de 93% pour le 1er cycle. Les résultats de la recherche cognitive pourraient avoir une incidence positive importante pour l'amélioration de la qualité des données si le taux de réponse est beaucoup plus élevé que le taux actuel. Il est donc important d'effectuer des recherches afin de déterminer les éléments particuliers (comme: une augmentation du taux de réponse, des interviews de groupe et une augmentation de l'utilisation de dossiers par les répondants) de la recherche cognitive qui permettent d'améliorer la qualité des données. Il faudrait aussi effectuer des recherches sur l'utilisation de la MTI dans le cadre de la méthode cognitive.

Quatre-vingt-seize pour cent des employeurs ont retourné le «Employer Questionnaire and Authorization Form», alors que seulement 40% des répondants ont signé la formule. Cette expérience a permis d'obtenir le renseignement fort valable suivant, soit que les personnes qui réalisent des enquêtes peuvent obtenir des données additionnelles auprès des employeurs. Toutefois, il faut effectuer un effort additionnel pour amener les répondants à signer la formule. De plus, il faudrait effectuer des recherches pour ajouter d'autres données tirées de dossiers administratifs aux données d'enquête. Le fait de recueillir des renseignements supplémentaires améliorera grandement la base de données en lui ajoutant des données de qualité sans augmenter considérablement le fardeau de réponse.

Le don d'un cadeau symbolique a un effet positif sur le taux de réponse pour l'enquête SIPP. Toutefois, cet effet est faible et certaines mesures administratives permettent d'obtenir le même résultat. Il faudrait réaliser des recherches afin d'évaluer l'effet que l'on obtiendrait si l'on donnait un cadeau plus dispendieux ou si l'on offrait des cadeaux deux fois ou plus.

La recherche portant sur la MTI a montré qu'il est possible de mener à bonne fin une enquête de nature complexe et délicate en maximisant l'interview téléphonique. La MTI a eu une incidence minime sur les estimations transversales et longitudinales. Pour l'enquête SIPP, on a commencé à utiliser la maximisation des interviews téléphoniques en février 1992. Nous surveillons les données de près. Nous n'avons pas observé de problème important jusqu'ici, sauf que le taux d'utilisation des cartes questionnaires par les répondants a diminué de 100% à entre 20 et 30%. Il faudrait effectuer des recherches en vue d'accroître l'utilisation des cartes questionnaires et des dossiers dans le cadre de la MTI.

De plus, on devrait effectuer des recherches sur la façon de réaliser l'interview téléphonique centralisée assistée par ordinateur pour rendre les opérations de l'enquête SIPP plus flexibles sans réduire la qualité des données. La flexibilité aidera à gérer la charge de travail qui fluctue sur le terrain.

La collecte de données pour un cycle manquant lors d'une interview ultérieure était considérée une bonne méthode de remplacement, mais elle n'a pas permis d'obtenir des données de qualité dans le cas des questions pour lesquelles on a recueilli des renseignements, probablement parce que la période dont l'enquêté devait se souvenir était plus longue (de 5 à 8 mois). Nous devrions effectuer des recherches à l'aide d'autres méthodes avec plus de rétroaction ou en demandant plus de précisions. Une méthode pour laquelle il vaudrait la peine d'effectuer des recherches consiste à recueillir des données jusqu'au moment de l'interview et à utiliser ces données pour demander des précisions lors de l'interview suivante. Une autre méthode consiste à fournir plus de rétroaction à partir de la dernière interview réalisée. L'interview sur place assistée par ordinateur (IPAO) constituera un véritable atout pour effectuer des recherches sur ces méthodes. La recherche dans ces domaines offrira aussi la possibilité de réduire l'effet de lisière.

L'étude effectuée à l'aide des ISS a fait ressortir un taux d'utilisation des dossiers beaucoup plus faible pour l'enquête SIPP que pour la recherche cognitive. Il faudrait effectuer des recherches afin d'étudier si une augmentation de l'utilisation de dossiers améliore la qualité des données. Si c'est le cas, il faudrait effectuer des recherches portant sur l'utilisation d'autres moyens comme une formation spéciale pour les intervieweurs, la modification de la norme de rendement des intervieweurs afin qu'elle comprenne l'utilisation des dossiers, l'acceptation d'un taux de réponse plus faible si ce dernier est accompagné d'une plus grande utilisation de dossiers, etc.

REMERCIEMENTS

Je désire exprimer ma vive reconnaissance à Sandy Carnegie et à Kathy Kreilick pour leurs travaux de dactylographie et à Rita Petroni, Andrea Meier, Daniel Kasprzyk et Nanak Chand pour leurs commentaires utiles qui ont permis d'améliorer la présente communication.

BIBLIOGRAPHIE

Adams, D. (1988). SIPP 85: Evaluation of the employer-provided benefits study. Note de service interne à Shapiro de Singh, datée du 28 octobre, 1988.

Butler, D. (1991). SIPP 87: Gift experiment results. Note de service interne du Census Bureau à Singh datée du 2 avril, 1991.

Cantor, D., Brandt, S., et Green, J. (1991). Results of first wave SIPP interviews. Note de service à Bowie, datée du 2 février, 1991.

Cantor, D., Green, J., Moesinger, K., Brandt, S., et Rose, P. (1992). Revised draft results of second wave of SIPP interviews. Note de service à Lampe datée du 9 septembre, 1992.

Durant, S., et Gbur, P. (1988). Testing telephone interviewing in the survey of income and program participation and some early results. Document de travail du SIPP, série n° 8824, U.S. Bureau of the Census.

Gbur, P., (1990). SIPP 86: Eighth interview study data analysis. Note de service interne du Census Bureau memorandum pour Shapiro de Singh datée du 9 mai, 1990.

Gbur, P., Cantwell, P.J., et Petroni, R.J. (1990). Effect of maximum telephone interviewing on SIPP topical module and longitudinal estimates. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Gbur, P., et Petroni, R. (1989). Preliminary evaluation of maximum telephone interviewing on the SIPP. *Proceedings of the Survey Research Section of the American Statistical Association.*

Huggins, V. (1987). Evaluation of missing wave data from the survey of income and program participation. *Proceedings of the Section on Survey Research Methods Section of the American Statistical Association.*

Kominski, R. (1990). The SIPP event history calendar: Aiding respondents in the dating of the longitudinal events. *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

Kominski, R. (1991). Record use by respondents. *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

Lamas, E.J., et McNeil, J.M. (1987). An analysis of the SIPP asset and liability feedback experiment. Document de travail du SIPP, série n° 8725.

Lessard, J., (1992). SIPP 90: Wave 5 results of the record check study. Note de service interne du Census Bureau pour le SIPP Research and Evaluation Steering Committee de Singh, datée du 15 juin, 1992.

McNeil, J., (1989). Quarterly estimates of core characteristics: 1984, 1985, and 1986 panels. Note de service interne du Census Bureau pour «The Record», datée du 21 juillet, 1989.

Moore, J., Bogan, K., et Marquis, K. (1992). A «cognitive» interviewing approach for the survey of income and program participation: Development of procedures and initial test results. Présenté au Symposium 92: Conception et analyse des enquêtes longitudinales, tenu à Statistique Canada, Ottawa, 2-4 novembre, 1992.

Nelson, D., McMillan, D., et Kasprzyk, D. (1985). An overview of the survey of income and program participation. Document de travail du SIPP, série n° 8401, mise à jour n° 1.

Petroni, R.J., Huggins, V.J., et Carmody, T.J. (1989). Research and evaluation conducted on SIPP. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census.*

Singh, R. (1988). General comments based on wave 8 experiment debriefing. Note de service interne du Census Bureau pour «The Record», datée du 23 septembre, 1988.

Singh, R., Huggins, V., et Kasprzyk, D. (1990). Handling single wave nonresponse in panel surveys. Document de travail du SIPP, série n° 9009, U.S. Bureau of the Census.

Weidman, L., King, K., et Williams, T. (1988). Further evaluation of the SIPP asset feedback experiment. *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

QUESTIONS MÉTHODOLOGIQUES RELATIVES À LA CONCEPTION DE L'ENQUÊTE «BRITISH HOUSEHOLD PANEL STUDY»

P.C. Campanelli et L. Corti[1]

RÉSUMÉ

On ne peut douter que les études par panel menées auprès des ménages offrent des occasions uniques de réaliser une gamme de projets de recherche méthodologique importants et innovateurs. En même temps, les études par panel posent des défis difficiles pour ce qui est de la conception ainsi que de la qualité et la nature complexe des données qu'elles permettent d'obtenir crée de nouvelles énigmes et de nouveaux problèmes tant pour les chercheurs dans les domaines spécialisés que pour les méthodologistes. Dans la présente communication, on décrit comment pour l'enquête British Household Panel Study (BHPS), on a traité de diverses questions relatives à la conception de l'étude. L'enquête BHPS est une enquête polyvalente à plusieurs cycles récente qui est menée auprès de 5 000 ménages en Grande-Bretagne. On traite de décisions prises par l'équipe qui dirige l'enquête BHPS pour ce qui est du travail initial de conception, de questions courantes en matière de contrôle de la qualité et du programme de recherche méthodologique.

MOTS-CLÉS: Études par panel; enquêtes; méthodologie.

1. INTRODUCTION

L'étude par panel polyvalente nationale menée auprès des ménages constitue un genre de plan de sondage longitudinal qui a connu une popularité croissante. Un exemple de ce genre d'étude est l'enquête «Panel Study of Income Dynamics» qui a été entreprise en 1968 à la University of Michigan (voir Morgan et Duncan 1986). Des études par panel nationales existent maintenant ou sont en cours d'élaboration dans la majorité des pays européens (y compris la Belgique, la France, l'Allemagne, la Grèce, la Hongrie, l'Irlande, le Luxembourg, les Pays-Bas, l'Espagne et la Suède). Toutes ces études ont des buts généralement semblables sans être identiques. L'enquête British Household Panel Study (BHPS) vient de s'ajouter à cet ensemble d'études nationales européennes.

1.1 La British Household Panel Study

L'equête BHPS est le plus gros projet unique jamais financé par le Economic and Social Research Council (ESRC) du R.-U. et elle a été conçue pour fournir des possibilités d'analyse uniques aux utilisateurs tant du milieu universitaire que de l'administration publique. L'enquête est administrée à partir des locaux du ESRC Research Centre on Micro-Social Change sur le campus de la University of Essex, une équipe interdisciplinaire composée de 41 personnes travaille à sa réalisation. Les interviews pour l'enquête ont commencé en 1991 et devraient, en attendant une revue, se poursuivre par cycles annuels jusqu'à, au moins, 1998. L'échantillon probabiliste obtenu de 5 538 ménages et de 10 303 personnes provient de partout en Grande-Bretagne. L'instrument d'enquête, qui est présenté à tous les membres adultes du ménage, comprend une interview en personne de 45 minutes avec chaque répondant, un bref questionnaire à remplir soi-même ainsi qu'un court questionnaire au niveau du ménage. Ces interviews et questionnaires portent sur divers domaines précis: la structure du ménage, le revenu et la richesse, l'expérience du marché du travail, les coûts et les conditions du logement, des questions de santé, le comportement en matière de consommation, le niveau de scolarité et la

[1] P.C. Campanelli et L. Corti, The British Household Panel Study, ESRC Research Centre on Micro-Social Change, University of Essex, Colchester, R.-U. CO4 3SQ.

formation ainsi que des valeurs socio-économiques. Les données sont structurées de façon à permettre aux chercheurs de décrire et d'analyser comment les particuliers, les familles et les ménages vivent les changements dans leur milieu socio-économique et comment ils réagissent face à ces changements. On trouve plus de renseignements sur les objectifs de l'enquête BHPS dans Rose et coll. (1991).

L'objet principal de la présente communication est de donner un aperçu de certaines des questions méthodologiques fondamentales qui surgissent lors de la conception et de la réalisation d'enquêtes par panel menées auprès des ménages et de la façon dont on aborde ces questions dans le cadre de l'enquête BHPS[2]. La communication portera sur des questions méthodologiques et de conception dans un sens multidisciplinaire très étendu avec des exemples tirés de tous les aspects du processus d'enquête depuis la conception initiale jusqu'à la diffusion inclusivement. La communication est divisée en trois sections: questions initiales, courantes et futures.

2. QUESTIONS MÉTHODOLOGIQUES INITIALES

Le fait d'entreprendre une nouvelle étude par panel soulève plusieurs questions fondamentales qui découlent du but théorique de l'enquête et de considérations de coût. Ces questions comprennent notamment quel genre de plan d'enquête par panel devrait être adopté; quel sera l'intervalle entre les cycles, le mode de collecte des données, les règles de conduite adoptées à l'égard des répondants, la taille de l'échantillon et le genre de plan d'échantillonnage; les sujets sur lesquels le questionnaire portera et comment nous nous y prendrons à ce propos; les méthodes que nous utiliserons pour régler les questions relatives à la non-réponse, etc. L'objectif général de l'enquête BHPS était de faire avancer notre compréhension du changement social et économique tant au niveau des particuliers que des ménages en Grande-Bretagne pendant les années 90. Pour l'enquête BHPS, les décisions relatives à la conception, les buts de la recherche ainsi que les considérations de coûts ont interagi des façons suivantes:

Le genre de plan d'enquête par panel

On peut utiliser plusieurs plans pour réaliser une enquête périodique portant sur des particuliers (voir Duncan et Kalton 1987; Kalton 1992). Ces plans comprennent le plan par cohortes, le plan avec renouvellement de panel, le plan à panel fractionné et les enquêtes transversales périodiques ainsi que le plan d'enquête par panel classique composé d'un échantillon représentatif d'unités ou de particuliers qui sont suivis pour une série de cycles. L'enquête PSID et les études nationales européennes semblables tendent à être du dernier genre. Pour l'enquête BHPS aussi, nous avons adopté ce plan, ce qui nous a mis en conformité avec ces autres études par panel menées auprès des ménages et nous a permis d'atteindre nos objectifs de mesurer diverses composantes du changement au niveau des particuliers et de déterminer des phénomènes transitoires ou persistants.

Intervalle entre les cycles

Il existe plusieurs raisons convaincantes pour réaliser les interviews à des intervalles de 12 mois (voir Rose, Buck et Corti 1991; van de Pol 1988). Des intervalles de moins de 12 mois signifient qu'il y a des cycles plus fréquents ce qui augmente les coûts de la réalisation de l'enquête sur le terrain et peut mener à une réduction dans la taille de l'échantillon pour compenser cette augmentation des coûts ainsi qu'à une augmentation du fardeau du répondant et du fardeau organisationnel. De même, il se peut que l'utilisation d'intervalles plus courts ne laisse pas suffisamment de temps pour que certains changements se produisent. Par contre, des préoccupations à propos des erreurs de mémoire constituent des arguments contre l'utilisation d'intervalles plus longs (p. ex., tous les deux ans). La position de compromis de 12 mois présente l'avantage additionnel de correspondre à un concept de délimitation significatif pour le répondant. Il n'est donc pas surprenant que pour l'enquête BHPS, l'enquête PSID et plusieurs études européennes, on ait choisi un intervalle de 12 mois.

---

[2] On peut trouver des descriptions plus complètes des questions et du travail relatifs à la BHPS dans Rose, Buck et Corti (1991) et dans Rose, Campanelli, Corti et Taylor (1992).

Mode de collecte des données

Des interviews en personne ont été réalisées pour le premier cycle de l'enquête. Plusieurs facteurs étaient implicites dans la décision de ne pas passer à l'interview téléphonique pour les cycles ultérieurs, en dépit des économies qui pourraient être réalisées. Un passage à l'interview téléphonique aurait entraîné des modifications dans les procédures et dans les questionnaires ainsi que la possibilité d'introduire des effets de mode entre les cycles. De plus, le téléphone n'est pas un mode d'interview bien accepté pour la réalisation d'enquêtes en Grande-Bretagne.

Règles de conduite adoptées à l'égard des répondants

Les objectifs de la recherche visant à étudier les structures et les processus à l'intérieur des ménages, à examiner les valeurs et les attitudes par rapport au comportement, ainsi que la nature complexe de certaines des questions ont eu une incidence sur la décision de réaliser des interviews en personne avec chaque membre adulte du ménage, plutôt qu'avec seulement le chef du ménage comme cela est fait pour l'enquête PSID et à n'accepter les déclarations par procuration qu'en dernier recours. Cette décision a été considérée essentielle, en dépit de ses conséquences financières importantes.

Taille de l'échantillon

On considérait que la taille minimum de l'échantillon devait être de 5 000 ménages et de 10 000 personnes afin d'obtenir la précision désirée pour les estimations tant de la population dans son ensemble que de sous-groupes importants qui nous intéressaient tels que les familles monoparentales et les personnes âgées. Cet échantillon relativement considérable (selon les critères des enquêtes réalisées par des universitaires) était aussi nécessaire pour qu'il se produise suffisamment d'événements au fil des ans afin qu'on puisse effectuer l'analyse des transitions et pour tenir compte de l'attrition.

Genre d'échantillon

Les premiers travaux que nous avons effectués étaient dominés par le compromis entre les besoins du plan d'échantillonnage et les exigences liées à certains modèles longitudinaux. Il y a eu discussion partisane parmi les membres de notre groupe consultatif sur les plans d'enquête par panel relativement à l'à-propos d'utiliser un plan de sondage en grappes à plusieurs degrés, certains économétres et statisticiens soutenant que pour les modèles statistiques qu'ils désiraient utiliser il fallait employer un plan d'échantillonnage aléatoire simple (EAS) (voir Coxon 1992). Le fait que la dispersion des grappes dans le temps poserait des difficultés immenses pour la modélisation nous préoccupait particulièrement. Les personnes qui connaissent bien les coûts de réalisation, sur le terrain, des enquêtes avec interviews sur place ne seront pas surprises d'apprendre que la mise en application d'un plan avec EAS, même quand on peut disposer facilement du Postcode Address File (PAF) en Grande-Bretagne comme base de sondage, augmenterait les coûts du travail sur le terrain de 30 à 40 % uniquement à cause des déplacements des interviewers. Par contre, les avantages économiques considérables qui découlent de l'emploi d'un échantillon en grappes ne sont accompagnés que d'une petite perte dans l'efficacité de l'erreur-type. De plus, avec un budget fixe pour l'enquête BHPS, un plan avec EAS aurait tellement réduit la taille de l'échantillon que, pour de nombreux genres d'analyses statistiques, cela aurait eu des conséquences critiques. Ironiquement, il aurait été impossible de mettre en application un plan véritablement «sans grappe» puisque le groupement dû aux intervieweurs et aux personnes dans les ménages existerait toujours.

Quels sujets inclure et comment le faire

Le contrat nous liant avec notre organisme de financement nous obligeait à inclure certains domaines spécialisés. Toutefois, cela nous laissait encore une certaine latitude quant à la nature et au nombre de questions à inclure pour chaque sujet. Quatre autres questions étaient considérées pertinentes:

1) **Le besoin de compléter d'autres ensembles de données importants**

   La reproduction de questions peut permettre d'effectuer des comparaisons avec d'autres ensembles de données à des fins de contrevalidation, ainsi que de faire une analyse comparative, tant avec des données

transversales que des données longitudinales provenant d'autres pays. L'enquête BHPS a donc été conçue afin de maximiser la possibilité d'apparier des données, tant avec celles provenant d'enquêtes universitaires et gouvernementales en Grande-Bretagne que des données d'autres études par panel menées auprès des ménages en Europe et aux É.-U.[3]

2) **Le besoin d'obtenir des mesures continues**

Une des caractéristiques précieuses d'une étude par panel est la possibilité que cette dernière offre de recueillir des mesures du changement de façon plus fiable qu'au moyen d'histoires rétrospectives. Toutefois, de nombreuses questions doivent porter sur les événements qui se produisent entre les interviews et sont donc de nature rétrospective. La collecte de telles mesures continues revêt une importance capitale (Duncan 1992). Notre but était de produire des questions qui nous permettraient de construire des mesures continues du revenu, des antécédents de travail et de l'activité, de la structure des ménages et de la mobilité résidentielle pendant le cycle de la vie.

3) **Le besoin d'une composante variable**

On ne peut poser toutes les questions souhaitables au même moment. Le fait de diviser le questionnaire en une composante «principale» et une composante «variable» permet d'ajouter plus de questions et, ce qui est plus important, nous donne une certaine flexibilité pour inclure, dans le futur, des questions qui peuvent être pertinentes aux besoins changeants tels qu'on les verra à ce moment. Cette façon d'agir soulève toutefois la question de déterminer la fréquence appropriée pour les questions posées à intervalle.

4) **Le besoin d'avoir des caractéristiques spéciales pour le plan d'enquête**

La façon d'établir, de façon absolument fiable, le lien entre tous les membres du ménage sans avoir à revenir à un plan matriciel complet (voir Brynin 1992) constitue un exemple d'une question spéciale reliée à la conception.

Non-réponse

Les biais dus à la non-réponse à une question et à la non-réponse d'une unité sont des préoccupations essentielles pour toute enquête et prennent une signification particulière dans le cas d'une étude par panel. Afin d'assurer une représentativité continue, nous avons fait beaucoup d'efforts, dans le cadre de l'enquête BHPS, afin d'obtenir un taux de réponse élevé pour le premier cycle. Nous avons été assez chanceux pour obtenir des interviews (soit directement, soit par procuration) avec TOUS les membres d'un ménage dans 69 pour cent de tous les ménages admissibles. Cela représente une réussite remarquable pour une enquête non gouvernementale britannique, alors qu'un bon taux de réponse, pour obtenir une interview avec AU MOINS UNE personne dans un ménage, s'établit aux environs de 65 à 70 pour cent.

Nous nous sommes concentrés sur l'utilisation de méthodes de travail sur le terrain bien établies afin d'augmenter les taux de réponse et de motiver les intervieweurs ainsi que sur la conception du questionnaire et les stratégies de dépistage des répondants afin d'aider à minimiser la non-réponse tant à une question que par une unité. Citons comme exemple une expérience à groupe fractionné avec récompense donnée au début de l'étude et conçue pour évaluer les effets du versement d'une récompense individuelle (un bon-cadeau échangeable à un magasin à succursales national) sur la réponse initiale. À cause de variations au niveau des intervieweurs, il faut interpréter les résultats avec une certaine prudence, mais il semble que la récompense ait eu un effet positif sur la réponse. Ce résultat, plus les données et recommandations provenant d'autres études (p. ex., Department for Statistics of Income and Consumption 1984; Jean et McArthur 1987) ont suggéré l'adoption de récompenses remises aux répondants pour tout travail sur le terrain. Un autre exemple est la controverse, au niveau de la conception de l'enquête, qui a entouré le problème portant sur la façon de poser un ensemble détaillé de questions sur le revenu des particuliers sans courir le risque d'une non-réponse à une question ou par une unité. Après avoir consulté un groupe d'experts nous avons, pour l'enquête BHPS, choisi

[3] Tous les détails pertinents sur l'origine des questions, entre autres caractéristiques des questions, ont été consignés.

une option intermédiaire qui cherchait à établir un compromis entre l'étendue chronologique et le détail des renseignements sur la période la plus récente. Par exemple, les questions sur la richesse et les biens, considérées les plus délicates, seraient posées lors d'un cycle ultérieur quand les membres du panel seraient suffisamment engagés dans l'enquête. De même, nous avons intégré des questions sur les gains à d'autres mesures de caractéristiques de l'emploi afin de les rendre moins importunes. Bien qu'une expérience à groupe fractionné doive être réalisée pour évaluer ces méthodes, nous avons obtenu, dans le premier cycle, des taux de non-réponse pour le revenu semblables à ceux des enquêtes gouvernementales, soit environ 10 pour cent. Ce chiffre est encourageant pour une enquête universitaire.

Réalisation de la collecte des données

Il y avait aussi des questions portant sur la façon dont la collecte de nos données serait réalisée. Avions-nous les moyens de constituer et d'entretenir notre propre équipe d'intervieweurs? Si nous avions recours à la sous-traitance pour effectuer ce travail, comment pouvions-nous assurer la qualité du travail? La question de la qualité est devenue critique lorsque nous avons choisi un organisme qui connaissait mieux les études de marché que la recherche universitaire pour effectuer le travail sur le terrain. Nous avons donc inclus, dans le contrat, une série de procédures propres à l'enquête afin d'assurer la qualité. Plusieurs de ces procédures sont d'usage courant dans le domaine des enquêtes, comme l'examen détaillé des intervieweurs, le fait de prévoir une formation particulière dans les domaines où l'on pense que les intervieweurs pourraient avoir des points faibles, l'utilisation d'interviews fictives pré-établies lors des séances de formation, le fait d'accompagner les intervieweurs quand ils commencent à réaliser les interviews, l'obtention, deux fois par semaine, de renseignements sur les progrès réalisés par les intervieweurs, des procédures de conversion des refus, des réinterviews par la poste et par téléphone auprès d'un échantillon de répondants, etc. D'autres étaient plus innovatrices comme l'utilisation de vidéos spéciaux pour la formation (voir Smith 1992) et le fait de montrer à tous les intervieweurs comment adapter leur présentation initiale pour qu'elle réponde aux circonstances spéciales propres au répondant (voir Groves et Cialdini 1991).

3. QUESTIONS MÉTHODOLOGIQUES COURANTES

Le deuxième cycle de travaux sur le terrain a amené de nouvelles questions portant sur la réalisation et le contrôle de la qualité. De plus, nous désirions établir la recherche méthodologique pour elle-même.

3.1 Questions portant sur la réalisation

Règles de suivi

Pour la conception des études par panel classiques, il faut de bonnes règles de suivi afin de préciser quelles unités doivent être conservées lors de chaque cycle, quelles nouvelles unités doivent être ajoutées et quelles unités doivent être éliminées (Duncan 1992). En termes simples, nous voulions suivre les membres de l'échantillon original dans le temps même s'ils allaient former de nouveaux ménages et nous désirions ajouter de nouveaux membres à l'échantillon à la suite des naissances et des mariages, afin que l'échantillon conserve une certaine représentativité. Toutefois, la question se complique rapidement. Citons comme exemple le cas où le statut de membre «dans l'échantillon» et «non dans l'échantillon» est recoupé avec le fait que la personne a été un répondant ou un non-répondant. Par exemple, remontons-nous à un ménage où les membres qui ne faisaient pas partie de l'échantillon ont collaboré, mais où les membres de l'échantillon original ne l'ont pas fait? Il y a aussi une limite pratique quant aux membres de l'échantillon original qui peuvent être suivis. Nos règles pour le deuxième cycle n'excluaient que les membres de l'échantillon original qui avaient déménagé à l'extérieur du pays (à la grande déception de nos intervieweurs trop consciencieux) ou qui ont été envoyés dans des prisons ou admis dans des hôpitaux pour malades mentaux (les autres établissements institutionnels étaient inclus).

Il y a aussi la question du «ménage par opposition à la famille». Des économies peuvent être réalisées si on limite les interviews à une unité «familiale» à partir du deuxième cycle puisque les rôles strictement économiques, comme ceux de pensionnaire ou de chambreur présentent peu d'intérêt pour la recherche. Pour ce faire, il faudrait établir une définition de la «famille» qui est cohérente et utilisable sur le terrain. Toutefois, une telle définition semble poser des problèmes en elle-même, car le critère pour l'inclusion dans une famille

devrait être la nature des relations courantes ainsi qu'une évaluation de leur permanence vraisemblable. Pour le deuxième cycle, nous avons employé la définition normalisée d'un ménage afin de déterminer qui doit être interviewé à une adresse où l'on trouve des membres de l'échantillon original. Le concept de «famille» pourrait toutefois entrer en jeu dans les règles de suivi à partir du troisième cycle, puisque certaines catégories de membres qui ne font partie ni de l'échantillon ni de la famille et qui sont partis pour former des ménages distincts pourront ne plus être suivis.

Conception de la page couverture

On a rencontré plusieurs problèmes imprévus lors de la conception d'une page couverture pour le deuxième cycle. Le premier problème s'est posé parce que l'enquête était réalisée par interview sur place: essentiellement, comment fournit-on physiquement les renseignements dont l'intervieweur qui réalisera l'interview cette année aura besoin, comme les adresses et les renseignements démographiques de base? Ces renseignements devraient-ils figurer sur un document distinct, être imprimés sur des étiquettes qui seront apposées sur la page couverture ou imprimés directement sur la page couverture elle-même? Un deuxième problème était le fait de déterminer si l'on avait besoin de pages couvertures au niveau des particuliers ou des ménages. Toutefois, le concept d'un ménage longitudinal est futile, puisqu'à partir du deuxième cycle, l'échantillon est essentiellement un échantillon de personnes. Par contre, les intervieweurs s'occupent encore de ménages. L'utilisation de pages couvertures au niveau des particuliers aurait entraîné une prolifération de documents pour les intervieweurs et les procédures nécessaires pour s'occuper des nouveaux entrants seraient devenues compliquées. Éventuellement, nous avons adopté une page couverture au niveau des ménages. Un troisième problème portait sur la façon de désigner les membres «prévus» de chaque ménage. Cela n'était pas toujours aussi simple qu'on le penserait. Par exemple, si une carte de confirmation d'adresse nous permettait de découvrir qu'un couple s'est séparé, nous supposions alors que les enfants étaient allés avec leur mère et nous les inscrivions comme membres prévus du ménage de leur mère. Toutefois, si l'intervieweur constatait qu'ils faisaient partie du ménage de leur père, ils lui sembleraient de «nouveaux entrants» parce qu'ils n'auraient pas déjà été inscrits. Sans un mécanisme spécial pour traiter de tels cas, des erreurs importantes pourraient être créées au niveau de l'appariement des données.

Rappel de renseignements déjà fournis (interview avec rétro-information)

Le fait de simplement poser les questions à nouveau à chaque occasion peut entraîner une surestimation du changement véritable. Prenons, par exemple, le changement illusoire bien connu relevé, dans le cadre de la CPS, à propos des flux bruts de données sur les professions et les branches d'activité (Collins 1975). Il y a aussi les problèmes de frontière entre les cycles embarrassants rencontrés avec les panels de la SIPP et de l'enquête PSID (Moore et coll. 1992; Duncan et Mathiowetz 1985). Il s'agit d'une tendance qu'ont les répondants à *surdéclarer* les changements dans l'état et dans les montants reçus entre des mois civils adjacents inclus dans les périodes de référence correspondant à des interviews différentes, et à *sous-déclarer* les changements entre des mois qui font partie de la période de référence pour une seule interview, c.-à-d. une tendance qu'ont les changements à se grouper à la frontière entre les cycles. Une procédure pour minimiser ces difficultés consiste à rappeler à un répondant la réponse fournie lors du cycle précédent puis à lui demander s'il y a eu des changements.

Le rappel de renseignements présente de nombreux avantages, y compris le fait de réduire l'erreur de mesure et le changement factice, d'alléger le fardeau du répondant et de l'intervieweur, de diminuer les coûts suite à une réduction du temps d'interview et du codage, le fait de fournir une période de délimitation pour le rappel d'événements récents, l'aide pour stimuler la mémoire et d'inciter les répondants à se remémorer avec plus de soin les événements passés. Par contre, cette façon d'agir comporte le danger de réduire les estimations du changement véritable (il est facile de dire que rien n'a changé); de donner l'illusion d'un non-respect de la confidentialité; d'une collaboration négative; et il y a aussi des questions de coûts et de complexité. Une discussion complète des questions et méthodes relatives au rappel de renseignements déjà fournis est présentée dans Corti et Campanelli (1992).

Pour le deuxième cycle de l'enquête BHPS, nous avons décidé de fournir une quantité limitée de renseignements déjà déclarés. Ces informations comprenaient des renseignements de base pour communiquer de nouveau avec un répondant, comme l'adresse de ce dernier; les feuilles de visite pour les ménages où il a fallu effectuer de nombreuses visites de rappel; et des listes de dénombrement avec des renseignements démographiques. Nous avons aussi inclus une question spéciale de vérification pour préciser si les renseignements fournis l'année

précédente étaient exacts ou non afin d'aider à expliquer toute incohérence entre deux années. Le coût et la complexité de la conception d'un questionnaire papier et crayon qui nous aurait permis de rappeler des renseignements aux membres de l'échantillon original ainsi que d'inclure de nouveaux entrants nous ont dissuadé de rappeler des renseignements relatifs aux sujets visés par l'enquête jusqu'à ce que nous passions à l'IPAO.

Il y a toutefois un chevauchement intégré dans les périodes de référence de l'enquête BHPS. Par exemple, la période de travail sur le terrain tant pour le premier cycle que pour le deuxième dure essentiellement quatre mois (de septembre à décembre) et, dans chaque cas, on demande aux répondants s'il s'est produit des changements depuis septembre de l'année précédente. Ainsi, pour un répondant qui était interviewé en octobre les deux années, les renseignements de septembre sont disponibles avec une période de rappel de 13 mois et répétés avec une période de rappel d'un mois. Ce chevauchement peut être utilisé pour démêler des déclarations en double. De même, l'enquête BHPS comprend une vérification intégrée portant sur la profession. Bien que l'on demande aux répondants leur profession lors de chaque cycle, on leur demande aussi quand ils ont commencé à travailler dans le poste qu'ils occupent au moment de l'enquête. Nous pouvons donc comparer leur conception de l'évolution ou de la non-évolution de leur emploi avec les différences dans les descriptions de leur profession, en éliminant la variation attribuable à des interprétations du codage.

Codage des professions (essai du CASOC)

Le CASOC ou Computer-Assisted Standard Occupational Coding (codage normalisé des professions assisté par ordinateur) a été élaboré récemment au R.-U. dans le cadre d'une collaboration des universités de Warwick et de Cambridge. Le logiciel CASOC peut être utilisé pour effectuer le codage assisté par ordinateur et le codage automatisé et plusieurs fichiers utilitaires qui y sont associés permettent de recoder automatiquement les sorties détaillées en fonction d'une gamme étendue de classifications professionnelles et sociales. Nous avons réalisé notre propre expérience avec le CASOC afin de comparer le codage assisté par ordinateur au codage manuel des professions au niveau 3 chiffres tant pour ce qui est de la fiabilité que de la validité. Un échantillon aléatoire de 325 descriptions de la profession des répondants a été recodé indépendamment huit fois: le travail a été effectué quatre fois à la main par des codeurs possédant différents niveaux d'expérience, trois fois par les mêmes codeurs à l'aide du logiciel CASOC en mode assisté par ordinateur et une fois par un codeur qui a utilisé le logiciel en mode entièrement automatisé. La dernière addition est une colonne pour un «codeur expert» qui sera utilisée comme modèle de la «vérité». L'analyse des données n'est pas terminée, mais certains résultats préliminaires laissent supposer une fiabilité assez élevée avec des taux de concordance moyens de 0,79 parmi les codeurs manuels comparativement à 0,82 en mode assisté par ordinateur.

Questions relatives à la conception de la base de données, à la documentation et à la diffusion

D'autres questions importantes ont trait au meilleur genre de conception de base de données, pour des données longitudinales, qui satisfera une liste de besoins contradictoires comme la facilité d'accès pour le traitement, la facilité de manipulation des données pour des genres particuliers d'analyses, l'efficacité du stockage, etc. Comme pour plusieurs autres études par panel portant sur des ménages, nous avons choisi le logiciel de base de données relationnelles Scientific Information Retrieval (SIR).

Les usages relatifs à la documentation et à la diffusion peuvent aussi soulever des questions méthodologiques. On ne doit pas oublier la nécessité d'élaborer des façons d'aider les autres à utiliser et à exploiter les données recueillies au moyen d'un panel. Cela suppose un besoin de méthodes, qui ont fait l'objet de recherches approfondies, en matière de documentation et de diffusion ainsi que de bonnes politiques de formation, particulièrement pour les jeunes chercheurs, puisque les ensembles de données considérables et complexes découragent souvent les utilisateurs à cause de leur complexité tant intrinsèque que technique. Dans le cadre de ce processus, le Centre a élaboré une banque de données sur les questions employées précédemment, des interfaces informatiques pour les utilisateurs, a établi une série de séminaires de formation (conjointement avec deux autres universités) et créé sa propre *Research Resources Unit* qui joue un rôle central dans la gestion de l'information au sein du Centre et qui fournit divers services au personnel de ce dernier ainsi qu'à l'ensemble des chercheurs.

3.2 Contrôle de la qualité et sources d'erreurs

Le souci principal pour toute enquête doit être la qualité des données à toutes les étapes du processus d'enquête. Un souci plus général est celui d'étudier, quand cela est possible, l'importance de divers genres d'erreurs d'échantillonnage et non dues à l'échantillonnage afin que les utilisateurs ultérieurs puissent bénéficier des résultats de ces recherches.

En plus de surveiller l'organisme qui effectue nos travaux à contrat, le Research Centre a participé à des projets réalisés avant et après l'enquête. Par exemple, en plus des techniques standard d'essais préalables, certains des essais préalables faisaient appel à des méthodes spéciales comme le codage des interactions entre l'intervieweur et le répondant (Cannell et coll. 1989), des études de compte rendu auprès des intervieweurs et sur le terrain auprès des répondants (Campanelli, Martin et Rothgeb 1991) et des études avec des groupes fractionnés.

On a aussi consacré beaucoup d'efforts pour nous assurer que la collecte des données sur le revenu nous permettait d'obtenir les résultats les plus exacts. On a demandé aux répondants de consulter des documents quand cela pouvait se faire. Les codes fiscaux ont été recueillis quand les fiches de paie étaient disponibles afin de vérifier l'exactitude des données déclarées par les répondants. De même, nous avions des soucis particuliers à propos de la mise en application et de l'évaluation des procédures de mise à jour des panels, de suivi ainsi que de dépistage, et un programme informatique spécial a été écrit pour aider à la réalisation de cette tâche.

Une fois la collecte des données du premier cycle terminée, nous avons aussi participé à plusieurs étapes courantes, bien que pas toujours simples, y compris des vérifications de la qualité des processus de contrôle et de codage, l'épuration des données, l'élaboration de variables calculées, des problèmes de pondération et d'imputation ainsi que l'analyse de la non-réponse. Dans le cadre de la collecte des données pour le deuxième cycle, nous tentons de convertir les refus dans les ménages et de trouver et d'interviewer les personnes avec qui nous n'avons pu entrer en communication lors du premier cycle. Dans le cadre de l'analyse pour le deuxième cycle, nous étudierons les questions relatives à l'attrition des panels.

Notre plan visant à mesurer la **variance attribuable aux intervieweurs** constitue un projet important dans le cadre de nos efforts pour déterminer les sources d'erreurs. Un tel travail exige la randomisation de l'affectation des interviewés aux intervieweurs. C'est Mahalonobis (1946) qui a été l'un des premiers à utiliser cette méthode sous le nom d'«échantillon enchevêtré». À cause d'exigences relatives au travail sur le terrain et des coûts de déplacement, nous avons adopté une forme de randomisation sous contrainte dans laquelle les adresses ont été attribuées aux intervieweurs de façon aléatoire dans des ensembles géographiques. Par exemple, toutes les paires d'UPÉ distantes de moins de dix kilomètres ont été réunies en grappes d'une manière unique. Dans une grappe donnée, les adresses ont été attribuées de façon aléatoire aux intervieweurs. Nous avons l'intention de recueillir des renseignements sur l'importance de la variabilité entre les intervieweurs et d'établir un rapport entre cette variabilité et les caractéristiques des questions posées, les caractéristiques des intervieweurs eux-mêmes et, ce qui est le plus important, avec certains indicateurs de la mesure de l'erreur qui seront disponibles pour les données tels que les divergences dues à la mémoire. Comme ce travail a été réalisé dans le cadre du deuxième cycle, une composante implicite de l'essai sera l'analyse de l'effet dû à l'envoi du même intervieweur ou d'un intervieweur différent chez un répondant lors de la deuxième année d'une étude par panel. Nous attendons avec impatience l'occasion d'analyser ces données le printemps prochain au moyen de modèles de composantes de la variance à l'aide d'un nouveau logiciel hiérarchique appelé ML3, disponible au R.-U. (Goldstein 1991).

3.3 Recherche en méthodologie

Les projets traités dans la présente section représentent ceux qui, nous l'espérons, permettront de faire progresser notre connaissance des avantages et des limites des données obtenues dans le cadre d'enquêtes par panel et ils découlent surtout des points d'intérêt de nos chercheurs.

Contamination des réponses

La présence de tiers a-t-elle une incidence sur les réponses à des interviews réalisées dans le cadre d'enquêtes? Par exemple, se pourrait-il qu'une épouse ne veuille pas parler de ses anciens petits amis et fournir des données exactes si son mari est présent dans la pièce? Ce genre de contamination des réponses est une préoccupation

importante pour toutes les enquêtes, mais il est particulièrement pertinent dans le cas d'études menées auprès des ménages, où il arrive souvent que d'autres membres du ménage soient présents quand les répondants sont interviewés. De même, on pourrait aussi imaginer la situation d'une erreur de réponse qui varie dans le temps puisque les tiers qui sont présents peuvent changer d'un cycle à l'autre. Le questionnaire de l'enquête BHPS permet d'effectuer ce genre de recherche car on trouve, à la fin de chaque section du questionnaire, une question réservée à l'intervieweur qui permet à ce dernier d'indiquer qui était présent. Dans l'analyse de ces données provenant du premier cycle, Corti et Clissold (1992) ont trouvé certaines données qui laissent supposer que la présence d'autres personnes lors d'une interview peut effectivement avoir une incidence sur les réponses du répondant, particulièrement dans le cas de questions de nature délicate.

Calendriers et appel à la mémoire

Comme nous l'avons déjà mentionné, le plan pour recueillir des mesures continues du changement exige qu'un bon nombre des mesures recueillies dans le cadre de l'enquête portent sur des descriptions rétrospectives du changement au cours de la dernière année, plutôt que sur la situation qui existe au moment de l'interview. La question de concevoir un questionnaire afin de saisir ce niveau de détail des données avec un degré de qualité acceptable a causé du souci et entraîné des discussions. Pendant les deux années de planification intense en prévision de l'enquête, un certain nombre d'enquêtes pilotes et d'essais préliminaires ont été réalisés pour faire l'essai de diverses façons de recueillir ce genre de données faisant appel à la mémoire, y compris le fait de structurer le sens de l'appel à la mémoire de différentes façons, c'est-à-dire du passé au présent ou du présent au passé, l'utilisation de diverses aides visuelles visant à rafraîchir la mémoire, l'intégration de la collecte des données avec d'autres événements de la vie ainsi que le fait de saisir des détails du changement à l'aide de rappels libres mais chronologiques ou de l'emploi d'un carnet où seraient consignés les renseignements mois par mois ou semaine par semaine, comme ceux utilisés dans le cadre d'autres enquêtes par panel européennes (voir Corti 1992).

Méthodes analytiques afin de tenir compte de la complexité du plan d'enquête

Les données obtenues au moyen de plans d'enquête par sondage complexes ont traditionnellement été analysées à l'aide de diverses méthodes (comme le développement en série de Taylor, la méthode BRR, les répétitions selon la méthode du jacknife) pour tenir compte des effets du plan sur les erreurs-types (voir, par exemple, Kish et Frankel 1974). Récemment, on a proposé des modèles à plusieurs niveaux comme autre cadre pour effectuer l'analyse de ces données (Goldstein 1991). Les modèles à plusieurs niveaux présentent l'avantage de permettre l'étude explicite des éléments essentiels de la structure hiérarchique des données-échantillons et ils peuvent fournir des estimations plus efficaces que les méthodes traditionnelles. Taylor et Campanelli (1992) ont entrepris une comparaison empirique de ces deux méthodes pour des modèles de régression normale et logistique. Le but est d'évaluer les avantages et les inconvénients possibles liés à l'emploi de modèles à plusieurs niveaux plutôt que de techniques traditionnelles d'estimation de la variance. Un certain nombre de modèles dans lesquels on retrouve une gamme d'effets du plan de sondage pour la variable dépendante ainsi qu'une gamme d'effets du plan de sondage pour les variables indépendantes ont été étudiés. Les données préliminaires laissent supposer que les méthodes traditionnelles et à plusieurs niveaux donnent des estimations des résultats semblables tant pour les paramètres que pour les conclusions liées aux modèles.

Interviews qualitatives

L'interview elle-même est un domaine qui peut être étudié en soi. Un des principaux soucis du Centre est de lier l'enquête à d'autres méthodes de collecte de données, comme des interviews qualitatives plus en profondeur. Un de nos associés de recherche entreprend actuellement une étude qualitative de la retraite portant sur des membres de notre enquête-pilote par panel. Cette étude sera donc très utile comme essai de la validité des mesures provenant de notre enquête comparativement à celles qui découlent d'une méthode basée sur les cycles de vie. Dans une veine semblable, le projet des systèmes de répartition des ressources des ménages (voir, par exemple, Laurie 1992; Laurie et Sullivan 1991) combine une méthode qualitative avec les possibilités plus limitées mais plus représentatives de l'étude par panel dans ce domaine.

3.4 Recherche méthodologique portant sur les sujets visés par l'enquête

Une autre stratégie relative au travail portant sur l'enquête BHPS a consisté à explorer des problèmes méthodologiques en effectuant de la recherche sur les sujets visés par l'enquête. Deux exemples de ce genre de travail résultent de la participation de membres de notre équipe à la International Conference on Social Science Methodology qui s'est tenue récemment à Trente en Italie. Buck et Scott (1992), par exemple, ont découvert certaines questions méthodologiques importantes quand ils ont utilisé des analyses des histoires des événements. Leur but était de modéliser les jeunes qui quittent le domicile familial. Ils ont découvert que cela peut constituer un concept ambigu. Ils ont constaté que de légères variations dans la définition de leur variable dépendante menaient à des différences non négligeables dans leurs résultats à propos des sujets visés par l'étude. Un autre exemple vient du travail de Dex et Laurie (1992) qui ont étudié diverses questions méthodologiques rencontrées lors de leur analyse transnationale comparative du comportement des femmes sur le marché du travail.

4. QUESTIONS MÉTHODOLOGIQUES FUTURES

Plusieurs nouveaux projets sont prévus et actuellement à l'étude:

Test de la faisabilité de l'IPAO

Dans l'avenir, nous étudierons l'utilisation de la méthode d'interview sur place assistée par ordinateur (IPAO). L'IPAO est maintenant utilisée ou à l'essai pour de nombreuses études partout dans le monde, y compris certaines études très complexes (voir Costigan et Thompson 1992). Non seulement devrons-nous étudier des questions et des difficultés liées à la mise en application, il y a des questions de comparabilité des données. Dans quelle mesure les données recueillies dans le cadre d'une enquête assistée par ordinateur sont-elles comparables aux données recueillies à l'aide de méthodes plus traditionnelles? Un passage à l'IPAO au milieu d'une enquête par panel peut nous faire confondre le changement véritable avec un effet de mode possible (Olsen 1992). Par exemple, les répondants pourraient être plus ouverts selon un mode d'interview (ce qui se traduirait par une différence dans les moyennes au fil des ans) ou les données déclarées pourraient être plus exactes (ce qui pourrait mener à une réduction dans la variance).

La fiabilité des données rétrospectives

Les résultats de la recherche portant sur la qualité et la fiabilité des données faisant appel à la mémoire sont disponibles (voir l'étude de Dex 1992), mais pour les études par panel européennes, ils sont extrêmement limités, en partie parce que bon nombre des études par panel nationales menées auprès des ménages n'ont commencé que depuis le début ou le milieu des années 80. Pour l'enquête BHPS, nous espérons améliorer cette situation. Les données offrent deux genres différents d'occasions d'estimer l'importance de l'erreur de mémoire: 1) on demande des renseignements sur des événements du passé lointain dans deux enquêtes réalisées lors d'années différentes et 2) comme on l'a décrit ci-dessus, il y a un chevauchement entre les cycles de l'enquête causé par la durée de la période de travail sur le terrain. Cela offre certaines occasions intéressantes de faire des comparaisons. Nous pouvons voir si les écarts dans les déclarations varient selon le genre de question, le sujet, les difficultés de la tâche de déclaration, la pertinence probable des événements en question et diverses mesures des caractéristiques des répondants. Nous espérons aussi établir l'effet de telles erreurs sur les modèles réels et lier ce travail avec un projet semblable proposé par d'autres chercheurs britanniques, à l'aide de données basées sur des rappels de mémoire portant sur une période de 10 ans.

L'initiative d'analyse de grands ensembles de données considérables et complexes («Analysis of Large and Complex Datasets»)

Le Economic and Social Research Council du R.-U. a récemment reçu un appel de proposition pour un programme de recherche sur l'analyse de grands ensembles de données complexes. Le Centre espère participer à des projets réalisés en collaboration dans le cadre de ce programme afin d'élaborer des lignes directrices pratiques à l'intention des spécialistes en sciences sociales qui désirent analyser des données d'enquête complexes, des façons d'incorporer des «indicateurs» d'imputation dans l'analyse des données et accroître les possibilités de

modèles de survie en temps discret pour englober des effets aléatoires à plusieurs niveaux. D'autres projets considérés comprennent l'utilisation d'une nouvelle technologie pour le codage de questions ouvertes autres que celles portant sur la profession et pour une reconceptualisation radicale de la diffusion des données.

5. CONCLUSIONS

Plusieurs genres importants de projets n'ont pas été mentionnés jusqu'ici. Pour l'enquête BHPS, la priorité sera accordée à une étude de validation afin d'examiner l'exactitude des données autodéclarées par les répondants et ensuite à du travail additionnel centré sur la non-réponse. De plus, il pourrait être intéressant d'ajouter un nouvel échantillon transversal afin d'examiner les effets de conditionnement d'un panel dans le contexte de l'enquête BHPS. Il y a d'autres questions qui ne s'appliquent pas seulement à l'enquête BHPS. Comme méthodologistes, nous voudrions voir des recherches portant sur les questions relatives à la conception des questionnaires dans un contexte longitudinal. Cela comprendrait le fait de déterminer quelle est la véritable valeur d'une «question longitudinale» et d'élaborer des lignes directrices pour les concepteurs et les utilisateurs relativement à ce qui doit être fait quand on découvre qu'une question du premier cycle ne donne pas les résultats escomptés ou a vu sa signification changer dans le temps. De même, il faut effectuer du travail de conception dans le domaine des bases de données pour des genres particuliers d'analyses (p. ex., des analyses sur la chronologie des événements). Nous voudrions aussi voir certains des outils plus récents disponibles pour la recherche portant sur des enquêtes comme les techniques d'interviews cognitives (Tanur et Fienberg 1992) et un codage détaillé des interactions entre les intervieweurs et les répondants (Cannell et coll. 1989) appliqués à certains des problèmes particuliers de la collecte de données longitudinales. Un précédent dans ce domaine est l'utilisation qu'ont fait Moore et ses collègues des interviews cognitives afin d'améliorer les procédures pour la SIPP (1992). Un autre exemple serait l'étude portant sur ce que les travaux de recherche cognitive en laboratoire et le codage du comportement pourraient nous apprendre à propos du fait que les répondants sont «conditionnés» ou non et dans l'affirmative, comment et quand cela se produit.

Nous espérons que cette communication vous a donné une idée d'une partie du travail qui est présentement réalisé ou prévu dans le cadre de l'enquête BHPS. Comme il s'agit d'une enquête par panel relativement nouvelle, nous n'aurions pu aller aussi loin que nous l'avons fait sans les connaissances uniques acquises à partir d'autres études. Nous voulons exprimer notre vive gratitude à toutes les personnes qui nous ont aidés ou qui nous ont donné des conseils et attendons avec impatience vos commentaires et suggestions.

Nous reconnaissons avec gratitude l'appui tant du Economic and Social Research Council (R.-U) que de la University of Essex. Le travail mentionné dans la présente communication fait partie du programme scientifique du ESRC Research Centre on Micro-Social Change de Grande-Bretagne.

BIBLIOGRAPHIE

Brynin, M. (1992). Relating people by computer. Dans *Survey and Statistical Computing*, (éds. A. Westlake, R. Banks, C. Payne et T. Orchard), Amsterdam: Hollande du Nord.

Buck, N., et Scott, J. (1992). *Modelling household dissolution: An event history analysis of young people leaving home*. Discussion présentée à l'International Conference on Social Science Methodology, Trento, Italy, June. Colchester: ESRC Research Centre on Micro-Social Change.

Campanelli, P., Martin, E., et Rothgeb, J. (1991). The use of respondent and interviewer debriefing studies as a way to study response error in survey data. *The Statistician*, 40, 253-264.

Cannell, C.F., Oksenberg, L., Kalton, G., Bischoping, K., et Fowler, F.J. (1989). *New Techniques for Pre-testing Survey Questions*, Research Report, Ann Arbor MI: Survey Research Centre, Institute for Social Research.

Collins, C. (1975). Comparison of month-to-month changes in industry and occupation codes with respondent's report of change: CPS job mobility study. *Response Research Staff Report No. 75-5*, Washington, DC: U.S. Bureau of the Census.

Corti, L. (1992). Calendar and life history recall aids. *Discussion Paper 2*, Colchester: ESRC Research Centre on Micro-Social Change.

Corti, L., et Campanelli, P. (1992). The utility of feeding forward earlier wave data for panel studies. In *Survey and Statistical Computing*, (éds. A. Westlake, R. Banks, C. Payne et T. Orchard), Amsterdam: Hollande du Nord.

Corti, L., et Clissold, K. (1992). Response contamination by third parties in a household interview survey. *Working Paper 13*, Colchester: ESRC Research Centre on Micro-Social Change.

Costigan, P., et Thomson, K. (1992). Issues in the design of CAPI questionnaires for complex surveys. In *Survey and Statistical Computing*, (éds. A. Westlake, R. Banks, C. Payne et T. Orchard), Amsterdam: North-Holland.

Coxon, A.P.M. (éd.), (1992). Sample design issues in a panel survey: The case of the British household panel study. *Working Paper 3*, Colchester: ESRC Research Centre on Micro-Social Change.

Department for Statistics of Income and Consumption, (1984). *Notes on the Planning and the Scope of the Socio-Economic Panel Survey*, Netherlands: Central Bureau of Statistics.

Dex, S. (1992). The reliability of recall data: A literature review. *Working Paper 11*, Colchester: ESRC Research Centre on Micro-Social Change.

Dex, S., et Laurie, H. (1992). Comparative analyses using large scale national data sources of women's employment. *ESF Working Paper 37*, Colchester: ESRC Research Centre on Micro-Social Change.

Duncan, G. (1992). Household panel studies: Prospects and problems. Discussion présentée au International Conference on Social Science Methodology, Trento, Italie, juin.

Duncan, G., et Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 1, 97-117.

Duncan, G., et Mathiowetz, N.A. (1985). A validation study of economic survey data. *Mimeo*, Ann Arbor: Survey Research Centre, Institute for Social Research.

Goldstein, H. (1991). Multilevel modelling of survey data. *The Statistician*, 40, 235-244.

Groves, R., et Cialdini, R. (1991). Toward a useful theory of survey participation. Dans *Proceedings of the Section of Survey Research Methods*, Washington, DC: American Statistical Association.

Jean, A.C., et McArthur, E.K. (1987). Tracking persons over time. *SIPP Working Paper Series No. 8701*, Washington, DC: U.S. Bureau of the Census.

Kalton, G. (1992). Panel surveys: Adding the fourth dimension. Dans *Recueil du symposium 92: Conception et analyse des enquêtes longitudinales*, Ottawa, Statistique Canada, 115-126.

Kish, L., et Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, 1, 1-37.

Laurie, H. (1992). Multiple methods in the study of household resource allocation. Dans *Mixing Methods: Qualitative and Quantitative Research*, (éd. J. Brannen), Aldershot: Avebury.

Laurie, H., et Sullivan, O. (1991). Qualitative versus quantitative? The use of qualitative information in the British household panel study. *Sociological Review*, 39, 1, 113-130.

Mahalonobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Moore, J., Bogen, K., et Marquis, K. (1992). A cognitive interviewing approach for the survey of income and program participation: development of procedures and initial test results. Dans *Recueil du symposium 92: Conception et analyse des enquêtes longitudinales*, Ottawa, Statistique Canada, 35-47.

Morgan J.N., et Duncan, G.J. (1986). Experience with the panel study of income dynamics. Dans *Microanalytic Simulation Models to Support Social and Financial Policy*, (éds. G.H. Orcutt, J. Merz et H. Quinke), Holland: Elsevier Science Publishers.

Olsen, R.J. (1992). The effects of computer assisted interviewing on data quality. *Working Paper 36*, Colchester: ESRC Research Centre on Micro-Social Change.

Rose, D. et coll. (1991). Micro-social change in Britain: An outline of the role and objectives of the British household panel study. *Working Paper 1*, Colchester: ESRC Research Centre on Micro-Social Change.

Rose, D., Buck, N., et Corti, L. (1991). Design issues in the British household panel study. *Bulletin de Methodologie Sociologique*, 32, 14-43.

Rose, D., Campanelli, P., Corti, L., et Taylor, M. (1992). Methodology for household panels and longitudinal data analysis: Where are we and where do we go from here? *Discussion Paper 1*, Colchester: ESRC Research Centre on Micro-Social Change.

Smith, R. (1992). *Audio-visual aids in interviewer training.* Discussion présentée à la 47th Annual Conference of the American Association for Public Opinion Research, St. Petersburg, Floride, Mai. Colchester: ESRC Research Centre on Micro-social Change.

Tanur, J., et Fienberg, S. (1992). Cognitive aspects of surveys: Yesterday, today, and tomorrow. *Journal of Official Statistics*, 8, 1, 5-17.

Taylor, A., et Campanelli, P. (1992). *Accounting for the complexity of the survey Design: A comparison of traditional design-based procedures to multi-level modelling techniques.* Discussion présentée au annual meeting of the Royal Statistical Society, Sheffield, Angleterre, Septembre.

Van de Pol, F. (1988). *Design issues in panel studies.* Amsterdam: Sociometric Research Foundation.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

UNE ÉTUDE LONGITUDINALE BASÉE SUR UNE ENQUÊTE PERMANENTE NATIONALE: L'ENQUÊTE «LONGITUDINAL STUDY OF AGING»

M.G. Kovar[1]

RÉSUMÉ

Un supplément à la National Health Interview Survey (enquête nationale sur la santé) (NHIS), une enquête permanente sur la population civile hors établissements institutionnels des États-Unis, a été utilisé afin d'obtenir les données de référence pour une étude longitudinale. Dans cette étude, l'enquête «Longitudinal Study of Aging» (LSOA), on a aussi tiré profit d'appariements avec des dossiers administratifs pour établir l'occurrence et la cause des décès ainsi que l'utilisation des soins médicaux dispensés aux malades hospitalisés. Dans la présente communication, on décrit l'étude et on examine certains des avantages et des inconvénients liés à l'utilisation d'une étude conçue à d'autres fins comme base d'une étude longitudinale.

MOTS-CLÉS: Vieillissement; longitudinale; NHIS; LSOA.

1. INTRODUCTION

De nombreux pays, y compris le Canada, réalisent des enquêtes-échantillons nationales transversales basées sur la population. De nombreux pays effectuent aussi des études longitudinales basées sur un échantillon de la population. De plus, tous les pays disposent de dossiers administratifs renfermant des données qui, du moins théoriquement, peuvent être couplées avec ces enquêtes et études. Toutefois, la coordination entre les études transversales et les études longitudinales est rarement bien faite et les dossiers administratifs ne sont pas utilisés aussi souvent qu'ils pourraient l'être.

Dans la présente communication, on décrit une enquête transversale nationale utilisée comme la base d'une étude longitudinale, dans le cadre de laquelle des données tirées de dossiers administratifs ont été couplées avec les données de l'étude.

L'étude est l'enquête LSOA, qui était basée sur le supplément «Supplement on Aging» (SOA) à l'enquête NHIS de 1984. Les enregistrements sur les participants à l'étude ont été couplés avec un fichier d'actes de décès et avec un fichier d'enregistrements sur les demandes liées aux soins médicaux. En autant que je sache, il s'agissait de la première étude longitudinale qui avait été conçue pour être basée sur une enquête transversale nationale permanente et aussi pour tirer profit d'appariements avec des dossiers administratifs afin d'ajouter des renseignements aux réponses recueillies dans le cadre de l'enquête.

Dans la présente communication, on décrit tout d'abord l'étude puis on examine certains des avantages et des inconvénients de cette méthode.

[1] M.G. Kovar, National Center for Health Statistics, Hyattsville (Maryland), É.-U. 20782.

2. L'ÉTUDE

2.1 L'enquête de référence

L'enquête LSOA était basée sur l'enquête NHIS, une enquête transversale permanente portant sur la population civile hors établissements institutionnels des États-Unis. L'enquête NHIS est composée d'un questionnaire principal, ou de base, qui est révisé environ une fois par décennie et de suppléments portant sur des sujets spéciaux qui changent d'une année à l'autre. En 1984, il y avait deux suppléments à l'enquête NHIS, un supplément sur l'assurance-maladie et le supplément SOA. Le questionnaire principal de l'enquête NHIS ainsi que les deux suppléments ont fourni les données de référence pour l'enquête LSOA.

Le supplément SOA a aussi fourni la base de sondage et il comprenait des questions qui ont été posées dans les interviews de l'enquête LSOA réalisées ultérieurement.

2.1.1 Le questionnaire de base de l'enquête NHIS

L'enquête NHIS est la grosse enquête permanente, réalisée par le National Center for Health Statistics (NCHS), qui est conçue pour obtenir des renseignements sur les maladies chroniques et aiguës dont souffre la population civile hors établissements institutionnels de tous les âges vivant dans des ménages aux États-Unis. Cette enquête receuille également des données sur les blessures subies par cette population et sur son utilisation des services de santé.

Des détails à propos de l'enquête, y compris le plan d'échantillonnage et les procédures utilisées en 1984, ont été publiés (Kovar et Poe 1985). Il y a, cependant, certaines caractéristiques de l'enquête NHIS qui n'étaient pas ce que nous aurions choisi si nous avions eu la haute main sur tout le processus.

L'enquête NHIS est une enquête nationale, basée sur un plan d'échantillonnage complexe, avec des interviews réalisées dans des ménages chaque semaine (les échantillons hebdomadaires sont des échantillons nationaux indépendants).

L'enquête est basée sur des personnes qui répondent pour leur famille. Bien que l'on invite tous les adultes qui sont dans le domicile à participer à l'interview, tout adulte peut répondre pour tous les autres membres de la famille.

L'enquête est conçue afin d'obtenir des estimations de la prévalence et non pour étudier des relations entre variables. C'est pourquoi on utilise des sous-échantillons quand une fraction de l'échantillon suffit pour estimer la prévalence. Des sous-échantillons sont utilisés, par exemple, pour fournir des estimations de la prévalence d'états chroniques. Il y a six listes d'états, chacune étant utilisée dans un sixième des ménages.

2.1.2 Le supplément sur l'assurance-maladie

Le supplément sur l'assurance-maladie renfermait des questions à propos de la protection offerte par l'assurance-maladie tant publique que privée à tous les membres de la famille vivant dans chaque ménage. Ce supplément avait été utilisé auparavant dans l'enquête NHIS afin de surveiller les changements dans la protection en matière d'assurance-maladie. Le répondant pour la famille était aussi le répondant pour ce supplément.

2.1.3 Le supplément «Supplement on Aging»

Le supplément «Supplement on Aging» était un nouveau supplément. Il s'agissait de la première enquête nationale, aux États-Unis, conçue spécifiquement pour obtenir des renseignements auprès d'un échantillon représentatif de personnes d'âge moyen et plus âgées vivant dans la collectivité.

On peut considérer le supplément SOA comme une enquête nationale indépendante réalisée auprès des Américains plus âgés et les données peuvent être analysées comme si elles provenaient d'une enquête indépendante. Toutefois, les données furent recueillies dans le cadre d'un supplément à une enquête permanente et les procédures ainsi que le questionnaire relatifs au supplément ont dû être intégrés à ceux qui étaient prévus

pour le questionnaire de base de l'enquête NHIS. Les procédures ainsi que le plan d'échantillonnage utilisés pour le supplément SOA ont aussi été publiés (Fitti et Kovar 1987).

Le taux de réponse pour le supplément SOA était élevé; 96.7 % des personnes âgées de 55 ans et plus dans les ménages visés par l'enquête NHIS ont participé au SOA, soit en personne, soit par personne interposée. Par conséquent, le taux de réponse effectif pour le supplément SOA était de 93.2 %.

2.2 Les interviews longitudinales

Les interviews longitudinales ne visaient que les participants au SOA âgés de 70 ans et plus. Ces interviews étaient conçues afin de mesurer le changement dans le statut fonctionnel et dans la situation des particuliers dans le ménage, y compris le fait d'entrer dans des maisons de soins infirmiers (et d'en sortir). Ces interviews étaient aussi conçues pour être réalisées régulièrement à des intervalles de deux ans. On s'est conformé, avec une consistance remarquable, aux deux parties de la conception mais, pour les deux parties, on a aussi dû céder à des pressions externes. Les détails, y compris tous les questionnaires, ont été publiés (Kovar, Chyba et Fitti 1992), mais je désirerais étudier certains de ces détails ici.

Deux changements importants ont été apportés entre l'enquête de référence et les interviews réalisées ultérieurement. L'un d'entre eux était le changement dans la taille de l'échantillon et l'autre le changement dans le mode d'interview.

Le changement dans la taille de l'échantillon constituait un changement important.

Des 7 527 participants au SOA, âgés de 70 ans ou plus, seulement 5 151 étaient dans l'échantillon qui devait être interviewé en 1986. À ce moment, on avait peu d'argent à consacrer à un concept nouveau tel qu'une étude longitudinale réalisée par téléphone auprès des personnes plus âgées. Heureusement, le National Institute on Aging était prêt à prendre une chance, mais la seule façon de réaliser l'étude consistait à recourir à un sous-échantillon. Nous avons conçu le sous-échantillon de façon à maximiser les possibilités analytiques en incluant chaque participant au SOA qui était âgé de 80 ans ou plus, chaque participant de race noire âgé de 70 ans ou plus ainsi que tous les membres de la famille des personnes faisant partie de ces deux groupes et la moitié des autres participants au SOA âgés de 70 ans ou plus.

Lors de la planification de l'interview de 1988, il y avait plus d'intérêt pour les études longitudinales portant sur les personnes plus âgées et l'on disposait de plus d'argent pour réaliser l'interview. Par conséquent, on a pu inclure tous les 7 527 participants au SOA âgés de 70 ans et plus, ce qui a ajouté 2 276 personnes aux échantillons de 1988 et de 1990. À cause de la conception de l'échantillon de 1986, les personnes qui se sont ajoutées à l'échantillon étaient toutes âgées de 70 à 79 ans et ne vivaient pas avec une personne âgée de 80 ans ou plus ou avec une personne de race noire âgée de 70 ans ou plus.

Le changement dans la taille de l'échantillon a créé des problèmes analytiques. De plus, les taux de réponse des personnes ajoutées en 1988 étaient plus faibles que ceux des personnes qui faisaient partie de l'échantillon de 1986.

Le deuxième changement important avait trait au mode d'interview.

Les interviews de référence ont été réalisées sur place. Les interviews de suivi étaient des interviews téléphoniques avec envoi de questionnaires par la poste aux personnes qui n'avaient pas le téléphone ou avec lesquelles on n'avait pu entrer en communication par téléphone.

Les interviews téléphoniques ont été réalisées à l'aide de la méthode d'interview téléphonique assistée par ordinateur (ITAO) à partir d'un centre téléphonique du Bureau of the Census des États-Unis. La décision d'utiliser l'ITAO plutôt que de réaliser des interviews sur place était basée sur les coûts, mais on l'a prise seulement après qu'une étude de faisabilité a démontré que l'interview par téléphone était une méthode pratique pour les personnes plus âgées (Fitti et Kovar 1985). Les intervieweurs qui ont réalisé les interviews téléphoniques étaient formés et ils avaient suivi une journée de formation pour cette enquête, en plus de participer à des interviews simulées et de pratique, avant de commencer à interviewer les participants à l'enquête

LSOA. Cependant, ils n'avaient pas les années de formation et d'expérience que possédaient un bon nombre des intervieweurs qui ont travaillé à l'étude de référence. Toutefois, contrairement aux personnes qui ont réalisé les interviews sur place, les personnes travaillant au téléphone ont eu une surveillance constante, parce que les surveillants pouvaient suivre le déroulement de toute interview en tout temps et corriger les erreurs immédiatement.

Il y a eu trois conséquences remarquables découlant de l'utilisation du téléphone : les interviews devaient être plus courtes que lorsque les intervieweurs étaient dans les maisons; les taux de réponse ont été plus faibles que pour les interviews réalisées sur place et nous étions moins en mesure de choisir qui répondait aux questions.

Le contenu des interviews était limité aux deux mesures qui nous intéressaient principalement (les changements dans le statut fonctionnel et les changements dans la situation des particuliers dans le ménage) pour que les interviews soient les plus brèves possible. Les interviews portaient sur les mesures principales nécessaires aux fins de l'enquête et obtenaient peu de renseignements sur les covariables sauf pour les changements dans les covariables de première importance.

Parce que l'accent principal était mis sur le changement dans le statut fonctionnel, les batteries de questions sur les activités de la vie quotidienne (AVQ) et sur les activités instrumentales de la vie de tous les jours (AIVTJ) ont été posées lors de chaque interview, sans modification. Il y avait aussi une question additionnelle à la fin des questions pour chaque AVQ et AIVTJ, laquelle visait à déterminer s'il s'agissait d'un changement par rapport à la dernière fois où nous avions parlé au répondant. Les questions de Nagi ont aussi été posées exactement comme elles l'avaient été lors de l'enquête de référence. De plus, il y avait des questions à propos des motifs du changement pour les deux problèmes les plus courants - la difficulté à marcher un quart de mille et la difficulté à monter dix marches sans se reposer.

Les questions sur la situation des particuliers dans le ménage ainsi que sur l'état matrimonial n'avaient pas à être répétées chaque fois pour déterminer le changement parce que de tels changements sont très évidents et soudains, contrairement aux changements subtils dans l'habileté fonctionnelle. Il y avait plutôt des questions simples sur le changement et sur la date des changements.

Ainsi, les questions sur les deux points d'intérêt critique de l'étude longitudinale ont été traitées différemment. On n'a pas tenté d'obtenir la date du changement dans le statut fonctionnel parce que de tels changements sont souvent graduels, mais on a consacré beaucoup d'efforts pour détecter l'occurrence d'un changement de ce genre. Par contre, on a consacré beaucoup d'efforts pour déterminer la date du changement dans l'état matrimonial ou dans la situation des particuliers dans le ménage, parce que de tels changements sont des discontinuités, mais on a eu besoin de peu d'efforts pour déterminer l'occurrence d'un tel changement.

Les taux de réponse pour l'enquête réalisée par téléphone et par la poste en 1986 sont demeurés supérieurs à 90 pour cent même si on a tenu compte des personnes que l'on savait décédées. Toutefois, les taux de réponse étaient inférieurs en 1988 et en 1990, en partie à cause de l'intervalle de quatre ans pour les personnes ajoutées en 1988. Les taux d'autodéclaration étaient plus faibles pour les trois interviews de suivi parce qu'il était plus probable que les femmes répondent au téléphone et qu'il était moins probable que les personnes puissent répondre pour elles-mêmes plus elles avançaient en âge.

Le plan d'échantillonnage prévoyait des interviews aux deux ans. Le 1er juillet 1986 survenait deux ans après que la moitié de la période de l'étude de référence était écoulée. Toutefois, certaines personnes âgées déménagent avec les saisons de sorte que le déroulement des interviews a été prévu d'août jusqu'à octobre afin d'augmenter la probabilité d'atteindre les personnes. Tous les questionnaires expédiés par la poste, ceux adressés aux personnes qui n'ont pas fourni de numéro de téléphone et ceux adressés aux personnes qui n'ont pu être atteintes par téléphone, ont été expédiés après la fin des interviews téléphoniques. Par conséquent, les réponses à l'interview de 1986 étaient concentrées dans la période allant d'août à octobre, mais on a continué de recevoir des réponses jusqu'à la fin de 1986.

Le calendrier pour 1988 était identique à celui de 1986. En 1990, cependant, le centre téléphonique du Bureau of the Census était requis pour des travaux relatifs au recensement en septembre et en octobre. Les interviews pour l'enquête LSOA devaient prendre fin au début de septembre, de sorte que la période d'interview a

commencé en juillet. Par conséquent, les réponses à l'interview de 1990 étaient concentrées en juillet et en août et l'on a reçu des réponses aux questionnaires expédiés par la poste jusqu'à la fin d'octobre. Par conséquent, il arrivait rarement que les interviews soient réalisées à exactement deux années d'intervalle.

2.3 Enregistrements appariés

2.3.1 Le National Death Index

Le National Death Index (NDI) est un fichier informatisé constitué à partir des actes de décès qui est tenu à jour par le National Center for Health Statistics (NCHS). Toute personne qui désire utiliser ce répertoire à des fins de recherche (il ne peut être utilisé à des fins de réglementation) doit présenter une demande à la Division of Vital Statistics, NCHS. Une fois la demande étudiée et approuvée, ses auteurs présentent un fichier d'enregistrements à apparier. C'est pour les fichiers renfermant les dix éléments suggérés dans le NDI Users Manual (NCHS 1981), ou la majorité de ces éléments, que l'appariement donne les meilleurs résultats.

Parce que l'on avait prévu l'appariement avec le NDI quand l'enquête de référence a été planifiée, on avait demandé aux participants la permission d'effectuer l'appariement et les renseignements sur les dix éléments. Presque tous les participants (15,938 sur 16,148 pour le supplément SOA et 7,426 sur 7,527 pour l'enquête LSOA) ont fourni les renseignements demandés.

L'appariement au NDI permet d'obtenir une liste de tous les appariements possibles en ordre décroissant de probabilité. C'est le directeur de l'étude qui doit décider si l'appariement est exact. L'algorithme utilisé dans l'enquête LSOA pour prendre cette décision est décrit dans Kovar, Chyba et Fitti (1992). Quatre catégories correspondant à la probabilité de l'appariement pour les décès de 1984 à 1989 sont sur la bande à grande diffusion.

2.3.2 Cause du décès

Le NCHS tient aussi à jour un fichier informatisé d'enregistrements avec causes multiples du décès qui comprend le numéro de l'acte de décès. Les agents de projet du NCHS qui désirent utiliser ce fichier doivent présenter une demande à la Division of Vital Statistics, NCHS. Une fois la demande étudiée et approuvée, ses auteurs présentent un fichier d'enregistrements à apparier[2].

Parce que les enregistrements du SOA avaient été couplés avec le NDI et que seulement les enregistrements pour lesquels nous étions certains ou presque qu'il y avait eu décès seraient présentés, la permission d'effectuer ce couplage a été accordée.

2.3.3 Enregistrements de Medicare (régime public d'assurance-maladie)

La Health Care Financing Administration (HCFA) tient à jour un fichier de toutes les factures pour les services payés par Medicare. Les enregistrements de l'enquête LSOA sont couplés avec ces enregistrements de Medicare seulement si le participant a donné sa permission et fourni un numéro qui pouvait être utilisé pour effectuer le couplage - un numéro de sécurité sociale (Social Security), de pension de retraite des sociétés de chemins de fer (Railroad Retirement), ou un numéro de demande présentée en vertu de l'assurance-santé (Health Insurance Claims). Seul le numéro est envoyé à la Health Care Financing Administration afin de préserver la confidentialité.

Les appariements avec les enregistrements de Medicare ont été plus compliqués que pour le NDI, parce qu'il était nécessaire, tout d'abord, d'effectuer un appariement avec le fichier des inscriptions afin de s'assurer que la personne était inscrite, puis de vérifier l'enregistrement des inscriptions par rapport à l'enregistrement de l'enquête LSOA afin de s'assurer que l'appariement numérique correspondait effectivement à la personne dans

---

[2] L'accès à ce fichier est limité aux personnes qui réalisent des enquêtes pour le compte du NCHS qui a ce privilège parce qu'il a un contrat avec chaque service d'enregistrement. Ces contrats prévoient l'acquisition des données nécessaires pour publier les statistiques nationales de l'état civil et afin de tenir le NDI à jour.

l'échantillon. C'est seulement après ces étapes que le fichier des enregistrements de l'enquête LSOA était présenté à la HCFA afin d'obtenir les enregistrements d'utilisation du service.

Les renseignements (y compris la date, les codes de diagnostic, les codes de l'acte et les frais), sont extraits des enregistrements sur les soins dispensés aux malades hospitalisés («Medicare Part A»). Il y a, sur la bande à grande diffusion de l'enquête LSOA, un enregistrement pour chaque hospitalisation dans ce fichier. Le fichier principal des personnes renferme un indicateur qui précise s'il y a aucun, ou au moins un, enregistrement pour cette personne dans ce fichier.

Aucun renseignement détaillé n'est extrait des enregistrements pour les autres services couverts par Medicare. Il y a, toutefois, un fichier sur la bande à grande diffusion de l'enquête LSOA avec un indicateur qui précise s'il y a ou non une facture pour le service particulier au cours de chaque année civile.

Il y avait 11 497 personnes dans l'échantillon du SOA qui étaient âgées de 65 ans ou plus; on a effectué un appariement pour les enregistrements correspondant à 10 442 de ces personnes y compris 6 920 dans l'échantillon, de l'enquête LSOA, des personnes âgées de 70 ans et plus.

3. AVANTAGES, PROBLÈMES ET SOLUTIONS

3.1 Avantages

L'utilisation d'une enquête permanente nationale afin d'obtenir des données de référence pour une enquête longitudinale présente l'avantage évident de réduire le coût. Puisque l'enquête NHIS aurait été réalisée de toute façon, il n'a pas été nécessaire de payer pour une enquête de sélection préliminaire afin de trouver les personnes âgées de 70 ans ou plus. De plus, le questionnaire de base de l'enquête NHIS a permis d'obtenir des renseignements démographiques, sociaux et sur la santé qui sont nécessaires pour toute étude de ce genre.

Les taux de réponse élevés obtenus dans le cadre de l'enquête NHIS ont fourni un autre avantage. Parce qu'il s'agit d'une enquête permanente, les intervieweurs de l'enquête NHIS travaillent rarement à une autre enquête, ils suivent souvent des cours de recyclage et connaissent très bien le questionnaire de base. Les taux de réponse élevés sont un résultat de cette situation; en 1984, des renseignements ont été obtenus de 96.4 pour cent des ménages choisis pour l'échantillon. Cela constituait un avantage réel découlant de l'utilisation de cette enquête. Nous n'aurions pu embaucher et former des intervieweurs pour une enquête unique et obtenir des taux de réponse aussi élevés.

Le couplage avec les fichiers du NDI présente deux avantages, particulièrement pour une population plus âgée où les taux de mortalité sont élevés. Cela a réduit la proportion des personnes qui auraient été considérées comme «perdues pour un suivi» si nous n'avions pas eu ce couplage. Nous avons aussi pu obtenir une date de décès précise, ce qui est utile pour des études de survie.

Le couplage avec les fichiers de Medicare a fourni des renseignements que les répondants dans le ménage auraient eu de la difficulté à déclarer d'une manière exacte. Les personnes âgées ont de la difficulté à préciser exactement la date des événements et peu de personnes, quel que soit leur âge, peuvent déclarer avec exactitude des diagnostics ou des actes médicaux.

3.2 Compromis

Le fait d'ajouter le supplément SOA à l'enquête NHIS a créé plusieurs problèmes de procédure. Nous voulions une auto-déclaration, nous voulions imposer des conditions pour tout le monde, nous voulions plus de détails pour certains éléments qui faisaient partie de l'enquête de référence et nous voulions poser des questions qui avaient été utilisées dans d'autres enquêtes.

La solution à l'intégration a consisté à poser les questions du SOA après que le questionnaire de base de l'enquête NHIS et le supplément sur l'assurance-maladie aient été terminés et d'utiliser les mêmes périodes pour

lesquelles les répondants devaient se remémorer certains faits que pour l'enquête NHIS, mais de changer les règles relatives aux répondants et d'ajouter une liste spéciale d'états chroniques.

Le fait de poser les questions du SOA après de nombreuses autres questions portant sur la santé peut avoir constitué un avantage si cela permettait de mieux se souvenir des faits, puisque les répondants avaient eu l'occasion de penser à leur santé. L'utilisation des mêmes périodes pour lesquelles les répondants devaient se remémorer certains faits que pour l'enquête NHIS (deux semaines, un an) nous a obligés à modifier cette période pour certaines questions adaptées à partir d'autres études (Fitti et Kovar 1987).

Le fait de changer la règle utilisée pour l'enquête NHIS, qui permettait à tout adulte de répondre pour tous les autres membres de la famille, en une règle d'autodéclaration a donné des résultats tellement bons que 90 pour cent des participants au SOA qui avaient 70 ans et plus (par opposition à 80 pour cent pour l'enquête NHIS) ont répondu pour eux-mêmes. Les personnes qui n'ont pas répondu pour elles-mêmes étaient habituellement très âgées ou en perte d'autonomie. Le fait d'ajouter une liste spéciale d'états chroniques élaborée à partir d'états dont la prévalence était élevée parmi les personnes âgées lors des réalisations antérieures de l'enquête NHIS signifiait que tout répondant devait indiquer «oui» ou «non» pour chaque état sur la liste.

3.3 Inconvénients

Parce que nous désirions plus de détails sur certains sujets qui faisaient partie du questionnaire de base de l'enquête NHIS, nous avons dû répéter certaines questions afin d'établir le contexte approprié pour les autres questions sur ce sujet. Par exemple, nous désirions des renseignements sur le rapport qu'ont les autres membres de la famille avec la personne âgée, ce qui signifie qu'il fallait retourner à un sujet déjà traité. Un autre exemple est le fait que nous voulions un échantillon complet pour une liste spéciale d'états chroniques dont nous savons que la prévalence est élevée chez les personnes âgées. Les questions à propos d'un sixième de ces états avaient déjà été posées dans chaque ménage.

Nous n'avons pu employer des «batteries» de questions ou les périodes pour lesquelles les répondants devaient se remémorer certains faits qui avaient été utilisées sur le questionnaire original pour les questions adaptées à partir d'autres études. C'est pour cela que certains critiques ont dit que les questions n'étaient pas calibrées.

L'enquête de référence a été réalisée au moyen d'interviews sur place. Les interviews de suivi ont été effectuées par téléphone. Le changement dans le mode d'interview peut avoir eu une incidence sur les estimations du changement dans le statut fonctionnel entre l'enquête de référence et le premier suivi. Peu de recherches ont été effectuées sur les effets du mode lorsqu'on interviewe des personnes plus âgées. Des recherches ont démontré, toutefois, que l'autodéclaration et la déclaration par personne interposée ne donnent pas les mêmes réponses, et une proportion plus élevée de personnes interposées ont fourni les réponses lors des interviews téléphoniques.

Le fait que l'enquête NHIS est réalisée pendant toute l'année signifiait que la première interview longitudinale ne pouvait être réalisée deux ans après l'interview initiale. Nous avons fait un compromis, mais cela a rendu l'analyse de survie plus difficile.

4. RÉSUMÉ

L'enquête LSOA a eu beaucoup de succès.

On a pu réaliser cette enquête à un coût relativement faible parce que les interviews initiales sur place ont évité d'avoir à réaliser une sélection dispendieuse pour trouver la population relativement rare des personnes âgées de 70 ans et plus et parce que ces interviews ont permis d'obtenir beaucoup de renseignements qui étaient nécessaires pour réaliser l'étude longitudinale. Le fait que nous ayons utilisé le téléphone pour effectuer l'étude de suivi, ce qui aurait été plus difficile si nous n'avions pas eu ce contact initial sur place, a aussi permis de limiter les coûts.

L'enquête de référence a été avantagée parce qu'elle a eu, comme personnel sur le terrain, des intervieweurs du Bureau of the Census des É.-U. qui étaient extrêmement bien formés à la façon d'obtenir des taux de réponse élevés et aux concepts de l'enquête.

Toutefois, la raison principale pour laquelle nous jugeons que l'enquête LSOA a été couronnée de succès est que cette étude a été conçue pour les chercheurs et que ces personnes l'ont utilisée. Elle a été largement employée pour des articles dans des journaux avec évaluation par les pairs, dans de nombreux mémoires et thèses ainsi que dans des documents de travail et elle a fourni des données nationales pour une politique de santé.

REMERCIEMENTS

La Longitudinal Study of Aging n'aurait pu être réalisée sans mes collègues du National Center for Health Statistics, Joseph Fitti et Michele Chyba, ou sans notre agent de projet au National Institute on Aging, Richard Suzman. Les premiers ont assumé la responsabilité pour des aspects importants de l'étude, tels que la réalisation des interviews et l'appariement des enregistrements. Le dernier a fourni un appui et des conseils pendant toute la durée de l'étude.

BIBLIOGRAPHIE

Fitti, J.E., et Kovar, M.G. (1987). The supplement on aging to the 1984 national health interview survey. *Vital and Health Statistics*, 1, 21.

Fitti, J.E., et Kovar, M.G. (1987). A multi-mode longitudinal study of aging *Proceedings of the American Statistical Association Section on Survey Research Methods.*

Kovar, M.G., et Poe, G. (1985). The national health interview survey design 1973-84, and procedures 1975-83. *Vital and Health Statistics*, 1, 18.

Kovar, M.G., Chyba, M.M., et Fitti, J.E. (1992). The longitudinal study of aging: 1984-1990. *Vital and Health Statistics*, 1, 28.

National Center for Health Statistics (1981). *Users Manual, The National Death Index*, Hyattsville, Maryland: Public Health Service.

SESSION 7

Applications générales II

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

UNE ENQUÊTE LONGITUDINALE ET LA VÉRIFICATION DE LA RÉALITÉ SUR LA VALEUR DES AVOIRS FINANCIERS

C.D. Cowan[1]

RÉSUMÉ

L'enquête «Estimated Cash Recovery Survey» (ECRS) présente toutes les caractéristiques d'une enquête longitudinale polyvalente classique. Elle sert à estimer le revenu espéré de la vente d'institutions financières en faillite qui ont été mises sous séquestre par le gouvernement des États-Unis.

L'enquête permet de déterminer des valeurs de liquidation pour 17 catégories d'actif, avec une stratification en fonction des régions et de la taille des institutions. La population des actifs visés change rapidement au fil des trimestres à mesure que de nouvelles institutions sont mises sous séquestre et que des actifs financiers sont vendus et retirés de la population. La valeur des actifs qui constituent l'échantillon change aussi au gré de la conjoncture économique. Tous ces facteurs - polyvalence de l'enquête, conjoncture économique variable et variation de la composition de la population étudiée - font de l'enquête ECRS un cas complexe mais intéressant au point de vue du plan de sondage et de l'analyse.

MOTS CLÉS: Liquidation d'actifs et recouvrement; institutions sous séquestre.

1. INTRODUCTION

En juin 1991, la Resolution Trust Corporation (RTC) a procédé à une enquête sur les institutions financières qui ont été mises sous séquestre et qu'elle a prises en charge depuis sa création en 1989. L'enquête, intitulée «Estimated Cash Recovery Survey» (ECRS), devait servir à fournir au Congrès, à l'administration et au grand public une estimation trimestrielle du montant qu'on espérait recouvrer de la vente des actifs détenus par les institutions d'épargne et de crédit (aux États-Unis, les «Savings and Loan Institutions») en faillite. Pour chacune de ces institutions, un actif correspond à un prêt consenti à une fin précise (un prêt commercial ou de construction, par exemple), ou encore à un type de bien détenu ou reçu par l'institution en garantie d'un prêt. L'enquête a porté sur 19 catégories d'actif, dont la liste figure au tableau 1.

Le montant du recouvrement escompté représente le recouvrement total en dollars, pour l'ensemble des 19 catégories d'actif. D'autre part, on porte un intérêt presque aussi grand au taux de recouvrement, c'est-à-dire le montant total du recouvrement escompté, divisé par la valeur comptable actuelle des actifs en cause. Par valeur comptable actuelle, on entend la valeur initiale du prêt, moins les remboursements reçus. Pour un prêt productif, c'est-à-dire qui fait toujours l'objet de remboursements par l'entreprise ou le particulier à qui il a été consenti, la valeur comptable diminue à mesure que le capital est remboursé.

L'intérêt et la complexité de ce plan de sondage tiennent à plusieurs facteurs. D'abord, étant donné que le taux de recouvrement de chaque catégorie d'actif suscite un intérêt certain, le plan de sondage doit échanger l'estimation globale de l'ensemble pour une estimation distincte du total pour chaque catégorie d'actif. Si l'on cherchait à optimiser l'échantillon pour répondre séparément à chacun de ces objectifs, on obtiendrait deux plans de sondage fort différents.

---

[1] C.D. Cowan, Statisticien principal, Resolution Trust Corporation, Washington (DC), É.-U. 20434-0001.

Ensuite, la population de cette enquête évolue très rapidement. Chaque trimestre, un certain nombre d'institutions sont dissoutes par l'Office of Thrift Supervision, un organisme de réglementation, puis confiées à la RTC pour mise sous séquestre. Les déposants de chaque institution sont dès lors remboursés, et l'actif de l'institution est pris en charge par la RTC. Il y a donc un apport soutenu d'institutions et d'actifs dans la population. Parallèlement, la RTC procède à la vente de certains actifs des institutions qu'elle a déjà prises en charge. Selon leur catégorie, les actifs sont liquidés à des rythmes très différents; par exemple, les prêts relatifs aux maisons unifamiliales se vendent beaucoup plus rapidement que les prêts de construction. Par conséquent, il existe également une sortie soutenue d'actifs, qui se fait à un rythme bien différent de celui de l'entrée d'actifs.

Tableau 1: Catégories d'actif utilisées dans l'estimation du recouvrement escompté.

Actifs visés par l'enquête

(1) Prêts hypothécaires, 1-4 familles - productifs
(2) Prêts hypothécaires, 1-4 familles - non productifs
(3) Prêts hypothécaires, 5 familles ou plus - productifs
(4) Prêts hypothécaires, 5 familles ou plus - non productifs
(5) Terrain en friche - prêts productifs
(6) Terrain en friche - prêts non productifs
(7) Prêts de construction - productifs
(8) Prêts de construction - non productifs
(9) Hypothèques commerciales - productives
(10) Hypothèques commerciales - non productives
(11) Prêts commerciaux - productifs
(12) Prêts commerciaux - non productifs
(13) Prêts personnels - productifs
(14) Prêts personnels - non productifs
(15) Biens immobiliers détenus
(16) Mobilier, agencements et matériel
(17) Avoirs subsidiaires
(18) Prêts subsidiaires
(19) Autres actifs

Actifs non visés par l'enquête

(20) Obligations de pacotille
(21) Titres hypothécaires
(22) Autres titres garantis
(23) Jugements
(24) Radiations

Enfin, les registres comptables de la RTC renferment des renseignements pouvant servir à des fins estimatives pour faire baisser la variance des estimations au moyen d'un estimateur par quotient ou par régression. Le taux de recouvrement mentionné plus haut illustre bien le type d'estimateur par quotient susceptible d'être retenu. Cependant, à cause de la nature des procédés qui relient le recouvrement à la valeur comptable initiale de l'actif, il faudrait délimiter le taux de recouvrement par un plancher de zéro et un plafond correspondant à l'unité. Nous pouvons réduire la variance des estimations en faisant appel au rapport entre la valeur de recouvrement escomptée et la valeur comptable (puisque celle-ci est connue pour chaque membre de la population), puis en limitant les estimations entre zéro et l'unité.

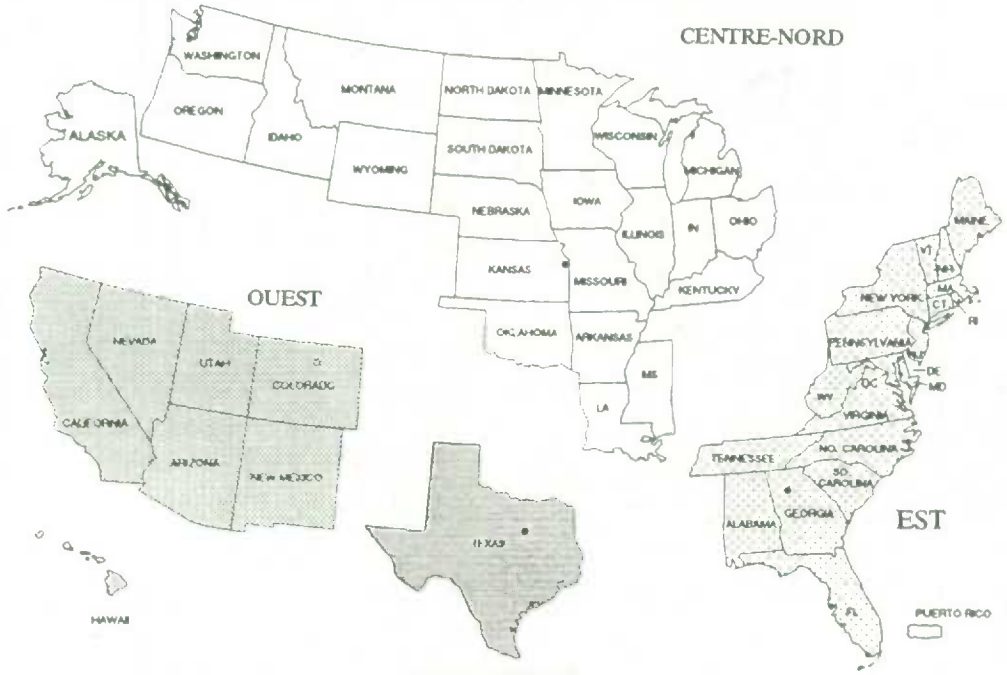
Le présent article expose les méthodes utilisées pour concevoir et réaliser l'enquête ECRS, tout en abordant les problèmes survenus au moment de la conduite de l'enquête. Il conclut en donnant des résultats provisoires de l'enquête, recueillis au cours des quatre premiers trimestres.

2. MÉTHODES

Avant de décrire le plan et l'analyse de l'enquête, il serait utile de préciser comment on entend recueillir les données, car ce facteur a une incidence économique (coûts à engager) sur le plan de l'étude.

On choisit d'abord un échantillon d'institutions et, au sein de chacune, un échantillon d'actifs au moyen de la méthode d'échantillonnage décrite ci-dessous. Les actifs de chaque institution sont évalués au moment de l'échantillonnage initial, puis ils sont successivement introduits dans l'échantillon et en sont supprimés par renouvellement selon un calendrier préétabli (qui est également décrit dans la section suivante). Le graphique 1 résume le processus.

Graphique 2: Régions de la Resolution Trust Corporation.



Une fois l'échantillon sélectionné, la liste des actifs à évaluer est remise à un cabinet d'experts-comptables dont les services ont été retenus pour déterminer le moment où chaque actif sera vendu, de même que la somme que la RTC peut s'attendre à recevoir. Les comptables calculent également l'apport de l'actif ou du bien au bénéfice d'exploitation, ainsi que les charges directes engagées pour la gestion de l'actif.

L'information sur l'apport escompté au bénéfice d'exploitation, les remboursements du prêt et les charges directes sont comptabilisés sur une base trimestrielle pendant les deux années qui suivent la date de collecte, puis sur une base annuelle au cours des trois années suivantes. Cette information trimestrielle favorise le «roulement» des estimations au cours des trimestres subséquents, lorsque l'institution se trouve supprimée de l'échantillon par renouvellement.

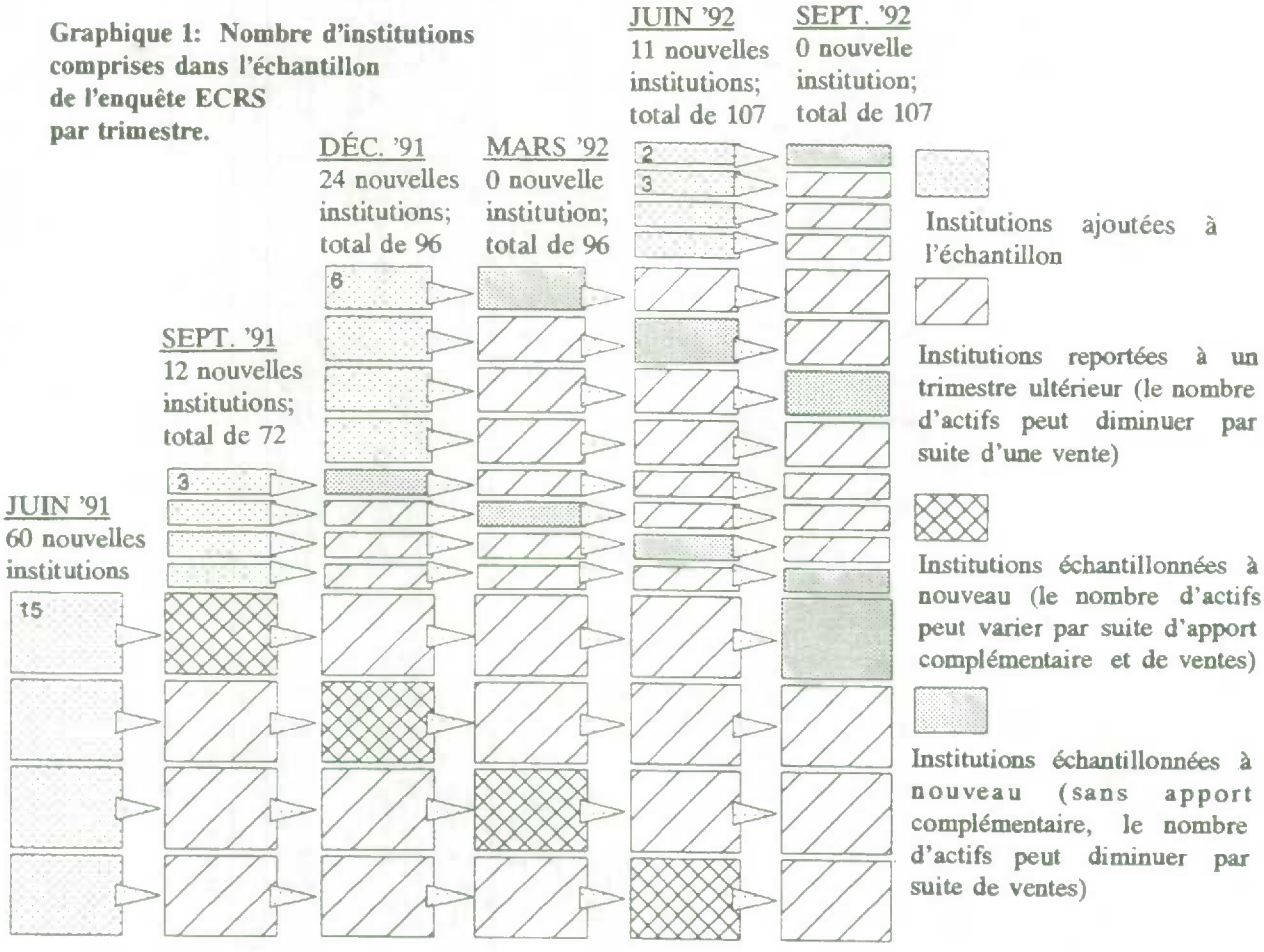
L'autre grand volet de la collecte consiste à déterminer si un actif compris dans l'échantillon est vendu au cours des trimestres subséquents. Dans l'affirmative, l'actif ne fait plus partie de l'enquête, car la RTC en a déjà réalisé le recouvrement. En effet, l'enquête a pour but d'estimer le recouvrement à réaliser dans l'avenir, alors que nous savons le montant exact recouvré pour un actif vendu; il est donc inutile d'échantillonner les actifs en question. Les spécialistes qui évaluent les actifs sur le terrain sont aussi chargés de signaler le prix obtenu pour les actifs vendus.

3. PLAN DE SONDAGE

Le plan consistait à mettre au point, à l'échelle nationale, un échantillon d'environ 5 000 actifs que devaient évaluer les experts-comptables. Par souci d'efficacité, tant sur le plan de la variance que sur celui des coûts, on a cherché à obtenir un échantillon de grappes stratifié à plusieurs degrés. De plus, chaque catégorie d'actif a été considérée comme une strate distincte lors de la deuxième étape de sélection.

À la première étape, on a procédé à la stratification au moyen d'un tableau à double entrée des institutions, lequel évoluerait à chaque trimestre en fonction de l'augmentation du nombre d'institutions mises sous séquestre. Les variables de stratification étaient la région - Est, Centre-Nord, Sud-Ouest et Ouest (catégories définies au graphique 2) - et la taille initiale de l'institution. Les catégories retenues pour la taille des institutions étaient les suivantes: «moins de 100 millions de dollars», «entre 100 millions et 500 millions de dollars» et «plus de 500 millions de dollars».

Graphique 1: Nombre d'institutions comprises dans l'échantillon de l'enquête ECRS par trimestre.



Les strates n'étaient pas équilibrées sur le plan de la taille des institutions, le nombre de petites institutions (selon la taille initiale) dépassant largement celui des grandes. Elles comportaient également un déséquilibre sur le plan de la taille globale (définie par la valeur comptable) et le nombre d'actifs, mais en sens contraire, les quatre strates ayant une taille initiale supérieure à 500 millions de dollars comptant pour plus de la moitié de l'actif total.

Il fallait tenir compte de plusieurs facteurs incompatibles au moment de la stratification et de l'estimation:

(1) la distribution des actifs était fortement asymétrique vers les grandes institutions;

(2) le nombre d'institutions était fortement asymétrique vers les petites institutions;

(3) le principal objectif de l'enquête consistait à produire une seule estimation nationale de recouvrement;

(4) toutes les institutions ne détiennent pas toutes les catégories d'actif, de sorte qu'un échantillon d'institutions où certaines strates sont faiblement représentées risque d'être complètement dépourvu de certaines catégories d'actif;

(5) les taux de recouvrement par région, par taille d'institution et par catégorie d'actif étaient d'une importance égale et ils suivaient de près l'estimation nationale en ce qui concernait l'usage qu'on allait faire des données de l'enquête.

D'ordinaire, la méthode la plus efficace consiste à établir une fonction objective (de variance, par exemple) à minimiser étant donné un coût fixe. Cependant, comme on ignorait presque tout du coût relatif occasionné par l'évaluation de l'actif et de la variation des estimations de recouvrement par strate, la méthode la plus efficace consistait en l'occurrence à choisir un nombre égal d'institutions par strate et un nombre à peu près égal d'actifs pour chaque institution de l'échantillon. Rétrospectivement, cette méthode s'avéra la bonne vu la difficulté à obtenir une liste complète des actifs de chaque institution comprise dans l'échantillon. Toute autre méthode faisant appel à des procédés d'échantillonnage de l'actif plus compliqués aurait considérablement ralenti le déroulement de l'enquête.

En juin 1991, on a échantillonné 60 institutions, soit environ cinq par strate (à cause de la répartition initiale des institutions, on a prélevé quatre institutions dans une strate et, pour compenser, six dans une autre). Les 60 institutions ont ensuite été réparties en quatre groupes de renouvellement, à raison de 15 par groupe, aux fins du suivi au cours des trimestres subséquents. La répartition cherchait à représenter chacune des 12 strates au moins une fois par trimestre et à éviter que la même strate soit représentée plus de deux fois par trimestre. On a obtenu pour chaque institution la liste de tous les actifs dans les 19 catégories en vue de l'échantillonnage à la deuxième étape, stratifiée par catégorie d'actif. On a ensuite choisi un échantillon d'au moins cinq actifs dans chaque catégorie qui en comptait au moins cinq. Lorsqu'une catégorie comportait moins de cinq actifs, ils étaient tous incorporés dans l'échantillon. Les actifs étaient ordonnés par valeur comptable (taille) au sein de chaque catégorie, puis échantillonnés systématiquement.

En septembre 1991, le plan s'est compliqué. On disposait désormais de trois sources ou listes d'institutions. Il y avait d'abord les institutions de l'échantillon initial de juin 1991, qu'il fallait rejoindre en septembre afin de réévaluer les actifs évalués en juin. En effet, la situation du prêt ou la conjoncture économique déterminant le prix de vente ont pu avoir changé dans l'intervalle, ou d'autres facteurs auraient pu avoir une incidence sur l'évaluation du prêt.

La deuxième source correspondait aux institutions échantillonnées à l'origine, soit en juin 1991, mais qu'on n'allait pas rejoindre en septembre. Les actifs de ce groupe d'institutions seraient reportés à une date ultérieure, comme nous l'avons vu plus haut. Les institutions incluses dans les deux premières sources représentaient toutes les institutions dissoutes en date de juin 1991.

En décembre 1991, nous nous retrouvions dans une situation à peu près analogue, sauf que la première source de données correspondait désormais aux 15 institutions du deuxième groupe de renouvellement défini en juin 1991, plus les trois institutions affectées au deuxième groupe de renouvellement en septembre 1991. À partir de la deuxième source de données, nous reportons les estimations des trois groupes de renouvellement restants de juin et septembre. La troisième source provenait encore une fois d'un nouvel échantillon d'institutions «dissoutes» entre août et novembre 1991, l'actif étant échantillonné de la même manière qu'en juin 1991.

La démarche a été exactement la même en mars et en juin 1992, sauf qu'aucune institution ne s'est ajoutée en mars en raison du nombre insuffisant de nouvelles institutions dissoutes. En juin 1992, nous avons ajouté 11 institutions, à raison encore d'une par strate. Aucune institution n'a été échantillonnée dans la strate 4, car aucune n'a été dissoute dans cette strate.

Enfin, à chaque reprise de l'enquête, on a vérifié chaque actif pour déterminer s'il avait été vendu. On a procédé de la sorte pour tous les groupes de renouvellement compris ou non dans l'échantillon, afin d'obtenir plus rapidement des renseignements sur les actifs vendus tout en compensant les éventuelles distorsions découlant de la vente de groupes d'actifs à des taux différentiels.

4. ESTIMATION

Le dernier élément du projet consiste en un schéma d'estimation. Si on a employé des méthodes standard pour la plupart des estimations de l'enquête, il a fallu adapter certaines méthodes pour déterminer les intervalles de confiance.

Pour la plupart des estimations du taux de recouvrement, nous disposons de beaucoup d'information pouvant servir à calculer des estimateurs par quotient ou par régression. Plus précisément, nous connaissons la valeur comptable de chaque actif et, pour les parties longitudinales de l'enquête, nous connaissons la variation de la valeur comptable. À tout le moins, nous pouvons utiliser la valeur comptable pour tous les actifs de la population et de l'échantillon, et le recouvrement prévu pour tous les actifs de l'échantillon, puis nous pouvons calculer l'estimateur par quotient stratifié classique pour un échantillon de grappes à deux degrés (Cochran 1963). L'estimateur par quotient et l'estimateur par variance (pour le quotient) sont tous les deux bien définis et connus depuis un long moment.

Étant donné que certains taux estimatifs de recouvrement sont exceptionnellement faibles ou élevés (respectivement proches de zéro ou de l'unité), il s'avère que les queues des intervalles de confiance produits dans le cadre de la procédure d'estimation de l'enquête sont inférieures à zéro (ce qui suppose un recouvrement négatif) ou supérieures à l'unité (ce qui sous-entend un recouvrement supérieur à la valeur initiale de l'actif). Bien que cela soit possible pour un actif donné dans des circonstances très exceptionnelles, c'est impossible lorsqu'il s'agit du taux de recouvrement pour l'ensemble de la population, à cause des méthodes retenues pour vendre les actifs. Cette restriction signifie que l'approximation normale, utilisée couramment pour la construction d'intervalles de confiance, convient à la présente enquête.

Comme solution de rechange, nous avons fait l'essai d'une approche fondée sur la méthode de Bayes (Box et Tiao 1973). Ainsi, nous avons supposé que le taux de recouvrement était un paramètre tiré d'une distribution antérieure, soit la distribution bêta. Nous avons fait appel à la méthode des moments pour estimer les paramètres de la distribution bêta, à partir de la moyenne et de la variance des estimations des taux de recouvrement provenant de l'échantillon. Enfin, nous avons déterminé, directement à partir de la distribution bêta, les limites inférieure et supérieure de l'intervalle de confiance afin de disposer de la limite de confiance la plus étroite qui soit, compte tenu de la restriction voulant que les limites inférieure et supérieure se trouvent dans l'intervalle allant de zéro à l'unité. Dans tous les cas, cette méthode a rétréci les intervalles de confiance par rapport aux résultats du calcul fondé sur la distribution normale, mais elle n'a provoqué aucune variation de la moyenne ou de la variance de l'estimation.

5. RECHERCHES COMPLÉMENTAIRES

À l'instar de tout bon programme statistique qui se prolonge indéfiniment, celui-ci a besoin d'autres recherches, plus précisément dans le domaine de la production d'estimations composites et de l'exploitation plus judicieuse du caractère longitudinal des données. Par ailleurs, il faudra approfondir les recherches pour savoir s'il est possible d'améliorer la méthode de Bayes (fondée sur la distribution bêta) en examinant la tendance descendante associée à la plupart des taux de recouvrement (en fonction du marasme global de l'économie).

RÉFÉRENCES

Cochran, W.G. (1963). *Sampling Techniques, Second Edition*. John Wiley and Sons, Inc., New York (NY).

Box, G.E.P., et Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley Publishing Company, Reading (MA).

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

PANELS D'ENTREPRISES ET CONFIDENTIALITÉ: LA MÉTHODE DES PETITS AGRÉGATS

D. Defays et P. Nanopoulos[1]

RÉSUMÉ

La nécessité de disposer de réseaux de collecte rapides et susceptibles de s'adapter facilement à des nouvelles demandes d'informations s'impose de plus en plus au niveau européen.

Les collecteurs d'informations (les instituts nationaux de statistique, généralement) n'étant pas toujours en mesure de communiquer des données individuelles, il importe de proposer des méthodes d'agrégation de ces informations, de définition d'unités virtuelles qui préservent la confidentialité des données tout en maximisant l'information micro-économique fournie.

Différentes techniques sont examinées et les résultats obtenus sont comparés.

MOTS CLÉS: Confidentialité; micro-agrégation; composantes principales; méthode de groupement floue.

1. INTRODUCTION

La réalisation d'analyses économiques fondées uniquement sur des données macro-économiques est, à tout le moins, problématique. Les tendances peuvent être difficiles à interpréter: des phénomènes de variation survenant à un niveau micro-économique (qui sont souvent la source d'une information utile) sont masqués par l'agrégation des données, et les méthodes d'agrégation appliquées aux données peuvent déterminer la nature des analyses qui peuvent être effectuées, ainsi que la nature et le nombre des modèles qui peuvent être testés.

Il est généralement impossible, une fois qu'un ensemble de tableaux a été établi de manière définitive, de répondre à des questions exigeant une exploitation de l'information sous des angles différents de ceux présentés; par exemple, lorsque des données sur les entreprises ont été regroupées selon des catégories de tailles d'entreprise exprimées en nombres d'employés, il est impossible de centrer l'attention sur une sous-population d'unités définies en termes de chiffre d'affaires. Les analyses exploratoires sont presque impossibles. Comment faire, dans ce cas, pour s'intéresser statistiquement à de petites entreprises innovatrices ou à des unités effectuant de la sous-traitance, si ces caractéristiques n'ont pas été incluses au moment de la définition initiale des tableaux statistiques?

Eurostat, le bureau statistique de la Communauté européenne, manque actuellement de données à caractère individuel sur les entreprises. Afin de résoudre ce problème, un projet visant à créer un réseau de panels d'entreprises nationaux coordonnés a été amorcé.

2. UN PANEL D'ENTREPRISES EUROPÉEN

Des spécifications visant à adopter une approche européenne commune en ce qui concerne les panels d'entreprises sont actuellement préparées. L'objectif est d'aider les panels nationaux existants, en Europe, à converger vers des normes communes. Toutefois, avant qu'on puisse songer à mettre en oeuvre de telles normes, différents problèmes devront être résolus. L'un d'eux consistera à étudier les ajustements qui pourraient être

[1] D. Defays et P. Nanopoulos, Eurostat.

apportés aux panels existants pour les rendre conformes aux exigences de la Communauté. En effet, certains pays ont déjà des panels d'entreprises, et le problème sera de déterminer dans quelle mesure ils sont compatibles avec les normes recommandées et d'examiner quelles modifications devraient leur être apportées pour améliorer leur comparabilité avec les autres panels nationaux. Un autre problème consistera à analyser les conditions dans lesquelles certains types de microdonnées pourraient être transmises à Eurostat. Les instituts nationaux de statistique (INS) des États membres de la Communauté européenne ne sont pas toujours en mesure de fournir des données sur des entreprises individuelles, car ils sont tenus de protéger la confidentialité des données qui leur sont transmises par les personnes physiques et morales qui constituent leurs sources d'information. Ils ne transmettent que des tableaux de données agrégées, et les analyses effectuées à l'échelle européenne sont en général confinées à cette source de données.

Le but de la présente communication est d'examiner comment nous pouvons maximiser la quantité de données transmises à Eurostat par les INS, tout en préservant le plus possible le caractère confidentiel des données.

3. LA QUESTION DE LA CONFIDENTIALITÉ

Le problème, évidemment, n'est pas nouveau. Toutefois, certains aspects de la façon dont il se pose pour Eurostat méritent qu'on s'y arrête et justifient peut-être l'adoption d'une nouvelle approche méthodologique. Premièrement, la gamme des usages que nous faisons des données transmises est particulièrement vaste, et certainement plus large que celle des instituts de recherche traditionnels, qui s'intéressent généralement à un type d'analyse bien précis. Deuxièmement, les données sur les entreprises ne doivent pas seulement répondre aux besoins de l'analyse économique, mais aussi aider la Commission à effectuer la gestion et le suivi de certains projets de la Communauté, ainsi qu'à évaluer l'impact de ses programmes. De tels besoins exigent, évidemment, une exploitation de l'information sous de nombreux angles différents.

En outre, la préservation de la confidentialité exige le pré-traitement de grandes quantités d'information dans douze États membres. Ces activités doivent être réalisées avec le maximum d'efficience. Enfin, puisque la préservation de la confidentialité est une question particulièrement délicate et parce que la transmission des données à Eurostat, qui est un organe interne de la Commission des Communautés européennes, est source de controverse, la méthode de protection des données doit être simple, facile à comprendre et parfaitement étanche. Il doit être possible de l'expliquer à des non-spécialistes.

La présente communication expose quelques résultats théoriques relatifs à l'agrégation d'unités dans des classes de tailles fixes et décrit une méthode simple d'agrégation minimale de données individuelles qui ne nuit pas aux analyses à effectuer, qui maximise l'information transmise et qui assure la préservation de la confidentialité statistique. L'impact de cette méthode sur l'analyse de données longitudinales est examiné. Plusieurs simulations portant sur des données relatives aux entreprises ont été effectuées et analysées.

4. ÉTAT DES CONNAISSANCES

Comme nous l'avons déjà indiqué, la protection des données individuelles n'est évidemment pas un problème nouveau. Diverses méthodes ont été étudiées et utilisées à cette fin. Une analyse comparative est présentée, par exemple, dans G. Paass (1988). Ces méthodes existantes consistent à perturber les données en leur ajoutant un «bruit», en construisant des unités par la permutation de blocs d'information (c.-à-d. de valeurs pour des sous-ensembles de variables) entre les unités originales, ou en construisant des micro-agrégats selon des axes qui diffèrent, toutefois, de ceux proposés dans la présente communication.

Les méthodes classiques de protection des données individuelles ne conviennent pas, selon nous, à notre situation, pour diverses raisons. Les méthodes de perturbation n'offrent pas une garantie adéquate de confidentialité, notamment avec des distributions aussi asymétriques que celles traitées dans le domaine des statistiques des entreprises. Elles sont en outre susceptibles d'entacher d'un biais l'estimation d'un certain nombre de paramètres (Adam et Wortmann 1989). Leur application à l'analyse de données longitudinales peut également poser des difficultés.

La permutation de données permet la préservation de certaines caractéristiques de la distribution multidimensionnelle originale, mais l'application de cette méthode peut se révéler très complexe. Elle semble offrir une protection efficace sur le plan de la confidentialité des données statistiques, mais son application à des analyses exploratoires comportant l'étude de la structure multidimensionnelle des données peut poser des problèmes, car cette structure, forcément, a été brisée par la construction des unités synthétiques. Il y a aussi un risque que l'application de ce genre de méthode à des données longitudinales n'introduise un biais d'ampleur importante.

En général, les règles permettent la transmission de données agrégées lorsque le nombre d'unités de l'agrégat dépasse un certain seuil k (normalement, $k = 2$) et qu'aucune des unités ne représente la quasi-totalité de l'agrégat. Une application stricte de cette règle nous permet d'obtenir, au lieu de données individuelles, de petits agrégats, ou encore les moyennes de ces micro-agrégats. Ceux-ci peuvent jouer le rôle d'unités fictives, que nous proposons de baptiser «pivots».

Pour des populations présentant un degré élevé d'homogénéité, les avantages de cette formule sont évidents. Par contre, la micro-agrégation aurait semble-t-il un effet significatif, dans certains cas, sur les propriétés statistiques des données. Elle oblige aussi à recourir à des méthodes de classification qui peuvent être coûteuses si elles visent de grands volumes de données.

Eurostat a également étudié une technique permettant la définition de prototypes (Bragard et coll. 1988). La méthode consiste, essentiellement, à prélever parmi une population un certain nombre d'unités virtuelles et réelles considérées comme «représentatives» de la population. Les techniques utilisées dans ces études s'inspirent d'une méthode de formation de grappes floues proposée par M. Roubens, qui consiste, essentiellement, à réduire les dimensions des données au moyen d'une analyse en composantes principales, puis à choisir des unités représentatives en se fondant, entre autres, sur les grappes floues. Dans chaque classe, l'élément ayant la fonction d'«appartenance» maximum est choisi comme prototype. Les degrés d'«appartenance» sont ensuite utilisés pour estimer les paramètres de la population. Une simulation a déjà produit des résultats prometteurs (mais non encore confirmés); toutefois, la méthode est relativement complexe et certains des critères de choix des prototypes demeurent plutôt arbitraires.

5. EXEMPLES DE BESOINS

Une question qui se pose souvent est celle de la définition de la taille d'une entreprise. Du point de vue administratif, la notion de «taille» d'une entreprise est liée à la notion d'unité (unité juridique, entreprise, unité locale, etc.) et à la mesure de plusieurs variables pertinentes comme le «nombre d'employés», le «chiffre d'affaires» ou l'«actif immobilisé», ou encore d'autres types de variables liées aux intrants et aux extrants de l'unité.

Il existe plusieurs façons de tenter de répondre à cette question. Nous en donnons ici trois exemples.

a) Définition de classes d'après les variables pertinentes

Le problème de cette méthode simple, c'est que l'utilisation de variables différentes produit des classifications différentes des unités. La façon normale de procéder, selon la législation de la CE, consiste à définir un seuil pour la variable X = «emploi» ou la variable Y = «chiffre d'affaires». Le problème qui se pose alors est de coupler le seuil X avec le seuil Y d'une façon raisonnable.

L'analyse de ce problème exige que l'on dispose d'un échantillon de taille relativement élevée, qui permette de comparer des méthodes fondées sur les quintiles avec des méthodes basées sur des scores.

b) Méthode des composantes principales

Si l'on s'entend sur les variables à considérer comme importantes, une façon raisonnable de résoudre ce problème est d'examiner la structure de corrélation de ces variables sur les diverses sous-populations, de façon à éviter les corrélations faibles attribuables à des disparités entre les classes d'activité.

L'utilisation de la première composante principale apporte une solution objective raisonnable au problème. Puisqu'il est possible d'ordonner toutes les unités et de définir des seuils selon la première composante principale, ces méthodes n'exigent pas de données individuelles; toutefois, puisque des annulations doivent être calculées pour un grand nombre de sous-populations, des données individuelles comportant une information catégorique se révèlent plus commodes.

c) Méthode des grappes

Les algorithmes de formation de grappes, comme celui des k moyennes (Hartigan 1975) peuvent être utilisés pour produire des groupes d'entreprises censés correspondre à diverses «classes de tailles». Une telle analyse ne peut se faire qu'à laide de données «individuelles».

6. PRÉSENTATION THÉORIQUE

a) Théorie du problème général des petits agrégats

Supposons que la population totale soit formée de N unités. À chaque unité correspond un vecteur X formé de p variables. Le but est de répartir l'ensemble Ω en n, $(n=N/k)$, groupes de k **points chacun**, $(G_1,..., G_n)$, de telle façon que les n groupes soient aussi homogènes que possible.

Pour être en mesure d'établir l'homogénéité des groupes, il nous faut une notion de proximité ou de distance $d(\omega,\omega')$ entre des points à l'intérieur de Ω, qui doit dépendre des variables observées, et de distance connexe $D(G,G')$ entre des groupes de points de Ω. Théoriquement, la quantité que nous devons minimiser est une sorte de «variance intra-groupe»:

$$\Psi(G_1,...,G_N) = \sum_i \psi(G_i), \tag{6.1}$$

où

$$\psi(G_i) = \sum_{\omega \in G_i} P(\omega)\, D(\omega,G_i)^2 . \tag{6.2}$$

Le problème de la répartition nxk diffère du problème classique et bien connu de formation de grappes (algorithme des k moyennes (*cf.* [3])), dans lequel le but est de répartir l'ensemble de la population en un nombre fixe de groupes. Dans ce cas, il n'y a pas de condition de cardinalité.

b) Le problème de la répartition nxk dans R^p

Le cas le plus fréquent et le plus important est celui où l'espace des valeurs de $X=(X_1,...Xp)$ est l'ensemble des nombres réels et la distance est la distance euclidienne habituelle dans R^p.

Les distances sont ainsi définies:

$$d(\omega,\omega') = \| X(\omega) - X(\omega') \|$$

$$D(G,G') = \| m(G), m(G') \| \text{ où } m(G) \text{ est la moyenne de } X \text{ sur } G.$$

Dans ce cas, l'expression (6.2) devient:

$$\psi(G) = \sum_{x \in G} p(x) D(x,G)^2 = \sum_{x \in G} p(x) \| x - m(G) \|^2, \tag{6.3}$$

ce qui, en termes de l'espérance conditionnelle du vecteur X sur le champ engendré par la répartition $G = (G_1,...,G_n)$ peut s'écrire

$$\Psi(G_1,...,G_n) = E(\| X - E(X/G) \|^2),$$

où

$$E(X/G) = \sum m(G_i) I_{G_i} \tag{6.4}$$

I_{G_i} étant la fonction indicatrice de l'ensemble G_i.

Supposons que le vecteur X soit centré $(E(X) = 0)$ dans $L_2(\Omega,\mathcal{A},P)$. Le théorème de Pythagore donne la décomposition suivante:

$$\begin{aligned} \Psi(G_1,...,G_n) &= E(\| X - E(X/G) \|^2) \\ &= E(\| X \|^2) - E(\| E(X/G) \|^2) \end{aligned} \tag{6.5}$$

et le problème prend la forme équivalente:

Maximiser l'expression

$$E(\| E(X/G) \|^2), \tag{6.6}$$

sous la contrainte $P(G_i)=k/N$.

c) Le problème de la répartition $2xk$ dans R^p

Dans le cas où $n = 2$, la répartition G est engendrée par un sous-ensemble A, $G = \{A, A^c\}$, et un calcul simple montre que la quantité à maximiser est donnée par:

$$\Psi(A) = \| E(I_A X) \|^2, \tag{6.7}$$

sous la contrainte $P(A) = 1/2$.

Supposons d'abord que $p = 1$. La quantité en (6.7) devient alors: $\Psi(A) = (\int I_A X dP)^2$.

D'après le lemme classique de Neumann-Pearson (voir [4]) sur la construction du test le plus puissant, l'ensemble produisant la maximisation est de la forme:

$$A^* = \{\omega \in \Omega \mid X(\omega) \geq \lambda\}, \tag{6.8}$$

pour une certaine constante λ.

Il en résulte, en pratique, que A^* est composé des $N/2$ éléments de Ω ayant les valeurs les plus élevées de X.

Dans le cas général ($p>1$) nous avons un résultat semblable:

Lemme A

Il existe un vecteur $c \in R^p$ et une constante λ tels que la solution de (6.7) soit de la forme:

$$A^* = \{\omega \in \Omega \mid \sum c_i X_i(\omega) \geq \lambda\}. \tag{6.9}$$

d) L'hyperplan dans le problème de la répartition en 2 groupes

Quelques autres aspects peuvent être abordés au sujet de la définition de l'ensemble optimal A^*.

La tâche la plus difficile est de déterminer le vecteur c car la constante λ peut ensuite être facilement établie d'après la condition latérale $|A| = N/2$.

Lemme B

Le vecteur $c \epsilon R^p$ qui définit l'ensemble optimal $A = \{\omega \epsilon \Omega \mid \sum c_i X_i(\omega) \geq \lambda\}$ dans (6.9) satisfait à la condition

$$c = E(I_A X). \tag{6.10}$$

Un corollaire immédiat de la preuve est que, si nous avons une répartition avec un ensemble A nous pouvons améliorer la valeur de Ψ en remplaçant A par l'ensemble $A^* = \{Y(A) > \lambda\}$, où $Y(\omega) = < X(\omega), E(I_A X) >$. Cette opération peut être répétée jusqu'à ce que $A = A^*$.

En fait, il est démontré dans la preuve du lemme ci-dessus que

$$\Psi(A) = E[I_A Y(A)] \leq E[I_{A^*} Y(A)] = E[I_A Y(A^*)] \leq E[I_{A^{**}} Y(A^*)].$$

Ces résultats nous permettent de proposer un algorithme pour le problème général dans R^p.

e) Le problème général dans R^p

Dans le cas général $(n>2)$, il est évident que dans chaque paire de groupes (G_i, G_j) il doit y avoir séparation par un hyperplan, comme il a été démontré dans la section précédente. Cet hyperplan peut être choisi de façon à être perpendiculaire à la ligne joignant les barycentres des deux groupes.

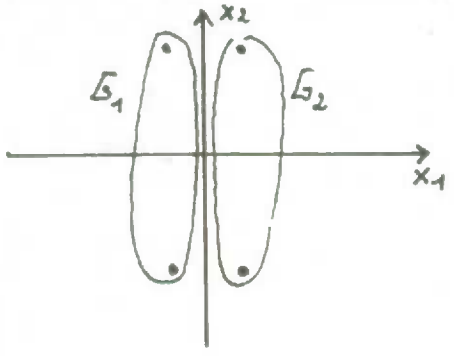
7. ALGORITHME PROPOSÉ POUR LE CAS GÉNÉRAL

Les résultats établis au paragraphe précédent incitent à adopter l'algorithme suivant.

a) Algorithme pour le cas $n = 2$

1. Commencer par effectuer toute répartition en 2 classes de «même» taille satisfaisant la condition de séparation par un hyperplan.
2. Ordonner les projections des points X sur la ligne joignant les moyennes des deux classes.
3. Prendre la nouvelle répartition définie par les premiers $N/2$ points sur cette ligne et les derniers $N/2$ points.
4. Si cette répartition est identique à la première, arrêter; sinon, reprendre à 2.

Il est facile de voir que cet algorithme convergera, mais aussi qu'il peut être piégé au niveau de minimums locaux, par exemple dans l'exemple ci-dessous, si la répartition de départ est la suivante:



b) Algorithme pour le cas général

Dans le cas général ($n \geq 2$), l'algorithme doit être transformé de la façon suivante:

1. Commencer avec toute répartition nxk satisfaisant à la condition de séparation par un hyperplan. On peut le faire, par exemple, en déplaçant un hyperplan parallèlement à un hyperplan fixe et en prenant des groupes de k points.

2. Prendre successivement toutes les paires (G_i, G_j) de groupes et optimiser cette répartition en deux groupes en suivant l'algorithme précédent.

3. Reprendre à l'étape (2) jusqu'à ce qu'aucune modification supplémentaire ne soit possible.

8. QUELQUES SIMPLIFICATIONS

Dans notre introduction, nous avons mentionné le besoin d'utiliser une méthode simple et peu coûteuse. La marche à suivre proposée jusqu'ici est peut-être encore un peu complexe et pourrait être coûteuse à mettre en oeuvre. Certaines simplifications sont souhaitables.

Nous proposons d'accroître le nombre de contraintes dans la définition de la répartition G. D'abord, comme nous l'avons déjà indiqué, nous précisons que chaque classe doit contenir le même nombre d'éléments. Ensuite, pour simplifier la construction de G, nous imposons la condition que cette répartition soit le fruit d'une simple mise en ordre des unités d'après une variable unidimensionnelle Y, par le regroupement d'unités contiguës. Plus précisément, nous déclarons qu'il existe une variable $Y{:}\Omega \rightarrow R$, et des bornes a_i, telles que

$$\omega \in G_i \Leftrightarrow a_i < Y(\omega) < a_{i+1}.$$

La répartition G est donc simplement définie en termes de $Q\ (i/n)$ de la distribution d'une variable Y.

Il est à signaler que la condition que nous imposons ici revient à supposer que les hyperplans de séparation obtenus dans le cas général sont tous parallèles.

En fait, cette hypothèse nous ramène à une situation pratique simple. En effet, il a déjà été proposé que dans le cas des statistiques sur les entreprises, par exemple, les unités soient ordonnées d'après le nombre d'employés et que G soit défini en formant des groupes consécutifs de k entreprises selon l'ordre obtenu.

Mais pourquoi choisir le nombre d'employés plutôt que le chiffre d'affaires? N'y a-t-il pas, parmi les combinaisons linéaires des variables X_i, une combinaison qui minimise la perte d'information résultant du remplacement de tous les éléments d'une classe par leur moyenne? Quelle combinaison des valeurs X_i minimisera la perte d'information occasionnée par la formation de groupes consécutifs de k éléments d'après l'ordre défini par cette combinaison?

Ce problème constitue donc une version simplifiée de notre problème initial.

9. SIMULATIONS

Plusieurs simulations nous ont révélé l'importance du choix de cette variable unidimensionnelle servant à ordonner les unités.

Les données que nous avons analysées ont été prélevées à même un échantillon d'environ 5 000 entreprises industrielles, caractérisées par 11 variables économiques: effectif employé, chiffre d'affaires, ventes à l'exportation, investissements corporels, valeur ajoutée, rémunération des employés, marge d'exploitation brute, montant de la sous-traitance confiée à des tiers, dépenses de publicité, nombre d'établissements et nombre d'activités économiques de l'entreprise.

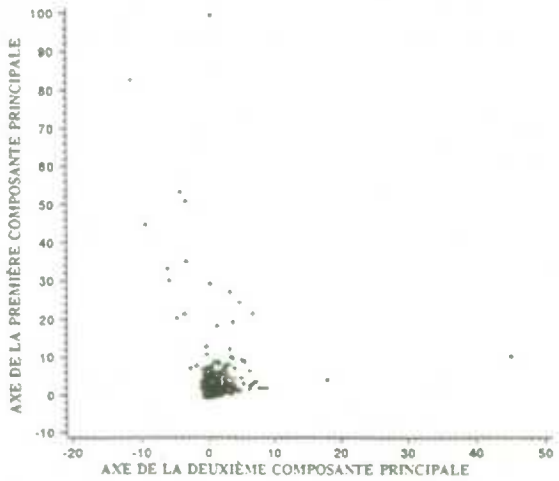
Les figures qui suivent présentent les principaux résultats des simulations: structure des données, pertes d'information (en pourcentage) attribuables à la formation de classes de k unités, pour différentes valeurs de k et différents choix de la variable Y, mesurées par le coefficient défini par

$$g(X/G) = E(\| X - E(X/G) \|^2) \; / \; E(\| X \|^2)$$

et indication quant à l'effet de la formation de groupes de 3 unités sur la structure interne des données.

La figure 1 montre les données dans l'espace des deux premières composantes principales. Environ 40 «valeurs aberrantes» ont été retranchées de la population afin d'éviter l'effet lié à l'asymétrie excessive des distributions, et toutes les variables ont été réduites (moyenne = 0 et écart-type = 1).

Figure 1: Diagrammes de la première composante principale par rapport à la deuxième composante principale.



Diagrame des données avant la suppression des «valeurs aberrantes»

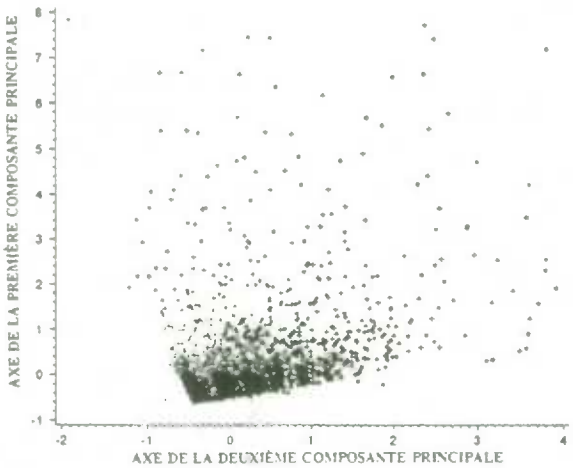
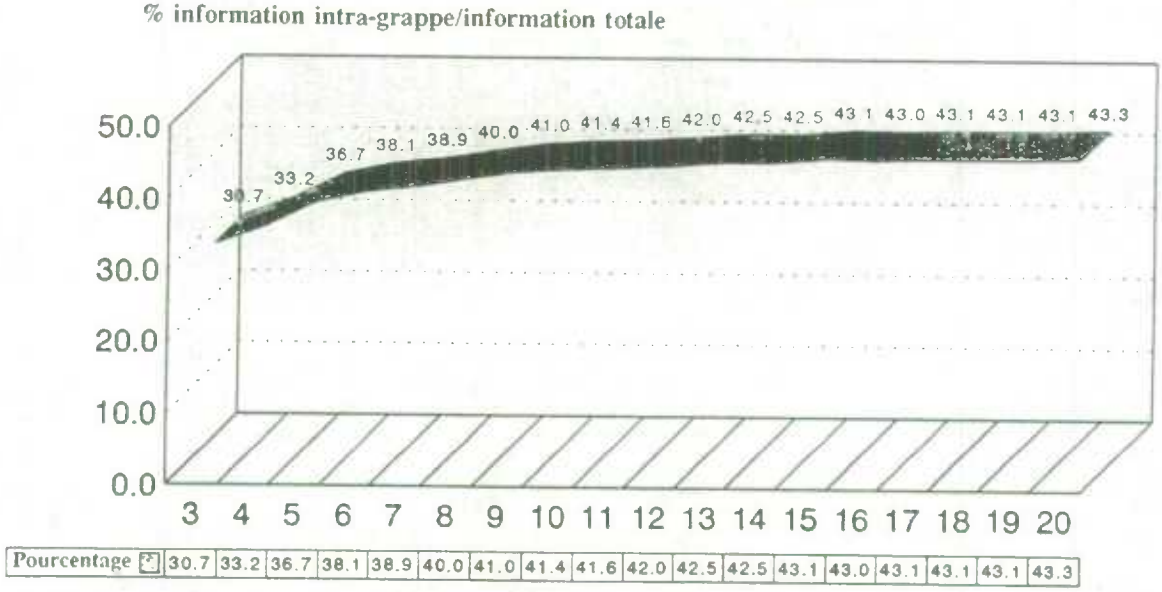


Diagramme des données après la suppression des «valeurs aberrantes»

La figure 2 montre l'accroissement des pertes d'information mesurées par $g(X/G)$ avec l'augmentation de k, c.-à-d. la taille des groupes. Dans cette figure, les entreprises ont été ordonnées selon leur nombre d'employés (qui joue donc ici le rôle de variable Y) et regroupés k par k le long de cette dimension.

Figure 2: Rapport (en pourcentage) de l'information intra-grappe sur l'information totale (valeurs aberrantes supprimées).

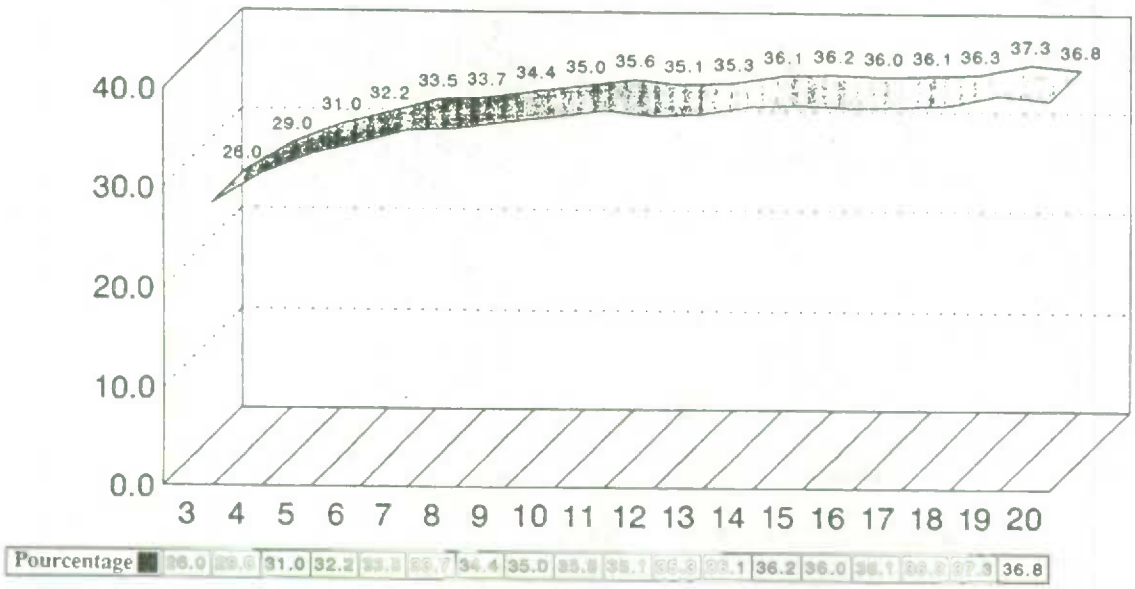


Nombre d'observations dans chaque grappe

Classement selon le nombre d'employés (en ordre croissant)

Les figures 3 et 4 sont semblables à la figure 2, sauf que la dimension (variable *y*) qui définit les groupes est respectivement la valeur ajoutée et la première composante principale. Il convient de souligner que, comme prévu, les résultats obtenus avec la composante principale sont supérieurs à ceux obtenus aux figures 2 et 3.

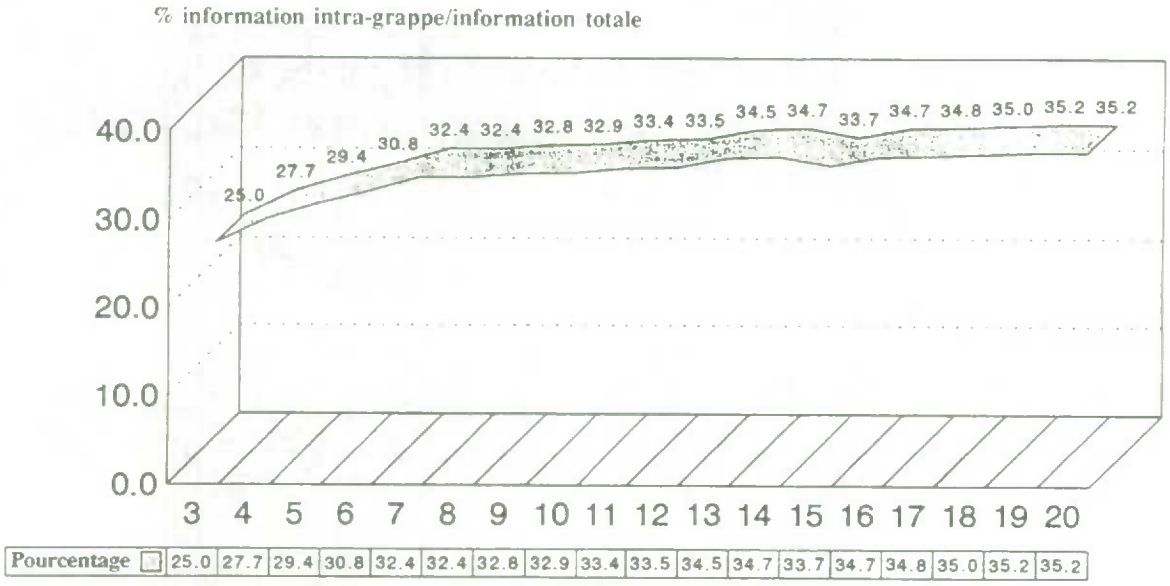
Figure 3: Rapport (en pourcentage) de l'information intra-grappe sur l'information totale (valeurs aberrantes supprimées).



Nombre d'observations dans chaque grappe

Classement selon la valeur ajoutée (en ordre croissant)

Figure 4: Rapport (en pourcentage) de l'information intra-grappe sur l'information totale (valeurs aberrantes supprimées).



La figure 5 présente les corrélations entre les variables initiales et leurs versions correspondantes une fois qu'il y a eu agrégation 3 par 3 le long de la première composante principale. Les valeurs indiquent principalement comment le processus d'agrégation influe sur chaque variable.

Figure 5: Corrélation entre les données par groupes de *k* unités et les données originales.

Variable:	Corrélation:
Effectif Employé	0.94357
CAHT	0.96582
Ventes à l'exportation	0.84760
Investissements Corporels	0.81040
VAHT	0.98181
Frais de Personnel	0.96182
EBE	0.86066
Montant de la sous-traitance confiée	0.70223
Dépenses de Publicité	0.65148

BIBLIOGRAPHIE

Adam, et Wortmann (1989). Security control methods for databases. A comparative study. *ACM Computing Surveys*, 21, 4.

Bragard, L., Roubens, M., Libert, J., et Gailly, B. (1988). Examen d'une méthode d'échantillonnage par la selection de prototypes. Rapport interne Eurostat.

Hartigan, J.A. (1975). *Clustering Algorithms*, Wiley, New York.

Lehmann, E.L. (1959). *Testing Statistical Hypothesis*, John Wiley & Sons, New York.

Paass G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business & Economic Statistics*, 6, 4.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

LA CONTRIBUTION DE L'INSTITUT DE RECHERCHE ÉCONOMIQUE IFO À L'ÉTUDE DES DONNÉES DE PANEL: LES TOUTES DERNIÈRES INNOVATIONS EN RECHERCHE APPLIQUÉE ET MÉTHODOLOGIQUE

G. Nerb et H. Seitz[1]

RÉSUMÉ

L'institut IFO réalise ses enquêtes-entreprises mensuelles depuis plus de 40 ans (environ 10 000 questionnaires remplis dans l'industrie du commerce et de la construction). Outre l'utilisation plus traditionnelle de ce genre de renseignements (systèmes d'indicateurs, inclusion dans des modèles économétriques), on a entrepris plusieurs projets de recherche visant à exploiter le contenu informationnel des micro-données dans des études longitudinales. On a démontré, par exemple, que les ventes de produits industriels ne réagissent que très faiblement à des variations inattendues dans les coûts alors que les changements imprévus dans la demande amènent des ajustements significatifs à court terme dans la production.

MOTS CLÉS: Données longitudinales; estimation par panel.

1. INTRODUCTION

Les *données de panel*, ou données longitudinales, sont un ensemble d'observations enregistrées pendant plusieurs périodes de temps sur un nombre (habituellement) élevé d'unités, par ex.: des ménages ou des entreprises (le nombre de périodes étant normalement beaucoup plus petit que le nombre d'unités). La forme d'analyse économique qui utilise des données de ce genre est devenue très populaire ces dernières années. On peut expliquer l'intérêt grandissant que l'on porte à l'étude analytique des données de panel ainsi que l'importance croissante de ce genre d'étude par le fait que dans des enquêtes comme l'enquête de l'IFO menée auprès des entreprises, on peut recueillir des données sur des variables qui ne font pas l'objet de statistiques officielles, par exemple des données sur les projets, les évaluations, les prévisions, les innovations, etc.; voir, par exemple, Vogler (1977), Anderson et Strigel (1981), Oppenländer et Poser (1989).

Les données de panel présentent des avantages évidents par rapport aux données transversales ou aux données chronologiques; voir, par exemple, Hsiao (1986) ou Ronning (1991) pour une analyse plus détaillée:

- les séries de données de panel renferment normalement un grand nombre d'observations, ce qui accroît considérablement l'efficience de l'estimation de paramètres économétriques;

- le fait d'utiliser des données individuelles plutôt que des données agrégées élimine une bonne partie des problèmes de multicollinéarité qui se posent ordinairement dans l'analyse économétrique;

- comparativement aux données transversales, les données longitudinales permettent d'analyser les caractéristiques dynamiques du comportement des entreprises ou des ménages;

- les données longitudinales favorisent l'examen de problèmes économiques que les données chronologiques ne permettent pas d'envisager, si ce n'est avec des hypothèses a priori plutôt strictes, car les données

[1] G. Nerb et H. Seitz, ifo Institut für Wirtschaftsforschung e.V., Postfach 860460, 8000 München 86.

longitudinales recueillies dans des enquêtes fournissent de l'information sur des variables qui ne font pas l'objet de statistiques officielles (par ex.: prévisions);

- grâce aux données longitudinales, on peut éviter le problème d'agrégation qui caractérise les études empiriques fondées exclusivement sur des données chronologiques; voir, par exemple, Seitz (1992);

- en ce qui a trait à l'interdépendance de la théorie économique et de la recherche économique appliquée basée sur la théorie, les données longitudinales permettent d'examiner les théories économiques au niveau où sont formulées ces théories, c'est-à-dire au niveau du ménage ou de l'entreprise pris individuellement; voir Nerlove (1983).

2. MÉTHODES DE MODÉLISATION POUR DONNÉES DE PANEL

Avant de faire une brève analyse des méthodes de modélisation qui ont été élaborées pour les données longitudinales, nous allons définir sommairement les types de données qui sont recueillies dans des enquêtes. En pratique, toutes les données qui sont recueillies dans les diverses enquêtes de l'IFO, et dans d'autres enquêtes aussi, appartiennent à l'une ou l'autre des catégories suivantes:

i) *Variables continues:* la collecte de données continues dans des enquêtes est plutôt exceptionnelle. Voici des exemples de données continues recueillies dans les enquêtes de l'IFO: taux d'utilisation de la capacité, volume des stocks, carnets de commandes, etc.

iia) *Variables discrètes: variables dichotomiques (0,1):* données dichotomiques ou binaires, par exemple les réponses que donnent les entreprises à des questions qui visent à déterminer si elles ont mis au point un nouveau procédé ou un nouveau produit, si elles prévoient accroître le nombre de leurs employés, etc.

iib) *Variables discrètes: variables trichotomiques (+,=,-):* la plupart des données recueillies par l'IFO, plus spécialement dans l'enquête menée auprès des entreprises, sont trichotomiques, c'est-à-dire que les entreprises indiquent si la valeur d'une variable particulière a augmenté (+), a diminué (-) ou est demeurée à peu près la même (=) ou si elle est censée prendre l'une ou l'autre de ces directions.

En fait, il existe plus que deux catégories de variables discrètes mais la plupart des enquêtes mettent en évidence uniquement les variables binaires ou les variables trichotomiques.

2.1 Méthode de l'agrégation

Dans cette méthode, les micro-données ne sont pas utilisées pour l'estimation mais servent plutôt à constituer des séries chronologiques agrégées. En ce qui concerne les données continues, on peut, par exemple, calculer des estimations du taux moyen d'utilisation de la capacité, du volume moyen des stocks, des carnets de commandes, etc. à l'aide des résultats de l'enquête de l'IFO menée auprès des entreprises. Les données binaires peuvent servir à établir des estimations de la proportion d'entreprises qui répondent soit «oui» ou «non» à des questions particulières, par exemple la proportion d'entreprises qui ont réussi à mettre au point un nouveau produit ou un nouveau procédé. En ce qui a trait aux variables trichotomiques, les notions les plus courantes sont celles de la «fréquence marginale» et de la «fréquence de compensation». La fréquence marginale est le pourcentage d'entreprises qui répondent '+', =' ou '-' à une question particulière, tandis que la fréquence de compensation est simplement la différence entre la proportion d'entreprises qui ont répondu '+' et la proportion d'entreprises qui ont répondu '-'.

Les données trichotomiques sur les prévisions et les projets relatifs aux activités de production, les variations de prix, etc. qui sont recueillies dans de nombreuses enquêtes de l'IFO servent à produire des variables instrumentales quantitatives pour des séries chronologiques de variables de prévision. La méthode pertinente, appelée «technique de quantification» (pour une analyse de cette méthode, voir Seitz, 1989a), a été élaborée par Anderson et Theil dans les années cinquante; voir Anderson (1951, 1953) et Theil (1966). Elle a été modifiée récemment par Ronning (1986), Seitz (1987, 1988) et Tödter et Werfel (1988), pour n'en nommer que quelques-uns qui l'ont ensuite appliqué à des données de panel de l'IFO. Par ailleurs, les données trichotomiques sont beaucoup utilisées pour la construction d'indicateurs de conjoncture avancés; voir, par exemple, Dormayer et Lindlbauer (1984), Nerb (1989) et Entorf (1990, 1991).

2.2 Analyse de tableaux de contingence et modèles probabilistes log-linéaires

L'analyse de tableaux de contingence et l'estimation de modèles probabilistes log-linéaires sont les premières méthodes à avoir été élaborées pour tirer profit des micro-données qualitatives. Un tableau de contingence fournit simultanément la répartition des unités statistiques dénombrées suivant les modalités de deux variables discrètes ou plus. Il existe des méthodes statistiques qui permettent de calculer des mesures du degré d'association entre les variables qualitatives pertinentes; pour une étude approfondie des diverses mesures du degré d'association, voir Reynolds (1977).

À une étape ultérieure de la recherche, on a amélioré l'analyse (descriptive) des tableaux de contingence par l'introduction de l'estimation de modèles probabilistes log-linéaires. Cette dernière méthode sert à convertir les éléments des tableaux de contingence en probabilités et à estimer des modèles paramétriques afin d'expliquer la concomitance de variables particulières, par exemple la probabilité que les entreprises qui font face à un accroissement de la demande augmentent le prix de leurs extrants de même que le volume de leur production. Pour une présentation méthodique du modèle probabiliste log-linéaire, se référer aux monographies de Nerlove et Press (1976) et de Fienberg (1977), et pour une brève étude préliminaire qui utilise des données de l'enquête de l'IFO menée auprès des entreprises, voir l'article de König, Nerlove et Oudiz (1982).

2.3 Méthode d'estimation par panel (méthode des microdonnées)

Comme on l'a montré plus haut, la méthode de l'agrégation enlève presque entièrement aux données de panel leur micro-caractère et l'utilisation des modèles probabilistes log-linéaires présente de sérieuses limites à l'égard de l'interprétation économique des résultats de l'estimation. La méthode d'estimation par panel est la seule technique économétrique qui exploite entièrement le caractère longitudinal et l'individualité des données d'enquête. Elle permet d'envisager simultanément les effets propres à l'unité et la dynamique du comportement individuel. Grâce aux recherches récentes en économétrie, on a pu créer un large éventail de techniques d'estimation qui utilisent des données de panel tant pour des variables continues que pour des variables discrètes.

3. RECHERCHE APPLIQUÉE FAITE À L'AIDE DE DONNÉES DE PANEL DE L'IFO

Il existe un très grand nombre d'ouvrages où l'on parle de recherche économique appliquée fondée sur des données de l'enquête de l'IFO menée auprès des entreprises et de l'enquête de l'IFO sur l'innovation. La liste suivante indique les grands thèmes sur lesquels portent les articles et les projets de recherche:

- modélisation de prévisions et de plans;
- vérification de théories économiques et examen des réponses concernant les prix, la production et les stocks;
- analyse des activités de R.-D. et de la structure du marché et évaluation des répercussions sur le marché du travail;
- données de panel de l'IFO et la construction de méso-indicateurs.

3.1 Modélisation de prévisions et de plans

Il est indispensable de savoir comment se fait l'élaboration de plans et de prévisions pour bien comprendre le fonctionnement du système économique, car les mesures prises aujourd'hui dépendent des prévisions que l'on fait sur l'évolution des variables étudiées. Dans une série d'articles, König et Nerlove (1980, 1983) et König (1979, 1980) ont étudié l'élaboration des prévisions relatives aux prix en se servant de micro-données de l'enquête de l'IFO menée auprès des entreprises. Non seulement ces auteurs ont cherché à déterminer par quel mécanisme on peut le mieux décrire l'élaboration de prévisions, mais aussi ils ont étudié l'élaboration combinée de prévisions relatives aux prix et à la production, faisant ainsi du modèle d'élaboration de prévisions à une variable un modèle multidimensionnel. Zimmermann (1986, 1988) a lui aussi utilisé des micro-données pour étudier les propriétés intrinsèques et la rationalité des prévisions des entreprises, et a établi une comparaison entre des entreprises françaises et des entreprises allemandes. Les résultats de ses études autorisent fortement à penser que les prévisions des entreprises sont biaisées; de fait, elles sont biaisées vers le bas, c'est-à-dire que les entreprises font des prévisions trop pessimistes.

3.2 Vérification de théories économiques et examen des réponses concernant les prix, la production et les stocks

Bon nombre des variables que recouvrent les diverses enquêtes par panel de l'IFO fournissent de l'information sur les prévisions, les plans et les jugements des entreprises, tous des éléments au sujet desquels les organismes de statistique ne recueillent pas de données. Voilà donc l'occasion de tester des théories économiques nouvellement échafaudées, comme les modèles de déséquilibre ou les modèles des anticipations rationnelles.

König et Zimmermann (1983) et Kawasaki, McMillan et Zimmermann (1982, 1983) ont testé des théories modernes du comportement de l'entreprise qui tiennent compte de phénomènes de déséquilibre comme les restrictions du marché et l'inélasticité des prix. L'étude des théories du comportement de l'entreprise fondées sur le principe de déséquilibre nécessite des renseignements sur la condition des entreprises sur les marchés des intrants et des extrants. Dans les articles mentionnés ci-dessus, les auteurs ont montré qu'il était possible d'obtenir de tels renseignements à partir des données d'enquêtes-entreprises et d'utiliser avec succès ces renseignements dans le processus d'estimation.

Outre la vérification d'hypothèses sur le comportement de l'entreprise, il s'est fait beaucoup de recherche appliquée sur les stocks. L'enquête de l'IFO menée auprès des entreprises comporte deux questions qui ont trait aux stocks. À chaque mois, on demande aux entreprises si elles considèrent que leur stock du moment est trop élevé, raisonnable ou trop faible (question qualitative sur les stocks). En outre, depuis 1981 l'IFO demande aux entreprises d'indiquer à tous les trois mois le volume de leur stock de produits finis, exprimé en équivalents de semaines de production courante (question quantitative sur les stocks). Des questions semblables sont posées en ce qui regarde les carnets de commandes. On a utilisé amplement les deux types de variable, qualitative et quantitative, dans les études empiriques. À l'une des premières étapes de la recherche, König et Nerlove ont introduit les deux types de variable dans des modèles probabilistes log-linéaires et ont ajouté les stocks et les carnets de commandes à la série d'instruments de réponse des entreprises; voir, par exemple, König et Nerlove (1984, 1986). König et Seitz (1989, 1991) ont estimé des modèles à effets fixes simultanés pour panel qui tiennent compte de la production (mesurée par l'utilisation de la capacité), des stocks et des carnets de commandes (mesurés à l'aide des données quantitatives trimestrielles) et qui expliquent ces trois variables au moyen d'une série de variables exogènes et de caractéristiques dynamiques.

3.3 Analyse des activités de R.-D.: répercussions sur la structure du marché et le marché du travail

L'IFO recueille périodiquement des données sur les activités d'innovation des entreprises par une enquête spéciale sur l'innovation mais aussi par l'enquête régulière menée auprès des entreprises. Ces données sont de plus en plus recherchées dans l'analyse des activités de R.-D. La plupart des articles dans lesquels on utilise les données de l'enquête de l'IFO sur l'innovation traitent notamment la question de l'incidence des activités de R.-D. sur l'emploi et la structure du marché; pour un examen détaillé de l'utilisation des données de l'enquête de l'IFO dans les analyses portant sur l'innovation et l'emploi, voir Zimmermann (1990).

König et Zimmermann (1986) ont fait une analyse théorique et empirique des déterminants de l'innovation et ont examiné le rapport de celle-ci avec la taille de l'entreprise et le pouvoir de marché (mesuré par des indices de concentration ou l'indice d'Herfindahl). Les résultats de leur étude révèlent une forte corrélation positive entre l'innovation et les indices de concentration ainsi qu'une corrélation négative entre la taille de l'entreprise et la mise au point de nouveaux procédés. Cependant, l'étude de König et Zimmermann ne permet pas de tirer des conclusions sur le rapport qui peut exister entre la taille de l'entreprise et la création de nouveaux produits car la méthode d'estimation utilisée par ces auteurs n'a pas permis de définir les paramètres qui décrivent cette relation. Laisney, Lechner et Pohlmeier (1992a) tiennent compte entièrement de l'aspect longitudinal. Ces auteurs considèrent un modèle d'activités d'innovation en se servant d'un panel équilibré à cinq cycles de 1325 entreprises tiré de l'enquête de l'IFO menée auprès des entreprises pour la période de 1984 à 1988. Dans leur article, ils examinent un estimateur probit transversal de même qu'un estimateur probit longitudinal et ils arrivent à la conclusion que le second surclasse le premier.

3.4 Méso-indicateurs économiques tirés de données longitudinales de l'IFO

Comme le fait la Banque mondiale (1989) pour les indicateurs sociaux, on peut calculer aussi des *méso-indicateurs* à l'aide de données d'enquête. Les méso-indicateurs se rapportent à des groupes ou à des segments particuliers de la société ou de l'économie.

Il est possible de segmenter les entreprises selon une théorie quelconque et de construire les méso-indicateurs appropriés grâce à des questions spéciales incluses dans l'enquête de l'IFO menée auprès des entreprises; voir, par exemple, Nerb (1992). Cette approche emprunte des concepts à la théorie du déséquilibre; voir, par exemple, Malinvaud (1980). La théorie du déséquilibre classe les entreprises en fonction d'une combinaison de situations auxquelles elles font face sur les marchés des intrants et des extrants, ce qui donne les types de régime suivants: le '*régime classique*' (c'est-à-dire que la production n'est pas limitée par le niveau de la demande et les prix des intrants, par exemple les salaires réels, sont trop élevés pour stimuler la production); le '*régime keynésien*' (la production est limitée par le faible niveau de la demande); le '*régime d'inflation latente*' (la production est limitée par le caractère tendu des marchés d'intrants -- par exemple, pénurie de main-d'oeuvre -- tandis que la demande de biens demeure excédentaire); et le '*régime de sous-consommation*', plutôt improbable, (la production est limitée par le faible niveau de la demande et par une pénurie d'intrants).

Les données qui sont recueillies dans l'enquête de l'IFO menée auprès des entreprises peuvent servir à élaborer une typologie des caractéristiques d'entreprises qui se rapprochent sensiblement de ces concepts de déséquilibre. À l'aide de données longitudinales, on peut suivre l'évolution des divers groupes tout le long du cycle économique et, par conséquent, se faire une meilleure idée de la nature des cycles économiques et décider des moyens d'intervention à prendre pour combattre le chômage ou l'inflation.

Étant donné qu'à l'heure actuelle, la principale préoccupation économique de l'État allemand est le développement économique dans l'ancienne Allemagne de l'Est et que les économistes du monde entier suivent de près ce développement, nous allons illustrer brièvement les méso-indicateurs en comparant le développement économique en Allemagne de l'Est à celui en Allemagne de l'Ouest au moyen de méso-indicateurs du déséquilibre. Pour calculer ces indicateurs, nous créons les quatre grands groupes suivants, ainsi que six sous-groupes pour l'Allemagne de l'Ouest seulement:[2]

Groupe 1: «Faiblesse de la demande».

Sous-groupe 1.1 Faiblesse temporaire de la demande.
Sous-groupe 1.2 Faiblesse permanente de la demande.

Groupe 2: «Équilibre», c'est-à-dire aucun obstacle du côté de l'offre ou de la demande.

Sous-groupe 2.1 plus: situation économique considérée comme bonne ou acceptable.
Sous-groupe 2.2 plus: situation économique considérée comme mauvaise.

Groupe 3: «Engorgement de l'offre».

Sous-groupe 3.1 Engorgement temporaire de l'offre.
Sous-groupe 3.2 Engorgement marqué de l'offre.

Groupe 4: «Faiblesse de la demande et engorgement de l'offre».

Les figures 1 et 2 présentent les résultats de cette classification. À ce jour, la situation économique dans les Neue Länder n'a connu qu'une légère amélioration. En juillet 1992, 28% des entreprises industrielles ont indiqué que rien, du côté de l'offre comme du côté de la demande, ne faisait obstacle à leur production; cette proportion était de 4% en juillet 1990 et est passée à 17% en juillet 1991. La proportion des entreprises qui attribuent leurs

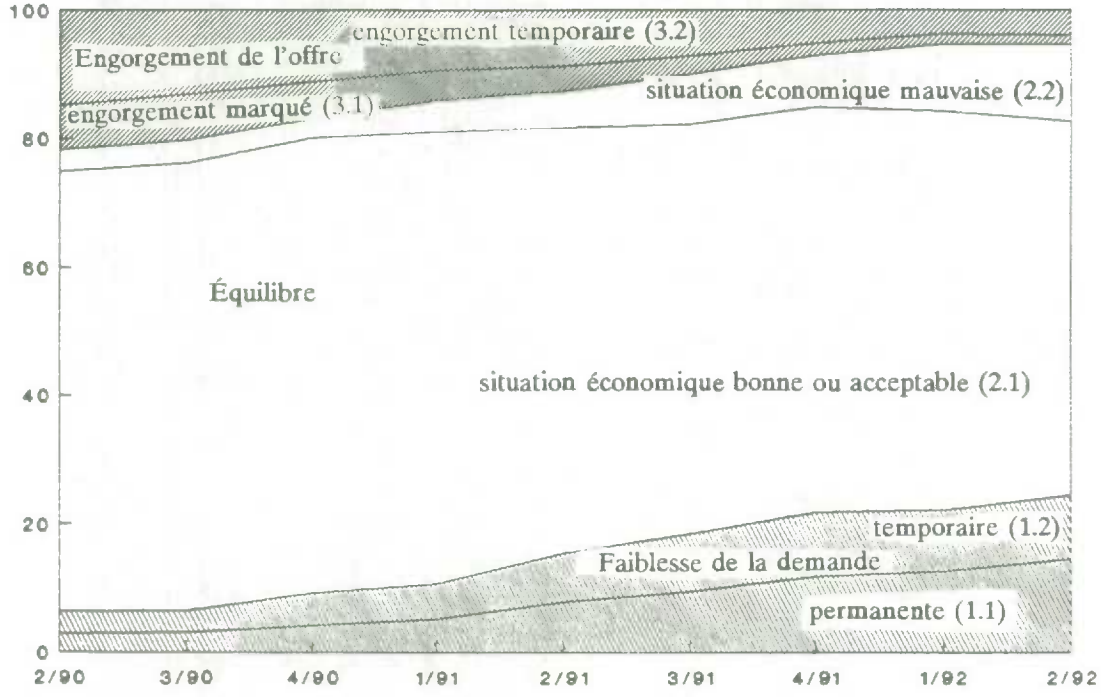
[2] Pour une description détaillée de cette classification, se référer à la version intégrale de cet article, que l'on peut se procurer en s'adressant aux auteurs.

difficultés exclusivement au faible niveau de la demande est passée de 29 % en juillet 1990 à 11 % en juillet 1992. On observe une tendance inverse chez les entreprises qui éprouvent des difficultés uniquement du côté de l'offre (20 % en juillet 1990 contre 33 % en juillet 1992; les principales sources de difficulté: programme de produits déficient, financement). Les entreprises qui formaient le plus gros groupe en juillet 1990 -- soit près de la moitié de la population -- étaient celles qui éprouvaient des difficultés aussi bien du côté de l'offre que du côté de la demande; depuis, la proportion représentée par ce groupe est tombée à 28 %. Une série aussi variée de réponses dans les enquêtes-entreprises n'est observée habituellement que pour les économies en voie de développement et les économies nouvellement industrialisées. Cela prouve que les problèmes économiques que connaît actuellement l'ex-RDA sont plus structurels que cycliques. Dans ces circonstances, il convient d'exclure l'idée d'un programme de relance classique. En ce qui concerne l'ancienne Allemagne de l'Ouest, près des trois quarts des entreprises font encore partie du groupe 2 (aucun obstacle à la production) malgré le ralentissement économique observé actuellement. Une infime proportion d'entreprises éprouvent de sérieuses difficultés du côté de l'offre comme du côté de la demande (moins de 1 % des entreprises Ouest-allemandes). La proportion d'entreprises Ouest-allemandes qui disent éprouver des difficultés du côté de l'offre est passée de 22 % en juillet 1990 à 5 % en juillet 1992, ce qui dénote le fléchissement de la demande en Allemagne de l'Ouest.

4. CONCLUSIONS

Les données de panel, comme les données de l'enquête de l'IFO menée auprès des entreprises ou celles de l'enquête sur l'innovation, sont un outil efficace pour la recherche économique appliquée. Des économistes du monde entier en viennent à reconnaître les limites et les faiblesses des modèles chronologiques. Comme ce sont des individus et non une entité fictive quelconque qui prennent des décisions économiques, nous devons intégrer l'individu -- l'entreprise ou le ménage -- dans nos modèles de comportement économique. L'évolution rapide des techniques de l'informatique et l'existence de logiciels conviviaux standard font que l'estimation de modèles à données longitudinales n'est pas plus coûteuse ni plus complexe que l'était il y a quinze ans l'estimation de modèles chronologiques. Toutefois, quant à savoir si l'analyse de micro-données est utile pour l'évaluation des conséquences sur le plan de l'action, beaucoup de recherches restent à faire dans un proche avenir.

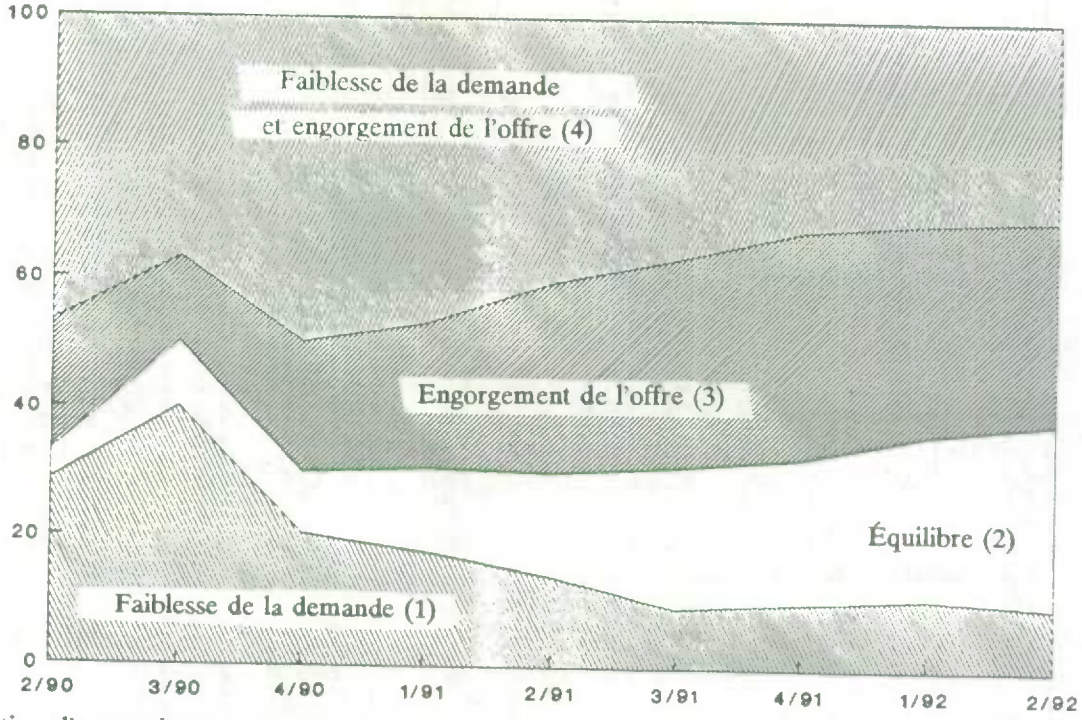
Figure 1: Types de régime pour les entreprises industrielles
Allemagne de l'Ouest / Total de l'industrie.



Proportion d'entreprises en %; trimestres

Source: Enquête de l'IFO auprès des entreprises

Figure 2: Types de régime pour les entreprises industrielles
Allemagne de l'Est / Total de l'industrie.



Proportion d'entreprises en %; trimestres

Source: Enquête de l'IFO auprès des entreprises

BIBLIOGRAPHIE

Anderson, O. (1951). *«Möglichkeiten und Grenzen einer Quantifizierung des Konjunkturtests» des Münchener ifo Institutes für Wirtschaftsforschung*, dans: Mitteilungsblatt für Mathematische Statistik, 3, 206-212.

Anderson, O. (1953). The business test of the IFO-Institut for economic research, Munich and its theoretical model. *Revue de L'institut International de Statistique*, 20, 1-17.

Anderson, O., et Strigel, W.H. (1981). Business surveys and economic research - A review of significant developments. H. Laumer et M. Ziegler (éds.). *International Research on Business Cycle Surveys*, 25-54.

Dormayer, H.-J., et Lindlbauer, J.D. (1984). Sectoral indicators by use of survey data. K.H. Oppenländer et G. Poser (éds.). *Leading Indicators and Business Cycle Surveys*, 467-498.

Entorf, H. (1990). *Multisektorale Konjunkturanalyse*, Campus Verlag.

Entorf, H. (1991). Das Ifo-Geschäftsklima, seine Komponenten und die Konjunkturprognose: Eine Regressionsstudie. *Ifo-Studien*, 37, 141-149.

Fienberg, S.E. (1977). *The Analysis of Cross-Classified Categorical Data*. Cambridge: The MIT Press.

Hsiao, C. (1986). *Analysis of Panel Data*, Cambridge University Press.

Kawasaki, S., McMillan, J., et Zimmermann, K.F. (1982). Disequilibrium dynamics: An empirical study. *American Economic Review*, 72, 992-1004.

Kawasaki, S., McMillan, J., et Zimmermann, K.F. (1983). Inventories and price inflexibility. *Econometrica*, 51.

Kõnig, H. (1979). Zur Bildung von Preiserwartungen: Ein log-lineares multivariates Wahrscheinlichkeitsmodel. *Kyklos*, 32, 380-391.

Kõnig, H. (1980). Über den mikroõkonomischen Zusammenhang zwischen Preiserwartungen und -realisationen. D. Duwendag et H. Siebert (éds.). *Politik und Markt, Festschrift für Hans Karl Schneider, Stuttgart.*

Kõnig, H., et Nerlove, M. (1980). Micro-analysis of realisations, plans and expectations in the Ifo-business test by multivariate log-linear probability models. W.H. Strigel (éd.). *Business Cycle Analysis, Westmead.*

Kõnig, H., et Nerlove M. (1983). Response of prices and production to unanticipated demand shocks: Some microeconomic evidence. K.H. Oppenländer et G. Poser (éds.). *Leading Indicators and Business Cycle Surveys*, 349-384.

Kõnig, H., et Nerlove, M. (1984). A recursive log-linear probability model of production plans and price anticipation. An empirical investigation for French and German firms, D. Vitry et B. Marechal (Eds.). *Emploi-Chômage Modélisation et Analyses Quantitative, Collection de l'Institute de Mathématiques Économiques*, 28.

Kõnig, H., et Nerlove, M. (1986). Price flexibility,inventory behavior and production responses. W. Heller, R. Storr et D. Starrett. *Equilibrium Analysis, Essays in Honor of Kenneth J. Arrow*, II, 179-218.

Kõnig, H., Nerlove, M., et Oudiz, G. (1982). Die Analyse mikroõkonomischer Konjunkturtest-Daten mit loglinearen Wahrscheinlichkeitsmodellen: Eine Einführung, *Ifo-Studien*, 28, 1, 155-191.

Kõnig, H., et Seitz, H. (1989). Zur Transmission von Nachfrage- und Kostenschocks auf Lagerhaltung, Preise und Produktion. *Jahrbücher für Nationalöknomie und Statistik*, 206, 421-433.

Kõnig, H., et Seitz, H. (1991). Production and price smoothing by inventory adjustment. *Empirical Economics*, 16, 233-252.

Kõnig, H., et Zimmermann, K.F.(1983). Mikroõkonomische Preis- und Produktionsplanung im Ungleichgewicht. H. Enke, W. Köhler et W. Schulz (Eds.). *Struktur und Dynamik der Wirtschaft*, 147-160.

Kõnig, H., et Zimmermann, K.F.(1986). Innovations, market structure and market dynamics. *Zeitschrift für die gesamte Staatswissenschaft*, 142, 184-199.

Laisney, F., Lechner, M., et Pohlmeier, W. (1992a). Innovation activity and firm heterogeneity: Empirical evidence from Germany. À paraître dans: *Economic Dynamics And Structural Change.*

Malinvaud, E. (1980). Macroeconomic rationing of employment. E. Malinvaud et J.P. Fitoussi, *Unemployment in Western Countries, MacMillan, 1980.*

Nerb, G. (1989). Zusammengesetzte Indikatoren und Indikatorsysteme, Chapter IV.2.1. K.H. Oppenländer et G. Poser (éds.). *Handbuch der IFO-Umfragen.*

Nerb, G. (1992). Neuer Ansatz zur Analyse von Konjunkturtestdaten. *CIRET-Studie Nr. 44* (forthcoming).

Nerlove, M., et Press, J. (1976). Multivariate log-linear probability models for the analyses of qualitative data. Discussion Paper No. 1, Center for Statistics and Probability, Northwestern University, Evanston.

Nerlove, M. (1983). Expectations, plans, and realizations in theory and practice. *Econometrica*, 51, 1251-1279.

Oppenländer, K. H., et Poser, G. (1989). Handbuch der IFO-Umfragen. Duncker & Humblot, Berlin 1989.

Reynolds, H. T. (1977). The analyses of cross-classifications. New York: The Free Press.

Ronning, G. (1986). Econometric approaches to the estimation of indifference intervals in business tendency surveys. K.H. Oppenländer et G. Poser (éds.). *Business Cycle Surveys in the Assessment of Economic Activity, Gower, Westmead (England)*, 175-209.

Ronning, G. (1991). *Mikroökonometrie*, Springer-Verlag Heidelberg.

Seitz, H. (1987). The estimation of inflation forecasts from business survey data. *Applied Economics*, 20, 427-438.

Seitz, H. (1988). An investigation into the reliability of business survey data. Discussion-Paper No. 358-387, Institut für Volkswirtschaftslehre und Statistik der Universität Mannheim.

Seitz, H. (1989a). Die Quantifizierung von Tendenbefragungs-daten: Ein Überblick, *Ifo-Studien*, 35, 1, 1-26.

Seitz, H. (1992). Still more on the speed of adjustment in inventory models: A lesson in aggregation. Discussion-Paper No. 377-388, Institut für Volkswirtschaftslehre und Statistik der Universität Mannheim. À paraître dans: *Empirical Economics*.

Theil, H. (1966). Applied economic forecasting, North-Holland, Amsterdam.

Toedter, K.-H., et Werfel, M.C. (1989). Quantification of indifference responses from business surveys with mixed data. À paraître dans: *IFO-Studien*.

Vogler, K. (1977). Content and determinants of judgemental and expectational variables in the Ifo business survey. W. H. Strigel (éd.). *Problems and Instruments of Business Cycle Analysis*, 73-114.

Zimmermann, K. F. (1986). On rationality of business expectations: A micro-analysis of qualitative responses. *Empirical Economics 11*, 23-40.

Zimmermann, K. F. (1988). Prognosequalität von Surveydaten: Mikroökonomische Evidenz. W. Franz, W. Gaab et J. Wolters (éds.). *Theoretische und angewandte Wirtschaftsforschung*, 261-274.

Zimmermann, K. F. (1990). Der IFO-Konjunkturtest in der arbeits- und industrieökonomischen Forschung. *IFO-Studien*, 36, 1-16.

SESSION 8

Analyse de données I

Recueil du Symposium 1992 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

MESURE DE LA ROBUSTESSE DES BARRIÈRES À L'ENTRÉE

J.R. Baldwin et M. Rafiquzzaman[1]

RÉSUMÉ

Des données longitudinales d'enquête par panel provenant du secteur manufacturier canadien sont utilisées pour modéliser le processus d'entrée et pour évaluer la présence de barrières à l'entrée. Dans cette communication, nous examinons la robustesse de résultats antérieurs 1) en utilisant diverses méthodes d'estimation, 2) en testant différentes spécifications de modèle et 3) en faisant varier les mesures utilisées pour évaluer l'importance de l'entrée.

MOTS CLÉS: Barrières à l'entrée; données de dénombrement; loi binomiale négative.

1. INTRODUCTION

Depuis les travaux fructueux de Bain (1956) et de Modigliani (1958), une grande attention a été portée par les économistes à la question de l'existence de barrières à l'entrée. Les modèles les plus largement utilisés sont ceux fondés sur le «prix limite»; ces modèles supposent que les niveaux de bénéfices au-delà desquels l'entrée est favorisée varient d'une industrie à l'autre et sont fonction des barrières à l'entrée. Les entreprises en place, dans des secteurs où les barrières sont élevées, peuvent hausser les prix sans favoriser l'entrée. Le niveau de prix au-dessus duquel il se produit une entrée est le prix limite.

Le modèle du «prix limite» suppose implicitement que la présence d'un entrant (nouveau venu dans une industrie) entraîne une augmentation de la production existante. À l'opposé, la vision de l'entrée fondée sur un «remplacement stochastique» se fonde sur l'hypothèse que l'entrée est un processus dynamique qui comporte le remplacement partiel ou complet d'entreprises existantes par des entrants (Baldwin et Gorecki, 1983). Cette perception axée sur le «remplacement» suppose qu'il peut se produire une entrée même lorsque le prix est égal au coût moyen à long terme et que les bénéfices de l'industrie sont nuls.

Les modèles du prix limite prétendent confirmer l'existence de barrière à l'entrée et, par conséquent, l'existence d'imperfections du marché[2]. Dans les modèles qui combinent le phénomène du remplacement stochastique et celui du prix limite, toutefois, les barrières à l'entrée revêtent une importance beaucoup moins grande (Baldwin et Gorecki, 1987).

Comme c'est souvent le cas en économique appliquée, l'interprétation de l'importance de ces différences est rendue plus complexe par le fait que les études antérieures diffèrent non seulement dans le choix du modèle, mais aussi dans leurs méthodes de mesure de l'entrée -- unités d'observation, unités de mesure, type d'entrants et périodes.

La présente communication cherche à déterminer dans quelle mesure l'importance observée des barrières à l'entrée dépend de la façon dont l'entrée est mesurée. À cette fin, on évalue jusqu'à quel point les techniques d'estimation, la spécification du modèle et les mesures modifient la conclusion selon laquelle l'entrée est entravée par certaines caractéristiques structurelles de l'industrie.

---

[1] J.R. Baldwin et M. Rafiquzzaman, Division d'analyse des entreprises et du marché du travail, Direction des études analytiques, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

[2] Voir Cable et Schwalbach (1991) pour une revue des résultats de modèles du type de celui de Orr.

2. MESURE DE L'ENTRÉE

L'entrée peut être définie comme la naissance d'une unité de production -- une nouvelle usine ou une nouvelle entreprise. Dans le premier cas, une entrée est définie comme une nouvelle usine dans une industrie particulière. Dans le second cas, elle est définie comme une nouvelle entreprise avec une nouvelle unité de production -- une entrée entièrement nouvelle. L'entrée peut aussi être définie comme la naissance d'une nouvelle personne juridique. Il se peut que de nouvelles personnes juridiques soient associées à la naissance d'usines, mais il se peut aussi que des entreprises entrent dans une industrie en faisant l'acquisition d'entreprises existantes. L'entrée peut être définie par une valeur brute ou par une valeur nette. Dans le premier cas, il s'agit du nombre total d'usines (ou d'entreprises) qui entrent dans une industrie au cours d'une période donnée. Dans le second cas, c'est la variation du nombre d'usines (ou d'entreprises) observée entre deux périodes différentes.

Dans la présente communication, nous utilisons une définition de l'entrée qui est implicite dans les travaux fondés sur le prix limite -- celle des entreprises entièrement nouvelles. Il s'agit d'une catégorie relativement homogène. La définition utilise les entreprises entièrement nouvelles plutôt que la somme des entreprises entièrement nouvelles et des entreprises résultant de fusions, car ces dernières n'engendrent pas, initialement du moins, d'augmentation de la production. Elle est axée sur les mesures brutes, plutôt que sur les mesures nettes qui tiennent compte non seulement de l'entrée, mais aussi de la sortie d'entreprises. Elle mesure l'entrée en termes de nouvelles entreprises qui construisent de nouvelles usines et non en termes de nouvelles usines comme telles. Cette dernière catégorie englobe à la fois les entreprises entièrement nouvelles et les usines qui sont construites par des entreprises déjà en place. Si l'on omet de faire la distinction entre les nouvelles entreprises et l'établissement de nouvelles usines par les entreprises déjà en place, on confond l'entrée avec les décisions d'expansion des entreprises en place. Des résultats (Baldwin et Gorecki, 1983, 1987) révèlent d'importantes disparités entre les facteurs déterminants des entreprises entièrement nouvelles et de celles résultant de fusions, des entreprises entièrement nouvelles et du processus de création d'usines par les entreprises existantes, et des mesures brutes et des mesures nettes de l'entrée.

L'importance des entreprises entièrement nouvelles est mesurée ici aussi bien du point de vue de leur nombre que du point de vue de leur taille. L'intensité des forces concurrentielles liées à l'entrée dépend probablement à la fois du nombre d'entrants et de la part du marché saisie par les entrants. Le pourcentage des livraisons d'un marché saisies par les entrants est égal au nombre de nouvelles entreprises multiplié par la taille moyenne des entrants, par rapport à l'ensemble des livraisons. Afin de mettre à l'épreuve la robustesse des résultats, les trois mesures ont été utilisées -- le nombre d'entrants, les livraisons des entrants et la taille moyenne des entrants -- comme variables dépendantes dans l'analyse de régression.

L'importance de l'entrée pour le processus concurrentiel et la performance du marché est également évaluée dans la présente communication aussi bien avec des données à court terme qu'avec des données à long terme. Les estimations à court terme ont été établies en fonction de deux points du temps voisins l'un de l'autre, tandis que les estimations à long terme se fondent sur deux points du temps éloignés l'un de l'autre. L'entrée à court terme a été estimée pour chaque année entre 1970 et 1979, puis une moyenne a été calculée. L'entrée à long terme a été calculée sous forme du nombre d'entreprises existant en 1979 qui ont fait leur entrée dans l'industrie depuis 1970. Il s'agit de la somme de toutes les entreprises ayant fait leur entrée chaque année à compter de 1970, moins les sorties d'entrants survenues au cours de cette période. Les entrées à court terme et à long terme ont été estimées au niveau des codes de 4 chiffres de la Classification type des industries (CTI) pour le secteur manufacturier canadien, au moyen d'une base de données longitudinale reflétant l'évolution des entreprises et des usines entre 1970 et 1979. Une description du fichier est donnée dans Baldwin et Gorecki (1990).

Enfin, une quatrième mesure -- le taux de succès des entrants -- a été calculée et soumise à une régression par rapport au même ensemble de variables explicatives que celles utilisées pour l'évaluation de l'entrée en termes du nombre d'entreprises, des livraisons et de la taille moyenne. Le taux de succès est défini comme la valeur de l'entrée à long terme divisée par la somme des valeurs de l'entrée à court terme. Il s'agit de la proportion de tous les nouvelles entreprises ayant fait leur entrée sur une période de dix ans qui existent toujours à la fin de la période; c'est une mesure directe de la survie de la population ou une mesure inverse de la disparition des entrants.

3. MODÈLES DE L'ENTRÉE

Les modèles de l'entrée les plus courants se fondent sur les travaux antérieurs de Orr (1974), qui supposent qu'une entrée se produira si les bénéfices sont supérieurs aux niveaux interdisant l'entrée. Sous l'angle adopté par Orr, le modèle s'exprime ainsi:

$$E_{it} = f(P_{it} - P^*_{it}), \tag{1}$$

où E_{it} est l'entrée dans l'industrie i au temps t, P_{it} est le bénéfice que l'entrant croit pouvoir réaliser après son entrée et P^*_{it} est le niveau de bénéfice interdisant l'entrée dans l'industrie i au temps t.

Le niveau de bénéfice interdisant l'entrée, P^*_{it}, dépend d'un vecteur de barrières à l'entrée, B, et d'une variable de risque du marché, R. P^* (sans indices liés au temps et à l'industrie) peut être écrit sous la forme $P^* = h(B,R)$. Par conséquent, le modèle de l'entrée présenté en (1) pour être ainsi exprimé :

$$E = f_1(P,B,R). \tag{2}$$

L'on s'attend à ce que E varie positivement en fonction du bénéfice attendu, P, et négativement par rapport à chaque composante de B et de R. Ce modèle fait donc l'hypothèse que le bénéfice favorise l'entrée, tandis que les barrières à l'entrée et le risque la réduisent.

Dans notre estimation, la rentabilité attendue (P_{it}) est représentée par deux variables. La première (PR) est la rentabilité moyenne des entreprises qui se maintiennent. Puisque la moyenne n'indique rien quant à la tendance susceptible d'influer sur les attentes de bénéfices futurs, la croissance des bénéfices (GP) sur l'ensemble de la période est également incluse. L'on s'attend à ce que l'entrée soit supérieure sur les marchés dont les bénéfices sont en croissance.

Les variables représentant les barrières à l'entrée (B) sont les économies d'échelle (MES), la concentration (CON), le degré de publicité (AD) et l'ampleur de la recherche et du développement (RD). Le risque du marché (R) est représenté par la volatilité de la croissance du marché (VMG).

L'équation (2) offre une description incomplète de l'entrée, car elle ne tient pas compte des aspects stochastiques de l'entrée. Selon l'approche fondée sur le «remplacement stochastique», une part importante de l'entrée est simplement attribuable à un remplacement des entreprises existantes, et survient même si la rentabilité économique est nulle. On suppose, selon cette approche, que le degré d'entrée dû au remplacement est fonction de la taille du marché. Si l'entrée est mesurée en terme de nombre d'entreprises, la taille du marché (S) est considérée comme étant le nombre d'entreprises de l'industrie (N). Si les livraisons des entrants sont utilisées comme variable dépendante, la taille du marché est mesurée en termes du total des livraisons de l'industrie (TVS). L'effet de la variable exprimant la taille du marché peut être mis en interaction avec les variables représentant les barrières à l'entrée, afin qu'on puisse déterminer si l'ampleur du remplacement stochastique est modifiée par les barrières à l'entrée.

Le nombre d'entrées devrait aussi dépendre de la facilité avec laquelle les entrants peuvent accéder à l'industrie et se tailler une part du marché. Un marché en croissance rapide se caractérise par un élargissement de la clientèle, et donc par une plus grande probabilité que les nouvelles entreprises accroîtront leur part de marché. Par conséquent, la croissance de l'industrie, G, est ajoutée à l'équation 2. Conformément à l'approche de Baldwin et Gorecki (1983, 1987), un modèle de l'entrée tenant compte à la fois de la vision du remplacement stochastique et de celle du prix limite peut être ainsi exprimé:

$$E = g(S,G,P,B,R). \tag{3}$$

S, G, et P sont des facteurs favorables à l'entrée, tandis que B et R tendent à l'entraver.

4. MÉTHODES D'ESTIMATION CONVENANT AUX DÉNOMBREMENTS

Beaucoup d'études empiriques de l'entrée réalisées antérieurement ont estimé une version linéaire et/ou log-linéaire de l'équation (3) ou une expression voisine, et se sont fondées principalement sur la méthode d'estimation des moindres carrés ordinaires (MCO). Puisque les données relatives à l'entrée sont des valeurs entières et s'écartent des hypothèses classiques de la régression, la spécification statistique de l'entrée exige le recours à une distribution de probabilité discrète. Pour répondre à cette exigence, nous définissons et estimons un modèle économétrique de l'entrée fondé d'abord sur l'hypothèse que chaque observation est tirée d'une distribution de Poisson, puis sur celle d'un tirage d'une distribution binomiale négative. Notre méthode s'apparente à celles de Hausman, Hall et Griliches (1984), et de Cameron et Trivedi (1986), qui utilisent à la fois la régression de Poisson et la régression binomiale négative pour des dénombrements des demandes de brevets des entreprises et de la demande de soins de santé des consommateurs, respectivement. Elle est également conforme à la démarche de Chappell et coll. (1990), Mayer et Chappell (1992) et Papke (1991), qui utilisent des régressions de Poisson pour étudier le phénomène de l'entrée dans diverses industries aux États-Unis.

Dans l'hypothèse où les données sur l'entrée sont tirées d'une distribution de Poisson, la probabilité d'obtenir un nombre d'entreprises, E_i, reflétant l'entrée dans l'industrie i est donnée par :

$$Pr(E_i) = Exp(-\lambda_i)\lambda_i^{E_i}/E_i!, \quad E_i = 0,1,2,..., \tag{4}$$

La moyenne et la variance de E_i sont égales à λ_i. Pour incorporer les variables explicatives, X_i, qui influent sur l'entrée, le paramètre λ_i est ainsi défini :

$$\lambda_i = E(E_i|X_i) = Exp(X_i\beta), \tag{5}$$

où $X_i = (S,G,P,B,R)$ et β est un vecteur de paramètres devant être estimé. La contrainte imposant l'égalité de la moyenne et de la variance en vertu de la loi de Poisson peut être surmontée; il suffit d'adopter la démarche de Gourieroux, Monfort et Trognon (1984a,b) et d'utiliser une version particulière de la loi binomiale négative[3].

Initialement, les régressions des MCO, de Poisson et de la loi binomiale négative ont toutes été estimées. Les résultats sont présentés au tableau 1. Les erreurs-types estimées en vertu des modèles de Poisson et binomial négatif sont sensiblement moins élevées que celles résultant de la méthode des MCO. Ces résultats sont conformes à ceux de Hausman, Hall et Griliches, ainsi que de Cameron et Trivedi.

Bien que les estimations ponctuelles selon le modèle de Poisson et le modèle binomial négatif soient de signes et de grandeurs analogues, les erreurs-types estimées avec le modèle de Poisson sont sensiblement plus faibles.

Pour tester l'hypothèse nulle selon laquelle le modèle sous-jacent est un modèle de Poisson par opposition à l'alternative selon laquelle le modèle répond à une loi binomiale négative, nous avons utilisé à la fois le test de Wald et le test du rapport des vraisemblances. Les statistiques des deux tests se sont révélées hautement significatives. Les données ont également permis de rejeter l'hypothèse d'égalité de la moyenne et de la variance, qui est la caractéristique fondamentale du modèle de Poisson. En conséquence, le modèle de Poisson a été rejeté en faveur du modèle binomial négatif.

5. RÉSULTATS

Dans la première section, nous examinons dans quelle mesure l'effet des barrières à l'entrée sur le nombre d'entrants est robuste selon que c'est le modèle du prix limite ou celui du remplacement stochastique qui est utilisé. La deuxième section présente une comparaison des effets des barrières à l'entrée sur diverses mesures de l'entrée; quatre variables dépendantes différentes sont utilisées: nombre d'entrants, livraisons des entrants, taille moyenne des entrants et taux de succès des entrants.

---

[3] Nous utilisons une forme particulière de la loi binomiale négative, dont la moyenne est égale à Exp(Xß) et dont la variance est égale à Exp(Xß)[1 + α Exp(Xß)]. Voir Cameron et Trivedi (1986) pour plus de détails.

Tableau 1: Comparison des méthodes d'estimation du modèle de l'entrée [1], [2]

	LONG TERME			COURT TERME		
	MCO	POISSON	BINOMIALE NÉGATIVE	MCO	POISSON	BINOMIALE NÉGATIVE
Constante	- 4.995 (0.500) [7.380]	3.587 (0.000) [0.087]	2.848 (0.000) [0.327]	- 73.218 (0.338) [76.190]	6.508 (0.000) [0.022]	6.005 (0.000) [0.359]
N	0.294 (0.000) [0.015]	0.003 (0.000) [0.000]	0.005 (0.000) [0.043]	2.482 (0.000) [0.066]	0.001 (0.000) [0.000]	0.003 (0.000) [0.002]
PR	1.479 (0.598) [2.795]	0.121 (0.000) [0.025]	0.094 (0.520) [0.146]	9.946 (0.744) [30.390]	0.108 (0.000) [0.008]	0.145 (0.510) [0.221]
GP	0.166 (0.911) [1.480]	- 0.045 (0.023) [0.020]	- 0.062 (0.372) [0.070]	- 2.598 (0.872) [16.030]	- 0.054 (0.000) [0.005]	- 0.080 (0.263) [0.072]
GS	0.883 (0.046) [0.439]	0.083 (0.000) [0.006]	0.090 (0.000) [0.022]	14.653 (0.002) [4.675]	0.081 (0.000) [0.002]	0.089 (0.000) [0.019]
CON	- 0.074 (0.427) [0.093]	- 0.023 (0.000) [0.001]	- 0.018 (0.000) [0.004]	- 1.280 (0.196) [0.986]	- 0.027 (0.000) [0.000]	- 0.030 (0.000) [0.004]
MES	1.466 (0.952) [24.060]	- 2.971 (0.000) [0.490]	- 2.038 (0.033) [0.955]	- 77.137 (0.768) [260.718]	- 5.684 (0.000) [0.174]	- 2.158 (0.014) [0.876]
RD	0.00004 (0.999) [0.074]	0.006 (0.000) [0.001]	0.005 (0.362) [0.006]	1.191 (0.141) [0.806]	0.011 (0.000) [0.000]	0.013 (0.005) [0.005]
AD	- 77.680 (0.311) [76.360]	- 9.904 (0.000) [1.357]	- 3.080 (0.399) [3.649]	- 515.700 (0.533) [824.740]	- 6.297 (0.000) [0.335]	- 0.408 (0.902) [3.298]
VMG	0.084 (0.018) [0.035]	0.003 (0.000) [0.001]	0.004 (0.012) [0.002]	1.247 (0.001) [0.382]	0.003 (0.000) [0.000]	0.003 (0.046) [0.002]
Paramètre de variance α			0.372 (0.000) [0.043]			0.487 (0.000) [0.057]
Adj R^2	0.81			0.93		
- Log L		1016.099	591.998		8711.917	997.528

[1] Les niveaux de signification d'un test bilatéral visant à rejeter l'hypothèse nulle selon laquelle le coefficient est zéro sont donnés entre parenthèses.
[2] les erreurs-types associées aux estimations sont indiquées entre crochets.

5.1 Comparaison entre les modèles du prix limite et du remplacement stochastique

Le choix de la régression binomiale négative plutôt que celle des MCO a permis de renverser des observations antérieures (Baldwin et Gorecki, 1987) selon lesquelles les barrières à l'entrée sont des facteurs déterminants positifs, mais non significatifs, du processus d'entrée. Avec un tel modèle fondé sur des valeurs entières, l'effet des variables «concentration» et «économies d'échelle» présente un degré de signification sensiblement plus élevé.

La méthode des MCO produit trois variables significatives, tant à court terme qu'à long terme (voir le tableau 1). Le processus d'entrée est relié positivement au nombre existant d'entreprises (N), à la croissance des livraisons (GS) et au risque, mesuré en termes de volatilité de la croissance (VMG). Les autres variables ne sont pas statistiquement significatives.

Chaque variable qui est significativement[4] différente de zéro dans l'estimation des MCO est également significative, et de même signe, dans la régression binomiale négative. En outre, la régression binomiale négative comprend, parmi les barrières à l'entrée, deux variables ayant un coefficient significatif et négatif, tant à court terme qu'à long terme, lesquelles n'étaient pas significatives dans l'équation des MCO. Il s'agit des économies d'échelle (MES) et de la concentration (CON). La recherche et développement a un effet positif sur l'entrée tant à long terme qu'à court terme, mais cet effet n'est significatif qu'à court terme. La publicité n'est pas significative, aussi bien dans la régression des MCO que dans le modèle binomial négatif.

Ces conclusions ne changent pas même si l'on fait varier la spécification du modèle. Trois variantes sont présentées au tableau 2. Comme au tableau 1, les colonnes 1 à 3 représentent les résultats à long terme, tandis que les colonnes 4 à 6 représentent les résultats à court terme. La première équation n'utilise que les variables provenant d'un modèle simple du type de celui de Orr, c'est-à-dire: rentabilité (PR), croissance de la rentabilité (GP), croissance des ventes (GS), concentration (CON), économies d'échelle (MES), recherche et développement (RD), degré de publicité (AD) et variabilité de la demande (VMG). La deuxième formulation comporte en outre la taille de l'industrie -- nombre d'entreprises (N). Dans la troisième formulation, des termes d'interaction entre la taille de l'industrie et les barrières à l'entrée -- concentration (CON), degré de publicité (AD), ampleur de la recherche et développement (RD) et économies d'échelle (MES) -- sont ajoutés.

Les barrières à l'entrée dont l'effet est significatif dans le modèle simple du prix limite (colonnes 1 et 4) ont également un effet significatif dans les deux modèles incorporant le phénomène de remplacement stochastique (colonnes 2 et 3; colonnes 5 et 6). La concentration (CON) et les économies d'échelle (MES) ont un effet négatif sur l'entrée dans tous les modèles.

Toutefois, en ce qui a trait à la concentration, le coefficient diminue d'environ 50 % lorsque le modèle de remplacement stochastique des colonnes 2 et 5 est utilisé. De plus, dans la troisième variante (colonnes 3 et 6), le fait que les économies d'échelle aient un coefficient positif lorsqu'elles sont mises en interaction avec le nombre d'entreprises signifie que l'effet des économies d'échelle diminue à mesure qu'augmente le nombre d'entreprises de l'industrie. En fait, dans le cas des résultats à long terme, quand le nombre d'entreprises (N) est supérieur à 30, les économies d'échelle n'ont pas de répercussions négatives sur l'entrée. Si l'on fait le même exercice pour la concentration, le point critique est d'environ 47 entreprises. Ces barrières jouent donc un rôle, mais pas quand le nombre d'entreprises est relativement élevé.

Il est intéressant de noter, par ailleurs, que la publicité est faiblement significative lorsqu'elle est mise en interaction avec la taille d'une industrie. Le taux de remplacement stochastique est moins élevé dans les industries ayant des ratios publicité-ventes élevés. Ce résultat contraste singulièrement avec celui de la première colonne, où l'on n'observe aucun effet significatif de cette variable, incluse pour représenter une barrière à l'entrée influant sur le niveau limite des bénéfices des entrants.

---

[4] Dans la présente communication, 5 % est utilisé comme seuil de signification.

Tableau 2: Comparison de différents modèles de l'entrée : Estimation de la régression binomiale négative[1],[2].

	(1)	(2)	(3)	(4)	(5)	(6)
Constante	4.561 (0.000) [0.292]	2.848 (0.000) [0.327]	2.498 (0.000) [0.257]	7.944 (0.000) [0.284]	6.005 (0.000) [0.359]	5.409 (0.000) [0.303]
PR	0.083 (0.534) [0.133]	0.094 (0.520) [0.146]	0.064 (0.626) [0.131]	0.031 (0.846) [0.160]	0.145 (0.510) [0.221]	0.162 (0.389) [0.188]
GP	- 0.107 (0.195) [0.083]	- 0.062 (0.372) [0.070]	- 0.071 (0.256) [0.062]	- 0.103 (0.220) [0.084]	- 0.080 (0.263) [0.072]	- 0.073 (0.257) [0.065]
GS	0.131 (0.000) [0.025]	- 0.090 (0.000) [0.022]	- 0.056 (0.001) [0.017]	- 0.108 (0.000) [0.022]	- 0.089 (0.000) [0.192]	0.052 (0.002) [0.016]
CON	- 0.036 (0.000) [0.004]	- 0.018 (0.000) [0.004]	- 0.014 (0.002) [0.005]	- 0.049 (0.000) [0.004]	- 0.030 (0.000) [0.004]	- 0.025 (0.000) [0.004]
MES	- 2.585 (0.032) [1.210]	- 2.038 (0.033) [0.955]	- 6.935 (0.000) [1.414]	- 2.160 (0.039) [1.047]	- 2.158 (0.014) [0.876]	- 5.642 (0.000) [1.082]
RD	- 0.009 (0.185) [0.006]	0.005 (0.362) [0.006]	0.005 (0.411) [0.006]	0.016 (0.001) [0.005]	0.013 (0.005) [0.005]	0.005 (0.147) [0.004]
AD	- 6.206 (0.116) [3.952]	- 3.080 (0.400) [3.649]	0.356 (0.941) [4.839]	- 4.552 (0.231) [3.803]	- 0.408 (0.902) [3.298]	- 0.696 (0.882) [4.695]
VMG	0.002 (0.370) [0.021]	0.004 (0.012) [0.002]	0.004 (0.002) [0.001]	- 0.001 (0.471) [0.002]	0.003 (0.046) [0.001]	- 0.004 (0.005) [0.001]
N		0.005 (0.000) [0.001]	0.004 (0.000) [0.001]		0.003 (0.000) [0.0002]	0.002 (0.002) [0.0005]
CON X N			0.00003 (0.487) [0.00005]			0.00002 (0.368) [0.00002]
AD X N			- 0.076 (0.064) [0.041]			- 0.015 (0.636) [0.031]
RD X N			- 0.0007 (0.313) [0.0007]			0.000003 (0.967) [0.00008]
MES X N			0.231 (0.000) [0.042]			0.144 (0.000) [0.024]
Paramètre de variance α	0.565 (0.000) [0.065]	0.372 (0.000) [0.043]	0.201 (0.000) [0.026]	0.713 (0.000) [0.086]	0.487 (0.000) [0.057]	0.311 (0.000) [0.039]
- Log L	620.304	591.998	556.235	1033.470	997.528	958.751

[1] Les niveaux de signification d'un test bilatéral visant à rejeter l'hypothèse nulle selon laquelle le coefficient est zéro sont donnés entre parenthèses.

[2] Les erreurs-types associées aux estimations sont indiquées entre crochets.

5.2 Autres mesures de l'importance de l'entrée

Dans la section précédente, l'entrée est mesurée sous forme du nombre de nouvelles entreprises. Dans la présente section, on se demande si les variables déterminantes de l'entrée demeurent les mêmes quand une autre unité de mesure est utilisée pour définir l'entrée. Quatre mesures différentes de l'entrée sont utilisées pour comparer la robustesse de nos conclusions sur l'importance des barrières à l'entrée. Dans chaque cas, on examine aussi bien l'entrée à long terme que l'entrée à court terme.

Ces mesures sont:

(1) le nombre d'entrants entièrement nouveaux (E),

(2) la quantité des livraisons des entrants (TVSE),

(3) la taille moyenne des entrants (ASE), et

(4) le ratio du nombre d'entrants à long terme au nombre d'entrants à court terme (RATIO).

Chaque variable dépendante est soumise à une régression par rapport au même ensemble de variables explicatives, à une exception près. La variable de normalisation pour le nombre d'entrants est le nombre d'entreprises de l'industrie (N); pour les livraisons des entrants (TVSE), c'est la valeur totale des livraisons de l'industrie (TVS); pour la taille moyenne des entrants (ASE), c'est la taille moyenne des entreprises existantes (ASF); pour le taux d'entreprises qui se maintiennent (RATIO), c'est la taille moyenne des entrants par rapport à la taille moyenne de l'industrie (RELSIZE).

Le tableau 3 présente les résultats relatifs à l'entrée à long terme. Les résultats sont qualitativement les mêmes pour l'entrée à court terme. L'estimation a été faite selon la loi binomiale négative dans le cas du nombre d'entrants, et selon les MCO dans le cas des autres variables[5].

Une comparaison des équations relatives au nombre d'entrants et aux livraisons de l'industrie révèle que la première comporte un moindre pouvoir explicatif. Une régression simple des MCO visant le nombre d'entrants (non présentée ici) donne un R^2 ajusté beaucoup plus élevé que la régression des MCO visant les livraisons -- .81 contre .41. La raison est que les variables explicatives parviennent mal à décrire la taille moyenne des entrants. Le R^2 ajusté pour l'équation utilisant la taille moyenne des entrants comme variable dépendante n'était que de .32. Malgré cette différence, la plupart des coefficients significatifs des équations relatives au nombre d'entrants et aux livraisons révèlent des effets analogues. L'entrée est reliée positivement à la taille et à la croissance des livraisons, et négativement à la concentration.

Bien que ces variables aient essentiellement le même effet sur le nombre d'entrants et les livraisons, elles n'influent pas toujours de la même façon sur la taille moyenne des entrants (ASE). La croissance a un effet positif sur la taille moyenne des entrants (ASE) et le taux de succès (RATIO), mais dans aucun des deux cas la relation n'est très significative. La croissance, par conséquent, influe sur l'importance des entrants en raison de son effet sur le nombre d'entrants et non parce qu'elle facilite l'entrée d'entreprises de taille moyenne plus élevée. La croissance influe aussi positivement sur le taux de succès (RATIO), mais le coefficient n'est pas significatif.

La concentration a également un effet différent sur le nombre d'entrants et sur la taille moyenne. Une concentration plus forte se traduit par un nombre d'entrants moindre, mais elle a un effet positif, mais non significatif, sur la taille moyenne des entrants. Il y a donc moins d'entrants dans les industries concentrées, mais les entrants ont tendance à être de plus grande taille -- probablement parce que les inconvénients, sur le plan des coûts, de l'entrée d'entreprises de petite taille sont supérieurs dans ces industries. La concentration est également associée à un effet positif significatif sur la variable RATIO -- le taux de succès des entrants.

---

[5] La régression binomiale négative a également été utilisée pour les livraisons, et une transformation logistique de la taille relative et du taux de succès de la population a aussi été employée; dans tous les cas, les mêmes résultats qualitatifs ont été obtenus.

Tableau 3: Comparison de différentes mesures de l'entrée: À long terme[1],[2].

	E*		TVSE		ASE		RATIO	
Constante	2.923 [0.280]	(0.000)	19.207 [4.529]	(0.000)	8.837 [5.828]	(0.132)	0.051 [0.027]	(0.060)
PR	0.125 [0.152]	(0.411)	- 0.863 [2.090]	(0.680)	- 2.514 [2.686]	(0.351)	- 0.010 [0.012]	(0.383)
GP	- 0.033]0.063]	(0.602)	- 2.348 [1.152]	(0.043)	- 2.854 [1.487]	(0.057)	- 0.008 [0.007]	(0.246)
GS	0.082 [0.018]	(0.000)	1.024 [0.327]	(0.002)	0.621 [0.417]	(0.138)	0.003 [0.002]	(0.153)
CON	- 0.018 [0.004]	(0.000)	- 0.015 [0.062]	(0.015)	0.095 [0.082]	(0.249)	0.002 [0.000]	(0.000)
MES	- 1.806 [0.872]	(0.038)	- 27.703 [18.060]	(0.127)	- 45.760 [23.204[	(0.051)	- 0.086 [0.103]	(0.406)
RD	0.004 [0.005]	(0.377)	0.094 [0.056]	(0.095)	0.168 [0.071]	(0.020)	- 0.001 [0.000]	(0.032)
AD	- 3.103 [3.069]	(0.312)	- 78.861 [56.700]	(0.166)	- 55.002 [72.518]	(0.450)	- 0.308 [0.318]	(0.335)
VMG	0.003 [0.001]	(0.012)	0.006 [0.027]	(0.828)	0.007 [0.034]	(0.984)	0.0003 [0.000]	(0.096)
N	0.005 [0.0004]	(0.000)						
TVS			0.000004 [0.000]	(0.000)				
ASF					0.0001 [0.000]	(0.001)		
RELSIZE							- 0.00002 [0.000]	(0.195)
R^2 ajusté			0.41		0.32		0.25	
F			13.35		9.32		6.92	
Degrés de liberté			(9,148)		(9,148)		(9,148)	

[1] Les niveaux de signification d'un test bilatéral visant à rejeter l'hypothèse nulle selon zéro sont donnés entre parenthèses.
[2] Les erreurs-types associées aux estimations sont indiquées entre crochets.
* Exclut toutes les valeurs nulles de la variable dépendante.

Bien que les bénéfices -- qu'il s'agisse de PR ou de GP -- soient rarement significatifs, les coefficients de cette variable sont de signes différents dans l'équation visant le nombre d'entrants et dans celle visant les livraisons. Les bénéfices sont reliés positivement au nombre d'entrants, mais négativement aux livraisons. Cela s'explique par le fait que des bénéfices plus élevés permettent à un plus grand nombre d'entrants d'accéder à une industrie, mais que ces entrants sont de plus petite taille.

En bref, les barrières à l'entrée qui ont été si souvent mises en évidence dans la littérature se révèlent moins importantes quand l'impact sur l'entrée est mesuré autrement que par un simple compte des entrants. Il se peut qu'il y ait moins d'entrants dans les industries concentrées, mais que ces entrants soient de taille plus grande, de sorte que la concentration n'a pas un effet aussi grand sur les livraisons des entrants que sur leur nombre. De plus, les entrants des industries concentrées ont une plus grande capacité de se maintenir en exploitation. Le nombre d'entrants qui survivent est supérieur dans les industries concentrées.

6. CONCLUSIONS

Dans la présente communication, nous avons vérifié la robustesse de conclusions selon lesquelles certaines caractéristiques structurelles constituent des barrières à l'entrée. Comme c'est souvent le cas dans les tests de robustesse, nous avons appris non seulement l'effet de telle ou telle variable et l'existence de tel ou tel phénomène, mais aussi les circonstances dans lesquelles l'effet s'exerce. Nous avons constaté que lorsque nous utilisons une méthode d'estimation plus complexe -- régression pour des données de dénombrement -- l'effet des barrières à l'entrée est plus facilement isolé de celui d'autres variables. Nous avons vu également que l'utilisation d'un modèle élargi confirmait l'importance des barrières, mais limitait les conclusions aux industries ne comportant qu'un petit nombre d'entreprises. Cet exercice nous a donc permis de confirmer que les barrières ont un effet non linéaire qui ne se reproduit peut-être pas d'une industrie à l'autre. Enfin, nous avons vu que les barrières n'ont pas le même effet sur le nombre d'entrants que sur la taille moyenne des entrants, et donc sur la part de marché qu'obtiennent les entrants. Certaines barrières structurelles réduisent le nombre d'entrants, mais parfois, ce nombre plus faible est compensé par une taille moyenne plus élevée des entrants.

ANNEXE: DESCRIPTION DES VARIABLES

PR	Taux de rendement brut du stock de capital, défini comme le total de la valeur ajoutée du secteur d'activité moins le total de la valeur des traitements et salaires du secteur d'activité, en 1970.
GP	Ratio (1979 par rapport à 1970) du taux pondéré des bénéfices des plus grandes entreprises (moitié supérieure en termes d'emplois), le taux des bénéfices étant défini comme le ratio pondéré marges/ventes.
GS	Taux de croissance entre 1970 et 1979 de la valeur totale réelle des livraisons du secteur d'activité.
CON	Indice de concentration - 4 premières entreprises
MES	Part de marché (en livraisons) de la plus petite entreprise nécessaire pour englober 50 % de l'emploi de l'industrie.
RD	Ratio entre le nombre de personnes travaillant en recherche et développement et l'ensemble des salariés.
AD	Ratio publicité-ventes.
VMG	Volatilité de la croissance du marché, définie comme l'erreur-type des résidus provenant d'une régression du logarithme des livraisons par rapport au temps.
N	Nombre existant d'entreprises dans une industrie.

TVS Valeur de l'ensemble des livraisons de toutes les entreprises d'une industrie.

ASF Taille moyenne de l'ensemble des entreprises d'une industrie, en livraisons.

RELSIZE Taille moyenne des entrants par rapport à la taille moyenne de l'ensemble de l'industrie, en livraisons.

BIBLIOGRAPHIE

Bain, J.S. (1965). *Barriers to New Competition*. Cambridge, MA: Harvard University Press.

Baldwin, J.R., et Gorecki, P.K. (1983). Entry and exit to the canadian manufacturing sector: 1970-1979. Discussion Paper # 225. Ottawa: Economic Council of Canada.

Baldwin, J.R., et Gorecki, P.K. (1987). Plant creation versus plant acquisition: The entry process in Canadian manufacturing. *International Journal of Industrial Organization*, 5, 27-41.

Baldwin, J.R., et Gorecki, P.K. (1990). Mesure de l'entrée et de la sortie d'entreprises à l'aide de données longitudinales. *Recueil du symposium de 1989 sur l'analyse des données dans le temps*, (éds. A.C. Singh et P. Whitridge), Statistique Canada, Ottawa et l'université Carleton, 279-300.

Cable, J., et Schwalbach, J. (1991). International comparisons of entry and exit. *Entry and Market Contestability: An International Comparison*, (éds. P.A. Geroski et J. Schwalbach), Oxford: Blackwell, 1991, 257-281.

Cameron, A.C., et Trivedi, P.K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1, 29-54.

Chappell, W.F., Kimenyi, M.S., et Mayer, W.J. (1990). A Poisson probability model of entry and market structure with an application to U.S. industries during 1972-77. *Southern Economic Journal*, 56, 918-927.

Geroski, P.K., et Schwalbach, J. (1991). *Entry and Market Contestability: An International Comparison*. Oxford: Blackwell.

Gourieroux, C., Monfort, A., et Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52, 681-700.

Gourieroux, C., Monfort, A., et Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica*, 52, 701-720.

Hausman, J., Hall, B.H., et Griliches, Z. (1984). Econometric models for count data with an application to the Patents-R&D relationship. *Econometrica*, 52, 909-938.

Mayer, W.J., et Chappell, W.F. (1992). Determinants of entry and exit: An application of the compounded bivariate Poisson distribution to U.S. industries, 1972-1977. *Southern Economic Journal*, 58, 770-778.

Modigliani, F. (1958). New developments on the oligopoly front. *Journal of Political Economy*, 66, 215-232.

Orr, D. (1974). The determinants of entry: A study of the manufacturing industries. *Review of Economics and Statistics*, 56, 58-66.

Papke, L.E. (1991). Interstate business tax differentials and new firm location: Evidence from panel data. *Journal of Public Economics*, 45, 47-68.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

LE SUIVI D'ENFANTS DANS LE TEMPS: LE DÉVELOPPEMENT DES ENFANTS ET SES LIENS AVEC L'ÉVOLUTION DE LEUR SITUATION FAMILIALE, SOCIALE ET ÉCONOMIQUE

P.C. Baker et F.L. Mott[1]

RÉSUMÉ

À partir de données de l'enquête «National Longitudinal Survey of Youth», nous essayons de déterminer dans cette communication comment le faible revenu de la famille et l'emploi maternel peuvent influer sur le niveau des connaissances et le comportement d'un enfant. Pour l'analyse, nous avons recours à la méthode de la mesure du gain brut qui nous permet de comparer, dans un premier temps, les scores obtenus par les enfants lors de tests menés à deux moments successifs assez rapprochés (soit 1986 et 1988 ou 1988 et 1990) et aussi de suivre l'évolution de la situation familiale au cours de cet intervalle, en fonction des caractéristiques de départ des individus et des familles. Les résultats font ensuite l'objet d'une analyse sur une plus longue période, c'est-à-dire pour les trois moments d'enquête de la période 1986-1990, et semblent indiquer que les changements cognitifs et socio-émotifs observés chez les enfants peuvent varier en fonction de la durée de l'intervalle entre la mesure préliminaire et la mesure terminale et d'autres facteurs comme la race et le niveau de maturité des enfants.

MOTS CLÉS: Évaluation des enfants; changements familiaux; mesure du gain brut.

1. INTRODUCTION

Dans le but de mieux évaluer le lien entre la situation familiale et le développement de l'enfant, nous essayons dans cette communication de déterminer dans quelle mesure les facteurs familiaux et socio-économiques peuvent entraîner chez les enfants, des changements de comportement et de rendement à des tests courants d'aptitude en mathématiques et en lecture. Nombreuses sont les études qui ont montré l'incidence des caractères qualitatifs des parents sur le développement de l'intelligence de leurs enfants et qui se sont intéressées aux effets des différences de caractéristiques des parents tout autant qu'à leur capacité variable de fournir les bases nécessaires à l'apprentissage de leurs enfants. Des recherches récentes semblent démontrer que les familles qui peuvent fournir à leurs enfants une sécurité économique et un milieu de vie offrant des niveaux adéquats de stimulation intellectuelle et de soutien émotif exercent une influence sur le rendement à l'école et l'acquisition des aptitudes sociales de ces enfants (Parcel et Menaghan 1990).

Si d'autres chercheurs ont déjà étudié les liens entre les caractères qualitatifs des familles et les niveaux de rendement des enfants, peu ont tenté de déterminer dans quelle mesure les conditions familiales sont liées aux *changements* du bien-être des enfants. L'enquête «National Longitudinal Survey of Youth» (NLSY) et les données d'évaluation d'enfants qui en découlent offrent des possibilités uniques d'analyse des relations dans le temps entre les caractéristiques familiales et maternelles et le développement des enfants ayant fait l'objet de l'enquête. L'enquête NLSY est la source de mesures répétées du niveau des connaissances et du bien-être socio-émotif d'un échantillon de grande taille de jeunes Américains, mesures recueillies à trois moments, soit en 1986, 1988 et 1990. L'analyse présentée ici est d'abord fondée sur un échantillon groupé afin que nous puissions évaluer les changements entre deux points successifs dans le temps (de 1986 à 1988 ou de 1988 à 1990) dans les scores obtenus dans des tests administrés aux enfants ainsi que les changements de niveaux de bien-être socio-économique survenus dans le même intervalle, après que nous ayons neutralisé l'effet des caractéristiques familiales de départ. D'autres résultats, fondés sur un échantillon plus restreint et portant sur les données

[1] P.C. Baker et F.L. Mott, Center for Human Resource Research, 921 Chatham Lane, Suite 200, The Ohio State University, Columbus (Ohio), É.-U. 43-221-2418.

recueillies aux trois points de mesure dans le temps, soit l'ensemble de la période 1986-1990, révèlent que les changements cognitifs et socio-émotifs observés chez les enfants peuvent varier en fonction de la durée de l'intervalle entre la mesure préliminaire et la mesure terminale et en fonction d'autres facteurs comme le niveau de maturité des enfants.

2. MÉTHODE

2.1 Données sur les enfants recueillies grâce à l'enquête NLSY

Depuis 1979, l'enquête «National Longitudinal Survey of Youth» (NLSY) a permis de suivre un échantillon national de plus de 12 000 personnes au moyen d'interviews annuelles poussées. La cohorte, répartie à peu près également entre jeunes des deux sexes, constitue un échantillon national probabiliste de la population hors institutions âgée de 14 à 21 ans au cours de l'année de base. On a également constitué des échantillons superposés de jeunes militaires, de jeunes de race hispanique et de race noire et de jeunes blancs défavorisés économiquement[2]. L'enquête a permis de recueillir des données détaillées sur l'emploi, le niveau d'instruction, la formation et le vécu familial des répondants lors de leur passage de l'adolescence au monde adulte.

À compter de 1986, les enfants des répondants de sexe féminin de l'enquête NLSY ont été interviewés à intervalle de deux ans afin qu'on puisse mesurer l'intelligence générale, le développement socio-émotif et le milieu de vie de ces enfants. Les mesures varient selon l'âge des enfants, lesquels, au moment de l'interview de 1990, pouvaient être des nouveaux-nés aussi bien que des jeunes adolescents. Le taux de réponse s'est maintenu à plus de 90% à chacun des cycles de l'enquête.

2.2 Les échantillons utilisés pour l'analyse

L'analyse est fondée sur deux types d'échantillon des enfants, un échantillon groupé pour étudier l'évolution des résultats après deux ans et un échantillon plus restreint pour suivre les enfants sur une période de quatre ans. L'échantillon groupé était constitué de 2 010 enfants âgés d'au moins cinq ans au moment de l'interview de 1986 ou de 1988 et dont les scores aux trois tests menés en 1988 ou 1990 étaient valides. L'échantillon restreint comprenait 930 enfants âgés d'au moins cinq ans en 1986 auxquels on a fait passé les trois interviews, soit en 1986, 1988 et 1990. Comme les enfants de l'échantillon superposé des enfants blancs défavorisés n'ont pas été évalués en 1990, ils ont été éliminés de l'analyse des données fondées sur l'échantillon restreint de la période de quatre années.

En 1986, les mères de l'échantillon étaient âgées de 21 à 28 ans et, en 1990, de 25 à 32 ans. Étant donné que les enfants compris dans l'échantillon étaient âgés de 5 ans et plus en 1986, leur mère était relativement jeune au moment de leur naissance, soit une moyenne d'environ 19 ans dans le cas des mères noires et une moyenne de près de 20 ans pour les mères blanches. Les mères (et leurs enfants) *ne forment pas* tout à fait un échantillon transversal des mères et de leurs enfants, mais constituent plutôt un échantillon national de mères relativement jeunes et de leurs enfants[3]. Toutefois, ces mères sont loin d'être des valeurs aberrantes de la population; elles représentent vraiment une section transversale de la population des mères et de leurs enfants à un moment donné dans le temps, plusieurs années après la naissance de ces derniers, soit au moins sept au deuxième test.

---

[2] Des poids individuels ont été calculés pour chaque année d'enquête afin de rendre l'échantillon conforme aux totaux estimés de façon indépendante dans la population des personnes âgées de 14 à 21 ans au 1er janvier 1979. Les poids tiennent compte de la probabilité de sélection à la première interview, du taux de non-réponse lors de la mesure préliminaire et de la variation due à l'échantillonnage aléatoire. Les poids permettent la production d'estimations groupées de la population dans les totalisations.

[3] Les poids des enfants sont fondés sur les poids des mères et nous avons appliqué un facteur d'ajustement pour tenir compte des divers taux d'interview des enfants dans chacun des catégories d'âge, de race et de sexe. À cette fin, nous utilisons les chiffres des enfants dont on connaissait l'existence de même que des estimations des taux de fécondité des femmes sorties du champ de l'enquête. Cependant, en ce qui concerne les enfants ne se trouvant plus dans la base de sondage, nous n'avons pas procédé à un ajustement pour tenir compte des taux différents d'achèvement des tests.

2.3 Les mesures

Trois évaluations des connaissances et du comportement ont été sélectionnées dans les données de l'enquête NLSY de 1986, 1988 et 1990: une échelle de comportement remplie par toutes les mères d'enfants âgés de 4 ans et plus ainsi que deux tests de connaissances administrés aux enfants de 5 ans et plus. Les problèmes de comportement ont été mesurés par 28 points sur une échelle remplie par la mère et conçue pour l'évaluation de la nature et de la fréquence des problèmes de comportement observés chez les enfants dans les trois mois ayant précédé l'interview. Une augmentation des scores dans cette échelle est le signe de problèmes de comportement plus grands. On a également administré des sous-tests en mathématiques et en lecture du Peabody Individual Achievement Test (PIAT) [test de connaissances individuel de Peabody] pour évaluer les aptitudes des enfants dans ces domaines. Nous avons tenu compte de la différence de scores percentiles entre les deux points d'évaluation pour mesurer l'évolution des connaissances. Une augmentation des scores aux deux sous-tests PIAT témoigne d'un progrès. Les taux de non-réponse aux trois évaluations varie entre 10% et 15%, selon l'âge et la race ou l'origine ethnique des enfants (Baker et Mott 1989).

L'ensemble de données de l'enquête NLSY contient plusieurs mesures utiles à la définition opérationnelle des facteurs socio-économiques ou liés à l'emploi maternel qui, comme nous le supposons, auraient un effet sur les résultats des enfants. Le tableau 1 présente une liste de ces principaux facteurs et antécédents et des facteurs observés lors de la mesure préliminaire afin qu'ils servent de données de contrôle dans la détermination des caractéristiques actuelles des enfants, des mères et des familles.

Le changement de situation économique des familles est mesuré par la moyenne des niveaux de faible revenu pour l'ensemble de la période à l'étude, d'un point d'évaluation à l'autre. La variable représente le rapport du revenu total de l'unité famille au seuil de faible revenu officiel, lequel est défini comme le revenu permettant de satisfaire aux besoins essentiels de la famille, en fonction de la taille de la famille et de l'âge du chef du ménage. L'importance de l'emploi maternel est déterminée par la moyenne du total des semaines travaillées dans les années comprises entre la mesure préliminaire et la mesure terminale. Par exemple, nous nous sommes fondés sur le nombre de semaines pendant lesquelles les mères ont déclaré avoir travaillé en 1987 et 1988 au moment des interviews dont elles ont fait l'objet pour calculer la moyenne dont nous nous sommes servis pour les enfants évalués pour la première fois en 1986 et ensuite en 1988. Nous avons établi trois catégories: faible niveau d'emploi (moins de 20 semaines); niveau d'emploi moyen (de 20 à 39 semaines) et niveau d'emploi élevé (de 40 à 52 semaines). Le régime de travail à plein temps constitue la base de référence des équations.

Les ressources intellectuelles de la mère sont déterminées par son score au test AFQT (Armed Forces Qualification Test) [test d'admissibilité aux Forces armées], une mesure de ses capacités (Profile of American Youth 1982). Le niveau d'instruction de la mère, mesuré par le plus haut niveau de scolarité atteint au moment de l'interview de 1986, sert d'indicateur de la situation socio-économique antérieure de la famille. Des variables muettes, indiquant si la mère avait consommé de l'alcool ou fumé dans les douze mois ayant précédé la naissance de l'enfant, représentent un certain nombre de caractéristiques de «maternage». L'âge de la mère (en années) au moment de la naissance de chaque enfant est la variable dont nous nous sommes servis pour rendre compte de caractères qualitatifs antérieurs non observables. Le poids de l'enfant à la naissance (en onces) est un indicateur du développement potentiellement compromis que la mère a pu signaler au cours de la première interview après la naissance de l'enfant.

Deux mesures relatives à la composition des ménages ont été prises en compte: (1) le nombre d'années dans la période d'évaluation où un grand-parent a été présent dans le ménage dans lequel vivait l'enfant; et (2) le nombre d'années pendant lesquelles le mari ou le conjoint de la mère a été présent.

2.4 Modèle

En ce qui concerne les trois variables dépendantes, nous avons utilisé des modèles d'estimation simples dans lesquels les changements de résultats des enfants sont une fonction des niveaux de revenu de la famille et de l'emploi maternel dans la période intermédiaire, une fois neutralisé l'effet des caractéristiques des enfants et des mères observées lors de la mesure initiale (comme le poids à la naissance, les soins donnés par des personnes autres que la mère, le niveau d'instruction de la mère, l'âge à la naissance, les habitudes prénatales de la mère et la composition des ménages). Les caractéristiques pré-existantes de la famille ont ensuite été introduites dans

l'analyse pour neutraliser les effets de sélection que pourrait entraîner le seul examen des conditions familiales propres à la période d'évaluation.

Le choix du modèle le plus approprié à ce processus de changement posait des problèmes. Les caractéristiques d'un tel modèle ne satisfaisaient pas entièrement aux critères normalement associés au modèle de la mesure du gain brut, dans lequel $Y_2 - Y_1$ fait l'objet d'une régression en X, pas plus qu'à ceux du modèle de la variable explicative, dans lequel Y_2 fait l'objet d'une régression en Y_1 et en X (Allison 1990). Afin d'éviter de surestimer les effets possibles des changements ou encore de procéder à un sous-ajustement pour tenir compte des différences antérieures, nous avons décidé d'utiliser la méthode de la mesure du gain brut comme variable dépendante au lieu de la méthode fondée sur le modèle de la variable explicative. S'il est vrai qu'il peut y avoir des relations causales entre Y_1 et Y_2, rien ne nous oblige à supposer que les éléments caractéristiques de la période de Y_1 sont corrélés avec X. En posant comme hypothèse que les changements intermédiaires des conditions de la famille se sont produits après que la mesure préliminaire ait été prise, nous devrions pouvoir atténuer le problème des erreurs de mesure en Y_1. Enfin, comme un très grand nombre de recherches antérieures laissent supposer que l'évolution de la situation familiale et son incidence sur le développement des enfants peuvent être fort différentes selon qu'il s'agit d'enfants noirs ou d'enfants blancs, nous avons jugé bon de procéder à des analyses distinctes, après stratification de l'échantillon selon la race.

3. CONSTATATIONS

Le tableau 1 présente les caractéristiques moyennes de l'ensemble des variables explicatives et des variables de résultats utilisées dans l'analyse. En ce qui concerne les résultats en mathématiques comme les résultats de l'échelle des problèmes de comportement, la variation nette du score moyen au cours des deux années est très faible, mais il reste que de faibles changements nets masquent le plus souvent des changements bruts importants, à la hausse ou à la baisse, au niveau individuel. En revanche, dans l'échantillon global et en particulier chez les enfants noirs, nous constatons une diminution substantielle des résultats des tests de reconnaissance en lecture, d'un point d'évaluation à l'autre. Compte tenu du fait que ces résultats ont été comparés à ceux d'un échantillon national américain, ce déclin laisse entendre que les enfants de l'enquête NLSY avaient dans une certaine mesure perdu du terrain pendant leurs premières années d'école par rapport à leurs aptitudes en lecture antérieures et après comparaison avec les résultats d'un échantillon national transversal d'enfants américains des mêmes âges.

Tableau 1: Moyennes des principales statistiques fondées sur les variables explicatives et les variables de résultats, selon la race, échantillon groupé.

	NOIRS	NON-NOIRS
Différence de scores percentiles, test PIAT en lecture, mesure préliminaire et terminale	-8.7	-3.1
Différence de scores percentiles, test PIAT en mathématiques, mesure préliminaire et terminale	+0.4	-0.9
Différence de scores percentiles, problèmes de comportement, mesure préliminaire et terminale	+0.9	+1.9
% d'enfants n'ayant pas été élevés par la mère, trois premières années de vie	0.5	0.5
Poids à la naissance (onces)	108.7	116.6
% de mères comptant moins de 12 années d'études	39.6	46.2
% de mères comptant 12 années d'études	41.7	41.5
% de mères ayant consommé de l'alcool pendant la grossesse	31.5	41.2
% de mères ayant fait usage de tabac pendant la grossesse	32.1	42.1
Âge moyen de la mère à la naissance de l'enfant	18.8	19.6
% de mères ayant obtenu un score percentile inférieur à 50+ au test AFQT	91.7	70.2
Points d'enquête moyens, présence d'un grand-parent à la maison, mesure préliminaire et terminale	0.5	0.2
Points d'enquête moyens, présence du conjoint à la maison, mesure préliminaire et terminale	1.0	2.0
% de familles dont le ratio de pauvreté était inférieur à 1, mesure préliminaire et terminale	48.1	25.1
% de familles dont le ratio de pauvreté se situait entre 1 et 199, mesure préliminaire et terminale	25.3	30.8
% de mères ayant travaillé moins de 20 semaines par année, mesure préliminaire et mesure terminale	44.6	40.6
% de mères ayant travaillé entre 20 et 39 semaines par année, mesure préliminaire et mesure terminale	16.0	20.4
TAILLE DE L'ÉCHANTILLON	744	1 266

Compte tenu de leur relativement jeune âge, ces mères (et leurs familles) ont des caractéristiques qui ont pour effet de les désavantager dans une certaine mesure par rapport aux femmes américaines d'un échantillon national

transversal. À titre d'exemples, ces jeunes mères avaient moins de chances d'avoir terminé leurs études secondaires et d'avoir fréquenté le collège et risquaient davantage de vivre sous le seuil de la pauvreté. Les mères noires et leurs familles étaient pour leur part nettement plus défavorisées à cet égard que les mères blanches de l'étude.

3.1 Niveaux et changements de niveaux des résultats

Nous mettons ici l'accent sur l'examen des *changements* dans les niveaux des résultats par rapport à diverses variables explicatives, alors que la plupart des recherches à ce jour analysaient comment les *niveaux* des résultats étaient associés à divers facteurs familiaux et socio-économiques. Comme le montre le tableau 2, ces deux perspectives peuvent donner des résultats fort différents. L'équation relative aux «niveaux» utilise comme variable de résultat le score (percentile) obtenu à la dernière évaluation, tandis que dans l'équation relative aux «changements», le résultat correspond à la *différence* entre les deux scores percentiles (mesure préliminaire et mesure terminale).

Comme des chercheurs ont déjà démontré qu'il est très difficile d'établir un lien entre l'importance de l'emploi des mères et la réussite des enfants (Piotrkowski, Rapoport et Rapoport 1987), nous utilisons des variables muettes pour mesurer les deux facteurs en question et ainsi vérifier l'hypothèse d'un rapport linéaire entre ces facteurs et les résultats obtenus. Les catégories de référence omises sont celles qui ont trait à des revenus plus élevés (au moins le double du seuil de pauvreté fixé) et à un nombre de semaines d'emploi supérieur (au moins 40 semaines par année). Plutôt que de présenter tous les coefficients des équations complètes, nous mettons l'accent sur les coefficients qui présentent un intérêt pour l'étude, c'est-à-dire uniquement sur les coefficients relatifs à la pauvreté et à l'emploi maternel, une fois éliminés les effets des variables explicatives énoncées ci-dessus.

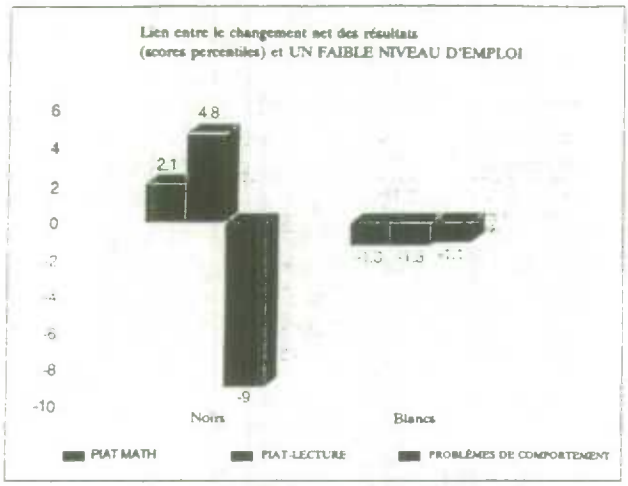
Tableau 2: Comparaison des effets de la pauvreté et de l'emploi maternel sur les niveaux et sur les changements de niveaux des résultats (scores percentiles) des enfants (moindres carrés ordinaires), échantillon groupé.

	PIAT MATHÉMATIQUES				PIAT LECTURE				PROBLÈMES DE COMPORTEMENT			
	Niveau		Changement		Niveau		Changement		Niveau		Changement	
Noirs												
Pauvreté												
Sous le seuil de pauvreté	-3.6	(2.9)	1.4	(3.2)	-5.6[c]	(3.3)	-3.4	(2.8)	3.6	(3.1)	4.2	(3.2)
Près du seuil de pauvreté	0.2	(2.7)	2.4	(3.0)	-2.7	(3.1)	-1.2	(2.6)	0.1	(2.8)	-3.6	(2.9)
Emploi maternel												
Mère travaillant moins de 20 semaines par année	-3.1	(2.3)	2.1	(2.5)	-1.3	(2.6)	4.8[b]	(2.3)	-2.1	(2.4)	-9.0[a]	(2.5)
Mère travaillant de 20 à 39 semaines par année	2.1	(2.6)	7.4[a]	(2.8)	3.4	(2.9)	7.2[a]	(2.5)	-3.9	(2.7)	-3.5	(2.8)
Non-Noirs												
Pauvreté												
Sous le seuil de pauvreté	-7.3[a]	(2.4)	0.6	(2.4)	-8.9[a]	(2.7)	-2.0	(2.3)	8.7[a]	(2.5)	1.2	(2.2)
Près du seuil de pauvreté	-6.3[a]	(1.7)	-0.1	(1.7)	-7.0[a]	(1.8)	-1.3	(1.6)	4.8[a]	(1.8)	-0.8	(1.6)
Emploi maternel												
Mère travaillant moins de 20 semaines par année	-2.0	(1.6)	-1.3	(1.6)	-0.6	(1.8)	-1.3	(1.6)	-3.2[c]	(1.7)	-1.1	(1.5)
Mère travaillant de 20 à 39 semaines par année	-5.4[c]	(1.9)	-5.5[a]	(1.9)	-0.4	(2.1)	3.2[c]	(1.8)	-4.1[b]	(2.)	1.3	(1.7)

Nota: Les effets des variables explicatives énoncées dans le tableau 1 ont été éliminés pour les variables relatives à la pauvreté et à l'emploi. Les variables relatives à la pauvreté et à l'emploi représentent des statistiques moyennes de la période d'évaluation. Nous avons exclu les catégories de référence suivantes: ratio de pauvreté $\geq$ 2.0 et emploi de la mère 40-52 semaines par année. Le ratio de pauvreté désigne le rapport du revenu total de la famille au seuil officiel de la pauvreté fixé, lequel est fondé sur la taille de la famille et l'âge du chef du ménage). Les données sont tirées d'équations distinctes pour les Noirs et pour les Non-Noirs.
a = coefficient significatif au niveau p $\leq$ .01; b = $\leq$ .05; c = $\leq$ .10.

Si nous examinons en premier lieu les liens entre la situation vis-à-vis de la pauvreté et les résultats des enfants, nous constatons, dans le cas des enfants noirs, qu'après avoir neutralisé l'effet de toutes les variables explicatives, le rapport entre la situation vis-à-vis de la pauvreté et les niveaux des scores pour tous les types de test est non significatif d'un point de vue statistique. Parallèlement à cette constatation, il *ne semble pas* que l'amélioration ou la détérioration des résultats de ces enfants au cours des deux années de référence ait un lien avec la situation vis-à-vis de la pauvreté de la famille de ces enfants. Par contre, dans le cas des enfants blancs, il existe une association importante entre la pauvreté, d'une part, et les niveaux de connaissance des enfants de même que les problèmes de comportement, d'autre part. Pour les trois types d'évaluation, nous remarquons des liens très importants entre le fait de vivre sous le seuil de la pauvreté (ou près de ce seuil) et de faibles scores en mathématiques ou en lecture ainsi qu'un niveau de problèmes de comportement supérieur à la moyenne.

Figure 1.



Cependant, comme pour les enfants noirs, nous ne pouvons démontrer qu'il existe un lien entre la situation vis-à-vis de la pauvreté et la détérioration des résultats dans le temps. Chez les enfants blancs comme chez les enfants noirs, les changements à court terme relatifs aux connaissances ou au comportement semblent liés à des facteurs autres que ceux représentés par les variables de substitution dans nos équations[4]. Il peut s'agir de facteurs reliés au milieu comme les caractéristiques du voisinage, le groupe d'amis et les caractéristiques de l'école fréquentée ou encore d'autres traits de caractère maternels ou familiaux non étudiés.

Le portrait de la situation change quelque peu lorsque nous passons de la pauvreté à l'emploi maternel. Chez les enfants noirs, il n'y a pas de relation entre l'importance de l'emploi maternel et le *niveau* de bien-être des enfants. Toutefois, comme on peut le voir clairement dans le graphique du bas de la figure 1, en ce qui concerne les deux résultats qui mesurent les connaissances, un niveau d'emploi moyen (de 20 à 39 semaines de travail par année) est associé à une amélioration des scores au bout des deux années à l'étude. Il est possible qu'on puisse attribuer le phénomène à un «équilibre» entre la quantité et la qualité du temps qu'une mère consacre à son enfant. Nous pouvons établir un lien entre le temps qu'une mère consacre à son enfant et les capacités intellectuelles de l'enfant, d'autant plus si la mère possède les capacités intellectuelles nécessaires à une telle interaction. Les compétences qu'une mère acquiert dans son travail, qu'elles aient ou non un lien avec la lecture ou les mathématiques, peuvent améliorer sa capacité d'enseigner à son enfant. Plus elle passe du temps avec son enfant, mieux elle est en mesure de fournir une telle formation. Dans le cas des familles noires, un niveau moyen d'emploi de la mère pourrait représenter l'équilibre optimal à atteindre à cet égard[5]. La situation est toute autre dans le cas des résultats relatifs aux problèmes de comportement, illustrée dans le graphique du haut de la figure 1. Il semble

[4] Soulignons que les équations comprennent deux autres variables de substitution de la famille, la première mesurant la présence du conjoint de la mère dans la maison et la deuxième la présence d'un grand-parent. Chez les enfants noirs, les variables de composition du ménage n'avaient pas d'effet significatif dans aucune des équations relatives aux changements; chez les enfants blancs, la présence d'un grand-parent est associée à des scores plus élevés en mathématiques.

[5] L'interaction entre la profession ou le niveau d'instruction et l'importance de l'emploi maternel pourrait permettre de clarifier les concepts distincts de qualité et quantité.

que chez les enfants noirs, nous pouvons associer le fait qu'une mère consacre plus de temps à son enfant à la maison (parce qu'elle travaille moins de 20 semaines par année) à une amélioration substantielle des problèmes de comportement (baisse du score). Une fois éliminés l'effet d'autres facteurs connexes (comme le niveau d'instruction et la présence d'autres membres de la famille), il semble que les enfants noirs dont la mère ne travaillent pas beaucoup sont nettement plus favorisés émotivement que les enfants noirs dont la mère travaille la majeure partie de l'année.

Aucun de ces effets dus à l'emploi n'est observé dans le cas des enfants blancs. En mathématiques, les enfants blancs sont désavantagés (tant du point de vue du niveau des scores que du changement de niveau des scores) lorsque leur mère travaille un nombre moyen de semaines, comparativement aux enfants dont la mère a un niveau élevé d'emploi (de 40 à 52 semaines de travail). Il est possible que cela soit attribuable à des biais de sélection différents pour les femmes blanches et noires qui travaillent. De plus, s'il est vrai qu'un niveau moyen d'emploi de la mère est associé à l'amélioration des capacités de lecture des enfants noirs, aucune tendance du genre ne peut être dégagée chez les enfants blancs. De plus, le fait que leur mère travaille moins influe très peu sur le comportement des enfants blancs; l'effet est si faible qu'on ne remarque pas de lien significatif entre l'importance de l'emploi de la mère et l'évolution du comportement de ces enfants blancs.

Des équations distinctes (non présentées ici) ont été utilisées pour évaluer l'effet de l'ensemble des variables de contrôle sur les coefficients des variables explicatives les plus importantes. Dans presque tous les cas, l'importance des coefficients ne change pas à la suite de l'inclusion (ou de l'exclusion) des variables de contrôle dans l'équation. Dans aucun des cas, l'ajout des variables de contrôle n'a eu pour effet de modifier de façon statistiquement significative l'importance d'un coefficient. Par conséquent, les changements de niveau des connaissances ou de comportement observés chez les enfants noirs dans le cas d'un niveau moyen d'emploi de la mère semblent indépendants des autres facteurs compris dans l'équation (du moins sous la forme précisée). Apparemment, l'absence de lien, dans les équations comprenant des variables de contrôle, entre la situation vis-à-vis de la pauvreté et les changements du niveau des connaissances ou du comportement des enfants noirs comme des enfants blancs, ne masque pas d'effets qui auraient pu se manifester entre deux variables dans les équations sans variables de contrôle.

3.2 Âge de l'enfant

Les auteurs de nombreuses études sur le développement de l'enfant laissent entendre que la capacité d'un enfant d'acquérir des aptitudes particulières ou de réagir à divers stimuli est sans doute étroitement liée à son stade de développement psychologique et intellectuel, lequel dépend à son tour de son âge physiologique ou chronologique. Comme les résultats d'un enfant sont souvent influencés par son âge, nous avons utilisé des équations par âge pour déterminer dans quelle mesure le niveau de maturité des enfants pouvait influer sur les relations entre la situation familiale et les changements de scores des enfants. Les résultats (non illustrés ici) révèlent que comme c'était le cas des relations globales entre la situation vis-à-vis de la pauvreté et les changements de résultats des enfants, l'étude selon l'âge ne fait pas ressortir de lien significatif. Pour les enfants blancs comme pour les enfants noirs, nous n'avons pas constaté de variation systématique selon l'âge pour ce qui est de la probabilité d'un effet de la situation vis-à-vis de la pauvreté sur l'évolution des niveaux des connaissances ou des problèmes de comportement.

En revanche, comme le montre le tableau 3, il pourrait de fait y avoir des variations importantes selon l'âge et la race en ce qui concerne l'influence de l'emploi maternel sur les changements de résultats des enfants[6]. Chez les enfants noirs, le fait que la mère travaille très peu à l'extérieur est associé à une amélioration des problèmes de comportement, quel que soit leur âge[7]. En outre, pour tous les enfants noirs sauf les plus âgés, le temps consacré à l'enfant à la maison est associé à une amélioration des capacités en lecture. Par contre, la relation globale entre un niveau d'emploi maternel moyen (20-39 semaines) et l'amélioration des résultats en

---

[6] Ce n'est qu'une supposition étant donné que nous n'avons pas fait de tests pour déterminer s'il existe des différences significatives du point de vue statistique entre les diverses catégories d'âge et de sexe.

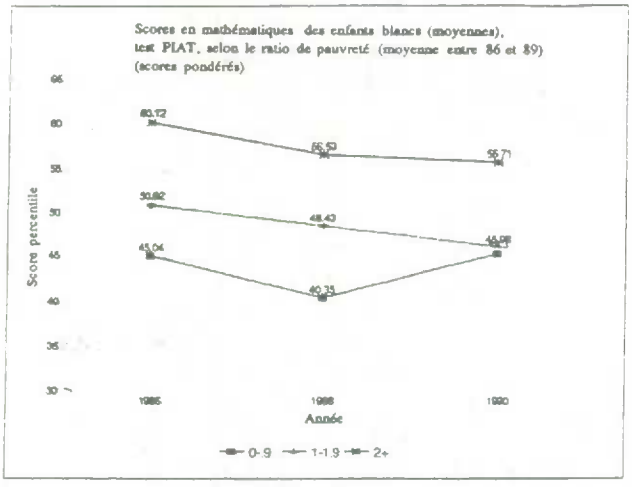
[7] D'autres analyses seraient nécessaires pour pouvoir préciser les relations de cause à effet entre ces facteurs. Il est fort possible que les mères d'enfants qui présentent de sérieux problèmes de comportement aient moins tendance à travailler (ou soient davantage portées à réduire leur temps de travail), comparativement aux autres mères.

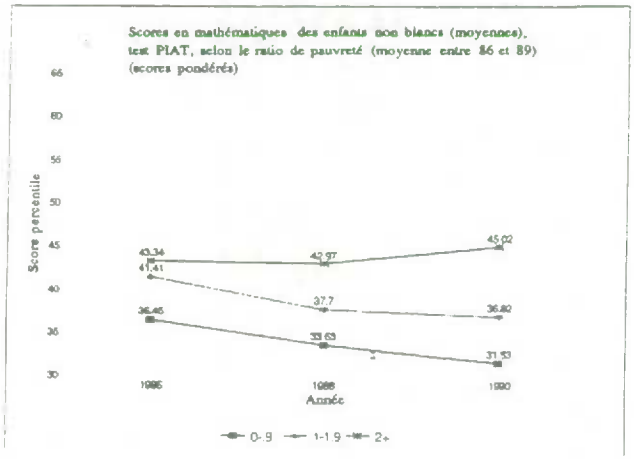
Tableau 3: Liens entre l'emploi maternel et les changements de scores (tests de connaissances et problèmes de comportement), selon la race et l'âge (moindres carrés ordinaires), échantillon groupé.

	TRAVAIL DE LA MÈRE > 20 SEMAINES PAR ANNÉE						TRAVAIL DE LA MÈRE 20-39 SEMAINES PAR ANNÉE					
	PIAT math.		PIAT lecture		Problèmes de comportement		PIAT math.		PIAT lecture		Problèmes de comportement	
Noirs												
7-8 ans	6.2	(4.1)	7.9[b]	(3.8)	-8.8[b]	(3.7)	11.5[a]	(4.5)	8.5[b]	(4.1)	-2.7	(4.1)
9-10 ans	-0.3	(4.3)	7.2[c]	(3.7)	-11.6[b]	(5.4)	5.9	(5.2)	12.1[a]	(4.5)	-6.4	(6.6)
11 ans et plus	-5.5	(5.0)	2.4	(4.2)	-12.0[b]	(5.0)	-1.8	(5.1)	4.5	(4.3)	-6.4	(5.1)
Non-Noirs												
7-8 ans	0.5	(2.3)	-2.2	(2.4)	1.6	(2.2)	-3.9	(2.6)	3.1	(2.7)	2.4	(2.5)
9-10 ans	-6.9[b]	(3.1)	-2.1	(3.0)	-5.6[c]	(3.0)	-5.6	(3.6)	4.8	(3.5)	-4.6	(3.5)
11 ans et plus	0.8	(3.4)	1.2	(2.3)	-4.2	(2.9)	-8.7	(4.1)[b]	3.5	(2.8)	6.5[c]	(3.5)

Nota: Les effets des variables explicatives énoncées dans le tableau 1 ont été éliminés pour les variables relatives à la pauvreté et à l'emploi. Les variables relatives à la pauvreté et à l'emploi représentent des statistiques moyennes de la période d'évaluation. Nous avons exclu les catégories de référence suivantes: ratio de pauvreté ≥ 2.0 et emploi de la mère 40-52 semaines par année. Le ratio de pauvreté désigne le rapport du revenu total de la famille au seuil officiel de la pauvreté fixé, lequel est fondé sur la taille de la famille et l'âge du chef du ménage). Les données sont tirées d'équations distinctes pour les Noirs et pour les Non-Noirs.
a = coefficient significatif au niveau p ≤ .01; b = ≤ .05; c = ≤ .10.

Figure 2.





mathématiques des enfants noirs mentionnée ci-dessus se traduit par un coefficient positif très important chez les plus jeunes. Cette constatation est conforme à l'hypothèse selon laquelle le maternage a davantage d'effets chez les jeunes enfants. Si nous passons aux enfants blancs, nous constatons que les effets faiblement négatifs observés dans le cas d'un niveau d'emploi moyen dans l'échantillon global se remarquent surtout chez les enfants plus âgés.

3.3 Résultats sur une période de quatre années

Il importe de souligner les contraintes associées à la durée très courte de la période de référence de l'analyse effectuée jusqu'ici. Le bien-être d'un enfant pourrait être étroitement lié au nombre d'années pendant lesquelles la famille a vécu dans une situation donnée vis-à-vis de la pauvreté ou encore pendant lesquelles la mère a connu tel ou tel régime de travail. La figure 2 illustre cette probabilité en comparant les scores obtenus en mathématiques aux trois tests de la période 1986-1990 par des enfants vivant dans diverses situations vis-à-vis de la pauvreté. Si l'on suit les enfants noirs de 1986 à 1988, on observe un certain élargissement des scores percentiles en mathématiques, que les enfants vivent ou non sous le seuil ou près du seuil de la pauvreté. Toutefois, l'écart entre les scores selon la situation vis-à-vis de la pauvreté s'accentue de façon considérable de 1988 à 1990. Par conséquent, il se pourrait que la durée du temps passé en situation de pauvreté stable ou encore plus défavorable influe négativement et de façon progressive sur les résultats,

sans que des facteurs extérieurs comme un enseignement d'appoint y changent quelque chose. Au contraire, chez les enfants blancs (dont les scores en mathématiques sont au départ nettement plus élevés) le portrait est très différent. Les résultats des enfants défavorisés par rapport au seuil de pauvreté diminuent au début (de 1986 à 1988) et connaissent ensuite un gain substantiel, peut-être en raison d'avantages comme de meilleures écoles ou un meilleur milieu de vie.

Pour mieux comprendre les tendances associées à l'évolution à long terme des enfants, nous nous sommes servis des résultats d'un échantillon plus restreint d'enfants évalués en 1986, 1988 et 1990 pour tenter de cerner les variables pouvant réagir à la durée du temps qu'un enfant passe dans une situation donnée. Comme on peut le voir dans le tableau 4, l'analyse à plusieurs variables portant sur la période de référence plus longue fait ressortir des tendances que ne montraient pas les équations relatives aux résultats sur deux ans. Chez les enfants noirs vivant sous le seuil ou près du seuil de la pauvreté, on observe une baisse importante des résultats en lecture et en mathématiques comparativement aux enfants de familles plus favorisées économiquement. Cette dernière constatation est conforme à la tendance décrite relativement à la figure 2. S'il est vrai que les enfants noirs dont la mère avait un niveau d'emploi faible ou moyen obtiennent un score en mathématiques plus élevé, l'emploi maternel ne semble exercer que peu d'effet sur les scores en lecture. L'amélioration du comportement des enfants noirs dont la mère travaillait un petit nombre de semaines durant l'année, déjà visible dans les équations pour deux années, s'accentue encore plus lorsque l'analyse porte sur une période de quatre années. Autre phénomène frappant que ne faisait pas ressortir les résultats sur deux années: l'effet négatif de la pauvreté sur les problèmes de comportement chez les enfants noirs. Ce résultat peut nous inciter à conclure que la durée du temps passé en situation de pauvreté finit par avoir des conséquences pour les enfants noirs, comme en témoigne l'augmentation importante des problèmes de comportement associés à la pauvreté chez ces enfants, dans l'équation relative à la période de quatre années. Les changements déjà observés chez les enfants blancs dans les équations des résultats sur deux années demeurent relativement stables à plus long terme.

Tableau 4: Liens entre l'emploi et la pauvreté et les changements de scores (percentiles), tests de connaissances et problèmes de comportement, selon la race, enfants évalués en 1986, 1988 et 1990 (moindres carrés ordinaires).

	PIAT MATHÉMATIQUES		PIAT LECTURE		PROBLÈMES DE COMPORTEMENT	
Noirs						
Pauvreté						
Sous le seuil de pauvreté	-11.5[a]	(4.6)	-11.7[a]	(4.5)	12.3[a]	(4.7)
Près de seuil de pauvreté	-6.7[c]	(3.9)	-7.2[c]	(3.8)	.2	(4.0)
Emploi maternel						
Mère travaillant moins de 20 semaines par année	7.9[b]	(3.8)	.88	(3.7)	-11.7[a]	(3.8)
Mère travaillant de 20 à 39 semaines par année	6.8[c]	(3.9)	-3.2	(3.8)	-1.5	(4.0)
Non-Noirs						
Pauvreté						
Sous le seuil de pauvreté	8.0[c]	(4.4)	.2	(4.2)	-3.8	(4.0)
Près du seuil de pauvreté	.4	(2.8)	-4.5[c]	(2.7)	-2.3	(2.9)
Emploi maternel						
Mère travaillant moins de 20 semaines par année	-4.9[c]	(3.1)	-4.4[c]	(3.0)	4.5[c]	(2.9)
Mère travaillant de 20 à 39 semaines par année	-5.5[c]	(3.1)	-.9	(2.9)	3.3	(2.8)

Nota: Les effets des variables explicatives énoncées dans le tableau 1 ont été éliminés pour les variables relatives à la pauvreté et à l'emploi. Les variables relatives à la pauvreté et à l'emploi représentent des statistiques moyennes de la période d'évaluation. Nous avons exclu les catégories de référence suivantes: ratio de pauvreté $\geq$ 2.0 et emploi de la mère 40-52 semaines par année. Le ratio de pauvreté désigne le rapport du revenu total de la famille au seuil officiel de la pauvreté fixé, lequel est fondé sur la taille de la famille et l'âge du chef du ménage). Les données sont tirées d'équations distinctes pour les Noirs et pour les Non-Noirs.
a = coefficient significatif au niveau $p \leq .01$; b = $\leq .05$; c = $\leq .10$.

3.4 Milieu de vie

L'ensemble de données de l'enquête NLSY comprend plusieurs éléments d'information et résultats de tests psychométriques sur lesquels nous pouvons nous fonder pour interpréter certains des résultats relatifs à la

pauvreté et à l'emploi maternel de la présente étude. L'échelle HOME (milieu de vie) de l'enquête NLSY est une version abrégée et modifiée d'une échelle d'utilisation très répandue dont l'objet est de mesurer la nature des interactions mère-enfant et la qualité du milieu de vie (Baker et Mott 1989). L'échelle HOME permet d'évaluer à la fois la stimulation du développement cognitif (par exemple, stimulation du langage, diversité des expériences, encouragement et renforcement des résultats de l'enfant) et le soutien émotif (par exemple, réceptivité, chaleur, incitation à la maturité). Comme le montre le tableau 5, les scores des enfants à ces égards sont nettement influencés par la situation vis-à-vis de la pauvreté de la famille. Il y a une relation étroite entre des scores élevés selon l'échelle globale HOME ainsi que selon les sous-échelles de la stimulation du développement cognitif et du soutien émotif et le revenu de la famille, chez les enfants blancs comme chez les enfants noirs. Bien que l'association soit moins grande, il existe aussi un lien positif entre des scores élevés selon l'échelle HOME et un niveau d'emploi maternel élevé, en particulier chez les enfants noirs. Apparemment, d'autres facteurs propres au milieu de vie de ces enfants noirs, lesquels sont sans doute corrélés positivement à une meilleure acquisition des connaissances, sont aussi associés à un niveau élevé d'emploi maternel[8].

Tableau 5: Scores percentiles moyens HOME selon la situation vis-à-vis de la pauvreté et l'importance de l'emploi maternel, enfants évalués en 1986, 1988 et 1990.

	Score HOME total	Taille de l'échantillon	Sous-score de l'échantillon	Taille de l'échantillon	Sous-score soutien émotif	Taille de l'échantillon
Noirs	36.9	404	43.1	393	37.0	350
Situation vis-à-vis de la pauvreté						
Ratio de pauvreté 0-.9	29.2	176	33.4	169	34.0	146
Ratio de pauvreté 1-1.9	36.9	109	41.4	107	40.5	95
Ratio de pauvreté 2 et plus	49.8	85	59.1	185	39.5	77
Emploi maternel						
< 20 semaines	33.1	161	38.7	154	33.3	139
20-39 semaines	37.6	79	40.8	79	38.3	70
40-52 semaines	40.3	164	48.4	160	40.0	141
Non-Noirs	57.4	492	55.7	484	57.2	463
Situation vis-à-vis de la pauvreté						
Ratio de pauvreté 0-.9	34.6	79	35.7	79	39.1	75
Ratio de pauvreté 1-1.9	50.4	147	49.1	144	53.2	137
Ratio de pauvreté 2 et plus	67.0	232	63.8	228	64.7	220
Emploi maternel						
< 20 semaines	52.5	177	51.0	174	54.9	172
20-39 semaines	56.6	115	52.0	112	59.1	106
40-52 semaines	61.8	200	61.3	198	58.1	185

Comme première mesure d'évaluation de l'importance en soi des variations observées relativement au milieu de vie des enfants, nous avons introduit dans les équations le score global obtenu à l'échelle HOME, à titre de variable explicative supplémentaire (les résultats de cette analyse ne figurent pas ici). Dans le cas des enfants noirs, la variable HOME ne modifie que très peu l'effet de la situation vis-à-vis de la pauvreté sur l'évolution des scores en lecture et ne renforce que légèrement son effet sur les scores en mathématiques. Pour les enfants blancs, les coefficients de pauvreté et d'emploi maternel ne changent pratiquement pas à la suite de la prise en compte de la variable HOME dans le modèle. Dans l'ensemble, la variable a peu d'effet sur la valeur des variables d'emploi maternel. Elle peut vraiment aider à prédire les niveaux absolus des résultats, mais *non pas* la façon dont ces résultats peuvent évoluer.

[8] Cette échelle comprend des éléments d'une validité sans doute plus élevée que d'autres pour ce qui est de prédire les compétences en mathématiques et expression orale. Nous prévoyons examiner la mesure dans laquelle ces éléments, qui contiennent beaucoup de renseignements sur les développement cognitif, pourraient aider à expliquer les relations entre l'emploi maternel et, en particulier, l'acquisition des compétences en mathématiques des enfants noirs.

4. CONCLUSIONS

Les principales constatations de cette étude sont que les liens entre la situation familiale et les résultats que les enfants obtiennent peuvent varier en fonction de facteurs comme la race et, dans une certaine mesure, la durée de la période d'observation. Lorsque nous passons d'une période de référence de deux ans à quatre années de résultats, nous constatons que certaines des tendances observées à court terme s'accentuent, en particulier dans le cas des enfants noirs. Sur une période de quatre ans, un grand nombre des enfants des âges à l'étude passent par plusieurs stades de développement, processus qui peut dépendre largement de la façon dont ces enfants réagissent à des stimuli de l'intérieur et de l'extérieur de la famille. Si la période d'observation est plus longue, l'enfant a plus de chances de connaître divers milieux de vie. Dans une certaine mesure, nous avons tenté de tenir compte dans nos analyses de changements qui concernent davantage «la vie à la maison», mais plus la période d'observation s'allonge, plus les facteurs «inobservables» du milieu extérieur peuvent prendre de l'importance. Les caractéristiques du voisinage, de l'école, du groupe d'amis peuvent influer sur les relations entre les résultats observés chez les enfants et les variables à l'étude ici. De telles influences peuvent être particulièrement importantes dans le cas des enfants noirs dont la situation vis-à-vis de la pauvreté de la famille peut aussi se vivre au niveau communautaire, davantage que chez les enfants blancs dont la famille est également défavorisée économiquement.

Lorsqu'ils sont examinés sur une plus longue période, les principaux facteurs à l'étude, soit la situation vis-à-vis de la pauvreté et, dans une moins grande mesure, l'emploi maternel, semblent afficher une plus grande variabilité à long terme. On pourrait pousser encore plus l'étude de tels changements en examinant la répartition dans le temps, c'est-à-dire à l'intérieur de la période de référence, des situations liées à la pauvreté et à l'emploi maternel. On pourrait aussi essayer de déterminer dans quelle mesure le degré de maturité et la race des enfants peuvent influer à la baisse sur la relation entre la répartition dans le temps des facteurs à l'étude et les résultats observés chez ces enfants. Dans cette analyse, nous avons accordé la même importance à la variable de la situation socio-économique de la famille et à la variable de l'emploi maternel, mais il existe peut-être des différences sensibles entre les deux qui pourraient modifier les résultats que nous avons présentés.

Malgré les nombreuses questions sans réponse, cette analyse a aider à clarifier certains points importants: les enfants noirs réagissent différemment à certaines situations ou caractéristiques de la famille. Il ne faut pas sous-estimer l'importance de l'emploi maternel sur le développement d'un jeune enfant, mais surtout, il importe de se rendre compte que les effets d'un tel facteur ne sont pas uniformes pour tous les types de résultats mesurés. On ne comprend pas encore très bien comment les avantages associés à un faible niveau ou à un niveau moyen d'emploi maternel peuvent être reliés à la qualité de l'emploi de la mère. Le plus surprenant dans le cas des enfants noirs est de constater la faible relation directe entre la pauvreté *en soi* et les niveaux ou les changements de niveaux des résultats à court terme des enfants sur le plan des connaissances ou du comportement. On assiste cependant à un revirement important de cette tendance lorsqu'on passe d'une période d'observation de deux à quatre années. S'il est vrai que les effets du régime de travail de la mère sur les résultats des enfants restent stables dans l'ensemble dans le temps, les effets d'autres variables de la famille sur les résultats des enfants pourraient être sous-estimés si l'on s'en tient exclusivement à des études à court terme.

RÉFÉRENCES

Allison, P.D. (1991). Change scores as dependent variables in regression analysis. *Sociological Methodology*, 20, 93-114.

Baker, P.C., et Mott, F.L. (1989). NLSY child handbook 1989: A guide and resource document for the national longitudinal study of youth 1986 child data. Columbus: Center for Human Resource Research, Ohio State University.

Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery. (1982). Washington, DC: Office of the Assistant Secretary of Defense.

Parcel, T.L., et Menaghan, E.G. (1990). Maternal working conditions and child verbal facility: Studying the intergenerational transmission of inequality from mothers to young children. *Social Psychology Quarterly*, 53, 132-147.

Piotrkowski, C.S., Rapoport, R.N., et Rapoport, R. (1987). Families and work. Dans M.B. Sussman and S.K. Steinmetz, Éds. *Handbook of Marriage and the Family*, 251-283. New York: Plenum.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

UTILISATION DE L'ENQUETE SUR L'ACTIVITÉ POUR L'ESTIMATION DES ÉCARTS DE SALAIRES ENTRE GRANDES ET PETITES ENTREPRISES AU CANADA

R. Morissette[1]

RÉSUMÉ

Même en tenant compte des caractéristiques observables des travailleurs et des capacités, constantes dans le temps et non observées, des hommes qui changent d'emploi, il reste un écart de salaires considérable entre les grandes et les petites entreprises. Ce qui est à la base de cet écart n'est pas évident. L'écart de salaires observé dans une équation de salaires à différence première peut être sujet à des problèmes importants d'autosélection; cet écart pourrait simplement refléter des différences dans les capacités de production propres à un secteur (c.-à-d. propres à la taille de l'entreprise). Par contre, selon les modèles de salaires basés sur le rendement, il se pourrait que les grandes entreprises versent des salaires plus élevés afin d'accroître l'effort des travailleurs ou pour réduire le roulement du personnel.

MOTS CLÉS: Salaires; taille de l'entreprise; revenu; marché du travail; emploi.

1. INTRODUCTION

Selon la théorie du capital humain et la théorie des écarts compensateurs, les salaires sont déterminés uniquement par le capital humain des travailleurs et par des aspects non pécuniaires des emplois. Une fois que l'on tient compte de ces facteurs, les écarts de salaires devraient disparaître. Un travail récent de Krueger et Summers (1988) sur les écarts de salaires entre les branches d'activité montre que ce n'est pas le cas; on trouve des écarts de salaires entre les branches d'activité même après que l'on a tenu compte de ces facteurs. Tout comme la structure de la branche d'activité, il semble que la taille de l'employeur ait une incidence sur les salaires. Des études récentes réalisées aux États-Unis (Brown et Medoff 1989; Idson et Feaster 1990) laissent supposer que les grands employeurs ont tendance à verser des salaires plus élevés. L'objet principal de la présente communication est de déterminer si l'on retrouve un tel rapport entre les salaires et la taille des entreprises au Canada.

Des données provenant de l'enquête sur l'activité (EA) de 1986 nous ont permis de constater que les grandes entreprises versent généralement des salaires plus élevés à des travailleurs équivalents, selon nos observations. Cela laisse supposer qu'une partie de l'écart de salaires entre les travailleurs canadiens dépend de facteurs qui ne sont pas liés aux attributs des travailleurs.

La présente communication est structurée de la façon suivante: dans la section 2 nous présentons le modèle théorique. Dans la section 3 on montre que, même quand on tient compte des caractéristiques observables des travailleurs ainsi que des effets propres à la profession et à la branche d'activité, les grandes entreprises paient encore environ 20% de plus que les petites entreprises. Une partie de cet écart de salaires peut être attribuable à des différences dans les capacités non observées des travailleurs. Nous utilisons des données longitudinales pour tenir compte des écarts dans les capacités constantes dans le temps et non observées et nous obtenons quand même un écart de salaires considérable pour les hommes qui changent d'emploi. Cela souligne le besoin d'autres explications du processus de détermination des salaires. Ces explications sont étudiées brièvement dans la section 4. Les conclusions sont présentées après la section 4.

---

[1] R. Morissette, Groupe d'analyse des entreprises et du marché du travail, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

2. LE MODÈLE

On suppose que les salaires dépendent du capital humain des travailleurs, d'aspects non pécuniaires des emplois, de la taille de l'entreprise et d'autres facteurs:

$$W_{it} = W(HC_{it}, CD_{it}, SIZE_{it}, OTHER_{it}, u_{it}) \quad (1)$$

où W_{it} est le salaire du travailleur i au moment t, HC_{it} est un vecteur des caractéristiques observables du travailleur i qui accroissent la productivité de ce dernier, CD_{it} est un vecteur de caractéristiques des emplois qui ont une incidence sur le niveau d'utilité du travailleur i, $SIZE_{it}$ est un vecteur de variables qui mesurent la taille de l'entreprise, $OTHER_{it}$ est un vecteur d'autres facteurs qui pourraient avoir une incidence sur les salaires (p. ex., le statut syndical, l'état matrimonial, la race) et u_{it} est une perturbation aléatoire. Selon la théorie du capital humain et la théorie des écarts compensateurs, $SIZE_{it}$ n'a aucune incidence sur W_{it}. Par conséquent, l'équation (1) est une spécification dans laquelle un modèle basé sur la théorie du capital humain et sur la théorie des écarts compensateurs est emboîté dans un modèle où la taille de l'entreprise a de l'importance. Il se peut que la taille de l'entreprise ait de l'importance à cause de l'hétérogénéité des caractéristiques des employeurs comme la puissance commerciale, les coûts de formation et les coûts de contrôle.

3. RÉSULTATS EMPIRIQUES

3.1 Les données

Dans la présente communication, une entreprise est définie comme l'ensemble de tous les établissements appartenant à un employeur donné au Canada. Les petites entreprises sont définies comme celles qui ont moins de 20 employés, les entreprises de taille moyenne sont celles qui emploient entre 20 et 499 travailleurs et les grandes entreprises ont 500 travailleurs ou plus. L'échantillon utilisé dans la présente communication comprend tous les emplois à plein temps occupés en 1986 par les travailleurs rémunérés du secteur commercial[2]. Les données sont tirées de l'enquête sur l'activité (EA) de 1986[3].

Au niveau agrégé, les grandes entreprises versent des salaires plus élevés d'environ 50% que ce n'est le cas pour les petites entreprises (Tableau 1). Cela soulève évidemment la question suivante: qu'est-ce qui est à la base de cet écart de salaires?

---

[2] Dans la présente communication, le secteur commercial comprend toutes les branches d'activité sauf les suivantes: 1) industries agricoles, 2) industries de la pêche et du piégeage, 3) industries des services d'enseignement et services connexes, 4) industries des services de soins de santé et des services sociaux, 5) organisations religieuses, 6) administration fédérale, 7) administration provinciale, 8) administration locale et 9) autres services des administrations publiques. Puisque nous nous limitons aux travailleurs rémunérés, l'échantillon résultant ne comprend pas les travailleurs familiaux non rémunérés et les travailleurs autonomes. En 1986, les heures travaillées dans des emplois à plein temps dans le secteur commercial étaient réparties de la façon suivante: 25 % dans les petites entreprises, 35 % dans les entreprises de taille moyenne et 40 % dans les grandes entreprises.

[3] Morissette (1992) compare les données sur l'emploi obtenues dans le cadre de l'EA à celles provenant de l'enquête sur l'emploi, la rémunération et les heures de travail (EERH). Il montre que la répartition de l'emploi selon la taille de l'établissement découlant de l'EA sous-estime considérablement la taille de l'établissement telle qu'obtenue à partir des données de l'EERH et que, par conséquent, cette répartition de l'emploi ne peut être utilisée pour évaluer si les grands établissements versent des salaires plus élevés que les petits établissements. Toutefois, la répartition de l'emploi selon la taille de l'entreprise, telle qu'établie à partir des données provenant des deux enquêtes, est semblable. Cela laisse supposer que l'on peut utiliser l'EA pour étudier si les grandes entreprises versent des salaires plus élevés que les petites.

Tableau 1: Salaires horaires moyens selon la taille de l'entreprise, emplois à plein temps.

Taille de l'entreprise (nombre d'employés)				
(1) 1-19	(2) 20-99	(3) 100-499	(4) 500+	(5) [(4) - (1) / (1)]
8.91	10.66	11.98	13.50	0.52

Source: Enquête sur l'activité (1986).

3.2 Prise en considération des caractéristiques observables des travailleurs

Selon la théorie du capital humain, les personnes qui sont plus instruites, qui possèdent davantage d'habiletés générales et propres à l'entreprise, reçoivent des salaires plus élevés parce qu'elles sont plus productives. Selon la théorie des écarts compensateurs, les entreprises doivent offrir des salaires plus élevés pour attirer des travailleurs d'une qualité donnée dans des emplois qui comportent de mauvaises conditions de travail. Pour vérifier la justesse de ces arguments, nous spécifions l'équation (1) de la façon suivante:

$$ln\ W_{it} = B_0 + B_1 * HC_{it} + B_2 * CD_{it} + B_3 * SIZE_{it} + B_4 * OTHER_{it} + u_{it} \qquad (2)$$

où: HC_{it} comprend: niveau de scolarité, âge, carré de l'âge, durée d'occupation d'un emploi, carré de la durée d'occupation d'un emploi

CD_{it} comprend: branches d'activité selon les codes à 2 chiffres, professions selon les codes à 2 chiffres

$SIZE_{it}$ comprend: taille de l'entreprise

$OTHER_{it}$ comprend: état matrimonial, statut relativement aux minorités visibles, statut syndical, noyau urbain, région.

Tableau 2: Écarts de salaires entre les petites et les grandes entreprises, emplois à plein temps[1].

	Hommes	Femmes
Taille de l'entreprise		
FIRM2[3]	0.0815	0.1007
	(0.0093)[2]	(0.0121)
FIRM3	0.1450	0.1650
	(0.0105)	(0.0137)
FIRM4	0.1853	0.2166
	(0.0095)	(0.0119)
R^2 corr.	0.4256	0.4388
Taille de l'entreprise	15 506	8 382

[1] La variable dépendante est le logarithme du taux de salaire horaire. Les régressions sont calculées à l'aide des moindres carrés pondérés; chaque emploi est pondéré par le nombre d'heures de travail.

[2] Les écarts-types des coefficients de taille sont entre parenthèses.

[3] FIRM2(FIRM4) correspond aux entreprises comptant de 20 à 99 employés (500 employés ou plus). Les entreprises comptant de 1 à 19 employés constituent le groupe de référence. L'écart de salaires en pourcentage entre les petites et les grandes entreprises est égal à l'antilogarithme du coefficient de régression moins 1 et exprimé en pourcentage.

• pas significatif au niveau 5%.

Source: Enquête sur l'activité (1986).

Comme dans la fonction des gains du capital humain de Mincer (1974), l'équation (2) renferme des variables qui reflètent les divers niveaux de scolarité (c.-à-d. l'instruction) et d'expérience sur le marché du travail (mesurée indirectement par l'âge) des travailleurs. La durée d'occupation d'un emploi est ajoutée pour tenir compte des différences dans l'expérience des travailleurs qui sont pertinentes pour l'emploi actuel[4]. Pour tenir compte des différences dans les aspects non pécuniaires des emplois, idéalement on voudrait disposer d'un indice de la qualité des conditions de travail qui pourrait être inclus dans le membre de droite de l'équation (2). L'EA ne nous donne pas de tels renseignements. À l'instar de Brown et Medoff (1989), nous ajoutons des éléments qui nous permettent de tenir compte de la branche d'activité et de la profession; cela peut nous aider à saisir une partie de la variation dans les conditions de travail qui se produit entre les branches d'activité et les professions. Parce qu'ils peuvent avoir un comportement différent pour ce qui est de l'abandon des emplois ou parce qu'ils peuvent faire face à de la discrimination, il se peut que les travailleurs non mariés et ceux qui font partie des minorités visibles reçoivent des salaires différents; deux termes de décalage de l'ordonnée à l'origine sont inclus pour tenir compte de ces possibilités. Le statut syndical est inclus afin de tenir compte de l'incidence de la syndicalisation sur les salaires. Quatre variables dichotomiques régionales et une variable dichotomique pour les régions métropolitaines de recensement sont incluses afin de tenir compte de la possibilité d'avoir des marchés du travail locaux distincts à cause de la mobilité géographique imparfaite des travailleurs.

Le tableau 2 présente l'écart de salaires entre les petites et les grandes entreprises calculé à l'aide de l'équation (2). Les résultats sont montrés séparément pour les hommes et pour les femmes. Dans les deux cas, les grandes entreprises paient des salaires plus élevés que les petites. L'écart de salaires entre les grandes et les petites entreprises varie de 20% à 24%[5].

3.3 Prise en considération des capacités, constantes dans le temps et non observées, des travailleurs

Comme cela se fait couramment dans les études portant sur l'effet de la syndicalisation (Freeman 1984), de la branche d'activité (Krueger et Summers 1988) ou de la taille de l'entreprise (Evans et Leighton 1989: Brown et Medoff 1989) sur les salaires, on peut soutenir qu'une partie de la variation dans les salaires est due au fait que les travailleurs ont des capacités non observées différentes. Plus précisément, si les travailleurs des grandes entreprises possèdent plus de ces capacités non observées, il est alors possible que l'écart de salaires relevé jusqu'ici ne reflète qu'une "différence non observable dans la qualité des travailleurs". Le calcul de la différence première de l'équation (2) nous permet de tenir compte de la partie de ces capacités non observées qui est constante dans le temps. Pour voir comment cela se produit, considérons l'équation de salaires suivante:

$$ln\ W_{it} = B_1 * X_{it} + B_2 * a_i + u_{it} \qquad (3)$$

où $ln\ W_{it}$, le logarithme du salaire du travailleur i au moment t, dépend d'un vecteur X_{it} de variables observables, de capacités constantes dans le temps et non observées a_i et d'un terme aléatoire u_{it}. Le calcul de la différence première de l'équation ci-dessus nous donne l'équation suivante:

$$ln\ W_{it} - ln\ W_{it-1} = B_1 * (X_{it} - X_{it-1}) + (u_{it} - u_{it-1}) \qquad (4)$$

dans laquelle les capacités constantes dans le temps et non observées ne figurent plus et l'équation en tient donc implicitement compte. L'équation (4) a été estimée aux États-Unis pour la période allant de 1973 à 1977 (Brown et Medoff 1989) à l'aide de l'enquête "Quality of Employment Survey" et pour la période allant de 1976 à 1981 [Evans et Leighton (1979) à l'aide de l'enquête "National Longitudinal Survey of Young Men"]. Alors que Brown et Medoff (1989) trouvent que l'effet de la taille demeure considérable même après qu'on a tenu compte des différences dans les capacités non observées, Evans et Leighton (1979) concluent qu'environ 60 pour cent de

---

[4] Le carré de la durée d'occupation d'un emploi et le carré de l'âge sont inclus afin de permettre qu'il y ait non-linéarité dans le rapport âge/salaire ou durée d'occupation d'un emploi/salaire.

[5] L'écart de salaires en pourcentage est l'antilogarithme du coefficient de régression moins 1, exprimé en pourcentage. Ainsi, pour les emplois à plein temps occupés par des hommes, l'écart de salaires de 20 % entre les grandes et les petites entreprises découle du calcul suivant: exp(0.1853) - 1.0.

l'effet salaire-taille est dû à l'hétérogénéité non observée quand on considère toutes les entreprises et environ 100 pour cent quand on considère les entreprises de 25 employés ou plus (p. 299).

Comme on l'a mentionné plus haut, nous nous basons sur la version 1986 du fichier de l'EA. Dans ce fichier, on trouve des renseignements sur jusqu'à cinq emplois occupés par une personne donnée en 1986. Nous nous concentrons sur le premier et le deuxième emploi occupé cette année-là par toutes les personnes qui ont changé d'emploi[6]. Cela nous donne un total de 1 539 et de 897 observations sur les écarts de salaires pour les travailleurs et les travailleuses, respectivement.

Nous estimons la version de différence première de l'équation (2)[7]. La variable dépendante est la différence entre le (logarithme naturel du) taux de salaire horaire dans le deuxième emploi occupé en 1986 et la valeur correspondante pour le premier emploi occupé en 1986. Nous ajoutons aussi une variable dichotomique pour faire la distinction entre les personnes qui changent d'emploi mais qui restent dans la même profession (selon le code à 2 chiffres) et les personnes qui changent de profession quand elles passent de leur premier à leur deuxième emploi. Parce qu'il est vraisemblable que les travailleurs qui appartiennent au premier groupe utilisent dans leur deuxième emploi une partie substantielle des connaissances qu'ils ont acquises dans l'emploi précédent, on s'attend à ce qu'ils touchent des augmentations de salaire net supérieures à celles des autres travailleurs. La troisième colonne du tableau 3 présente les écarts de salaires découlant de cette équation de salaires à différence première. Les deux premières colonnes montrent les estimations d'écarts de salaires obtenues à l'aide de l'équation (2) appliquée: 1) à tous les premiers emplois occupés par les personnes qui changent d'emploi et 2) à tous les deuxième emplois occupés par ces mêmes personnes.

Les deux premières colonnes du tableau 3 montrent que l'écart de salaires entre les grandes et les petites entreprises (FIRM4) varie entre 8% et 15% pour les hommes qui changent d'emploi. Si cet écart était simplement dû à des différences dans les capacités non observées des travailleurs, il disparaîtrait quand on utilise l'équation de salaires à différence première. Manifestement, ce n'est pas le cas; l'écart de salaires obtenu quand on emploie l'équation de salaires à différence première est égal à 9% (colonne 3). On pourrait soutenir qu'on devrait s'attendre à ce que les travailleurs qui quittent leur emploi reçoivent des augmentations de salaire net plus élevées que ceux qui sont mis à pied. Si c'est le cas, la raison pour laquelle on change d'emploi devrait être incluse comme variable explicative. Le fait d'ajouter deux variables dichotomiques pour les départs et les mises à pied à l'équation de salaires à différence première (colonne 4) ne modifie pas considérablement les coefficients de taille; l'écart de salaires entre les grandes et les petites entreprises reste à 9%[8]. De plus, bien que l'effet salaire-taille de l'entreprise disparaisse (quand on passe de l'équation (2) à sa version de différence première) pour les entreprises comptant entre 100 et 499 employés (FIRM3), il demeure important et assez constant pour les entreprises comptant entre 20 et 99 employés (FIRM2). Ainsi, il semble juste de conclure que les capacités non observées ne peuvent expliquer tout l'écart de salaires constaté pour les hommes qui changent d'emploi.

Les résultats pour les femmes qui changent d'emploi sont quelque peu intrigants. Alors que l'effet salaire-taille de l'entreprise demeure important pour les entreprises comptant entre 100 et 499 employés (FIRM3), il disparaît pour les entreprises comptant entre 20 et 99 employés (FIRM2). De plus, le coefficient de taille pour les grandes entreprises (FIRM4) n'est pas significatif pour le premier emploi occupé par des travailleuses en 1986. Nous n'avons pas d'explication simple à offrir pour ce comportement.

---

[6] Le nombre de personnes qui occupent plus de deux emplois pendant la même année est trop faible pour une analyse statistique probante.

[7] Les variables qui prennent des valeurs constantes durant une année (niveau de scolarité, âge, sexe, statut relatif aux minorités visibles) disparaissent quand nous passons de l'équation (2) à sa version de différence première.

[8] La variable dichotomique pour les départs est significative au niveau 5 % et laisse supposer que, comparativement au groupe de référence [c.-à-d. les travailleurs qui quittent leur emploi pour d'autres raisons (maladie, responsabilités personnelles ou familiales, température inclémente, conflits de travail, vacances non rémunérées, nature saisonnière de l'emploi, vente de l'entreprise ou de l'exploitation agricole, autres)], les travailleurs masculins qui quittent leur emploi reçoivent des augmentations de salaire net de 10 % supérieures. La variable dichotomique pour les mises à pied est significative au niveau 6 % et laisse supposer que les travailleurs masculins qui sont mis à pied reçoivent des augmentations de salaire net inférieures de 5 % à celles obtenues par les personnes qui quittent leur emploi pour d'autres raisons.

Tableau 3: Écart de salaires entre les petites et les grandes entreprises pour les personnes qui changent d'emploi, emplois à plein temps.

	Équation de salaires[1]: premier emploi occupé en 1986	Équation de salaires[1]: deuxième emploi occupé en 1986	Équation de salaires à différence première[2]	
			sans	avec variables dichotomiques pour les départs et les mises à pied
	(1)	(2)	(3)	(4)
Hommes				
FIRM2	0.0790	0.0546	0.0629	0.0604
	(0.0263)	(0.0261)	(0.0237)	(0.0234)
FIRM3	0.1315	0.1146	0.0337*	0.0237*
	(0.0345)	(0.0310)	(0.0313)	(0.0310)
FIRM4	0.0803	0.1423	0.0895	0.0866
	(0.0286)	(0.0295)	(0.0287)	(0.0284)
R^2 corr.	0.4771	0.5063	0.1799	0.1974
Taille de l'échantillon	1,539	1,539	1,539	1,539
Femmes				
FIRM2	0.1107	0.0326*	0.0113*	0.0093*
	(0.0423)	(0.0325)	(0.0372)	(0.0370)
FIRM3	0.1107	0.2478	0.1405	0.1420
	(0.0481)	(0.0361)	(0.0430)	(0.0428)
FIRM4	0.0365*	0.1235	0.0879	0.0928
	(0.0422)	(0.0297)	(0.0377)	(0.0376)
R^2 corr.	0.4325	0.5640	0.1085	0.1189
Taille de l'échantillon	897	897	897	897

[1] La variable dépendante est le logarithme du taux de salaire horaire. Les régressions sont calculées à l'aide des moindres carrés pondérés. Chaque emploi est pondéré par le nombre d'heures de travail.

[2] La variable dépendante est la différence première du logarithme du taux de salaire horaire. Les régressions sont calculées à l'aide des moindres carrés pondérés. Chaque observation est pondérée par son poids d'échantillonnage. Pour d'autres détails, voir le tableau 2.

Source: Enquête sur l'activité (1986).

En dépit de ce fait, les données présentées au tableau 3 laissent supposer, du moins pour les hommes qui changent d'emploi, qu'il reste un écart de salaires important entre les grandes et les petites entreprises, que nous tenions compte ou non des capacités non observées des travailleurs. Cela laisse entendre que des travailleurs qui sont équivalents, selon nos observations, peuvent recevoir des salaires différents selon la taille de l'entreprise pour laquelle ils travaillent. Alors pourquoi les grandes entreprises paieraient-elles des salaires plus élevés?

4. POURQUOI LES GRANDES ENTREPRISES PAIERAIENT-ELLES DES SALAIRES PLUS ÉLEVÉS?

L'économique offre de nombreuses raisons qui expliquent pourquoi les grandes entreprises verseraient des salaires plus élevés. Dans les sections précédentes, nous avons traité de certaines de ces raisons. D'après Brown et Medoff (1989), on peut soutenir que les grandes entreprises verseraient des salaires plus élevés parce que:

1) elles ont une main-d'oeuvre de qualité supérieure;
2) elles doivent compenser les travailleurs pour de mauvaises conditions de travail;
3) elles désirent éviter la syndicalisation;
4) elles ont moins de postulants par emploi et doivent augmenter les salaires pour attirer une qualité donnée de postulants (Weiss et Landau 1984).
5) elles ont une puissance commerciale supérieure (c.-à-d. que leurs courbes de demande sont plus inélastiques) et elles partagent une partie de leurs bénéfices (supérieurs à la moyenne) avec les travailleurs.

Alors que l'argument de la qualité du travail suppose que des entreprises de taille différente versent aux travailleurs possédant des caractéristiques identiques un salaire identique, les quatre autres hypothèses laissent entendre que des travailleurs identiques peuvent recevoir des salaires différents. On peut aussi utiliser des modèles de salaires basés sur le rendement afin d'expliquer pourquoi des employeurs verseraient des salaires différents à des travailleurs identiques (Yellen 1984). Tels qu'appliqués au rapport entre le salaire et la taille de l'entreprise, ces modèles pourraient être utilisés pour soutenir que les grandes entreprises verseraient des salaires plus élevés parce que:

6) elles ont plus de difficultés que les petits employeurs à découvrir les personnes qui manquent à leurs obligations et utilisent des salaires plus élevés comme mécanisme pour assurer la discipline parmi les travailleurs (Shapiro et Stiglitz 1984);
7) elles ont des coûts de formation plus élevés et utilisent des salaires plus élevés comme moyen de réduction du roulement du personnel (Salop 1979);
8) elles comptent plus sur le travail d'équipe que les petits employeurs et désirent élever les normes de travail de leurs travailleurs au-dessus du minimum requis en versant à ces derniers des salaires supérieurs au minimum requis (Akerlof 1982)[9].

5. CONCLUSIONS

Les données présentées dans cette communication laissent supposer que, du moins pour les hommes qui changent d'emploi, les différences dans les caractéristiques observables des travailleurs ou dans leurs capacités constantes dans le temps et non observées ne peuvent expliquer entièrement pourquoi les grandes entreprises versent des salaires plus élevés que les petites entreprises. Nous ne savons pas encore pourquoi cela se produit. L'écart de salaires observé dans l'équation de salaires à différence première peut être sujet à des problèmes de sélection qui pourraient être importants. D'après Heckman et Sedlacek (1985), on peut soutenir qu'il ne reflète que des différences dans les capacités propres à un secteur (c.-à-d. propres à la taille d'une entreprise). Les travailleurs qui vont volontairement des petites aux grandes entreprises auraient des capacités productives dont le prix unitaire serait **beaucoup** plus élevé dans les grandes entreprises que dans les petites alors que les travailleurs qui vont volontairement des grandes aux petites entreprises auraient des capacités dont le prix unitaire serait **légèrement** plus élevé dans les petites entreprises que dans les grandes. Comme Krueger et Summers (1988) l'ont fait dans le cas des écarts de salaires entre les branches d'activité, on pourrait tenir compte

[9] Une autre version des modèles de salaires basés sur le rendement (modèles de sélection défavorable: voir Weiss (1980)) laisse supposer que les entreprises ne peuvent déterminer les capacités des travailleurs (que l'on suppose être non observables) et doivent verser des salaires plus élevés pour attirer une meilleure réserve de candidats. Tels qu'appliqués au rapport salaire-taille de l'entreprise, ces modèles laisseraient entendre que les grandes entreprises versent des salaires plus élevés parce qu'elles désirent avoir des travailleurs très qualifiés. Puisqu'il reste un écart de salaires considérable, même quand on a tenu compte des capacités constantes dans le temps et non observables ainsi que des caractéristiques observables des hommes qui changent d'emploi, on ne peut utiliser ces modèles pour expliquer l'écart de salaires qui reste.

de cet argument en: a) vérifiant si les personnes qui passent des petites aux grandes entreprises reçoivent une augmentation de salaire semblable à la diminution de salaire que les personnes passant des grandes aux petites entreprises subissent (supposément), b) étudiant les changements dans le salaire des travailleurs déplacés, c.-à-d. les travailleurs qui changent d'emploi involontairement après avoir été mis à pied.

Il se peut aussi que l'écart de salaires qui reste soit dû à l'hétérogénéité des entreprises. Comme les modèles de salaires basés sur le rendement le laissent supposer si, pour les entreprises de taille différente, les coûts de formation de ces dernières, la façon dont elles comptent sur le travail d'équipe ou la facilité avec laquelle elles peuvent surveiller leurs travailleurs ne sont pas identiques, ces entreprises peuvent trouver profitable de verser des salaires différents à des travailleurs identiques. Parce que des explications différentes de l'écart de salaires peuvent mener à des implications différentes en matière de politique économique, la détermination de la source de l'effet salaire-taille de l'entreprise est une question d'importance pour les recherches ultérieures.

L'auteur désire remercier Ted Wannell de son aide et de ses commentaires précieux ainsi que Garnett Picot et John Baldwin de leurs commentaires utiles relativement à des versions antérieures de la présente communication.

BIBLIOGRAPHIE

Akerlof, G.A. (1982). Labor contracts as a partial gift exchange. *Quarterly Journal of Economics*, 97, 543-569.

Brown, C., et Medoff, J. (1989). The employer size-wage effect. *Journal of Political Economy*, 97, 1027-1059.

Evans D.S., et Leighton, L.S. (1989). Why do smaller firms pay less? *Journal of Human Resources*, 24, 299-318.

Freeman, R. (1984). Longitudinal analyses of the effects of trade union. *Journal of Labor Economics*, 2, 1-26.

Heckman, J., et Sedlacek, G. (1985). Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market. *Journal of Political Economy*, 93, 6, 1077-1125.

Idson, T.L., et Feaster, D.J. (1990). A selectivity model of employer-size wage differentials. *Journal of Labor Economics*, 8, 99-122.

Krueger, A.B., et Summers, L.H. (1988). Efficiency wages and the inter-industry wage structure. *Econometrica*, 56, 259-293.

Mincer, J. (1974). Schooling, experience, and earnings. (New York: Columbia University Press).

Morissette, R. (1992). Canadian jobs and firm size: do smaller firms pay less? *Canadian Journal of Economics*, à paraître.

Salop, S.C. (1979). A model of the natural rate of unemployment. *American Economic Review*, 69, 117-125.

Shapiro, C., et Stiglitz, J.E. (1984). Equilibrium unemployment as a worker discipline device. *American Economic Review*, 74, 433-444.

Weiss, A. (1980). Job queues and layoffs in labor markets with flexible wages. *Journal of Political Economy*, 88, 526-538.

Weiss, A., et Landau, H. (1984). Wages, hiring standards, and firm size. *Journal of Labor Economics*, 2, 477-499.

Yellen, J.L. (1984). Efficiency wage models of unemployment. *American Economics Association Papers and Proceedings*, 74, 200-205.

SESSION 9

Analyse de données II

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

CRÉATION D'UNE BASE DE DONNÉES COMPARATIVE INTERNATIONALE DE PANEL: LE PROJET COPA

G. Schaber, G. Schmaus et G.G. Wagner[1]

RÉSUMÉ

Il est difficile d'effectuer des recherches internationales à l'aide d'ensembles de données provenant de panels nationaux parce que chacun des ensembles de données nationaux est structuré de façon différente et utilise une présentation différente. Afin de surmonter ces problèmes, le CEPS/INSTEAD crée, au Luxembourg, en collaboration avec le DIW Berlin, une base de données comparative internationale qui renferme les données recueillies auprès de divers panels nationaux de ménages. Le but du projet COPA est d'élaborer des instruments afin d'analyser, de programmer et de simuler des politiques socio-économiques. Ce projet vise à faciliter la recherche comparative transnationale sur des questions de politique comme l'activité, la distribution du revenu, la pauvreté, les problèmes des personnes âgées et ainsi de suite.

MOTS CLÉS: Variables harmonisées; relationnelle structure.

1. INTRODUCTION

Pour saisir et analyser des phénomènes dans divers domaines tels que le marché du travail, la distribution du revenu, la pauvreté ou la protection sociale et aussi pour établir des plans en vue d'élaborer la politique publique, les spécialistes en sciences sociales et les économistes des années 80 et 90 font une grande utilisation de micro-données.

Les micro-données utilisées dans la majorité des études réalisées par ces personnes découlent, toutefois, d'enquêtes transversales. Dans le cadre de la LIS (voir Smeeding, T.M. et Schmaus, G. 1988; Smeeding, T.M. et Schmaus, G. 1990), on fait une tentative considérable pour grouper des fichiers transversaux et des renseignements sur le revenu. Aux fins mentionnées plus haut, les données transversales sont évidemment supérieures aux agrégats; néanmoins, pour traiter de PROCESSUS économiques ou sociaux ainsi que de la DYNAMIQUE qui les sous-tend, ces renseignements donnent des résultats plutôt médiocres.

C'est pourquoi certaines équipes de recherche, tout d'abord aux É.-U., puis en Europe, ont commencé à constituer des échantillons par panel et à recueillir des micros-données longitudinales qui permettent d'effectuer des analyses détaillées DANS LE TEMPS, le temps étant un élément essentiel dans toute tentative pour s'attaquer aux changements, aux processus et à la dynamique.

Les analyses par panel imposent une charge de travail considérable aux chercheurs qui devront passer beaucoup de temps à se familiariser avec la structure des données du panel et avec les procédures à appliquer pour exploiter ces données.

Un seul panel pourrait suffire et la majorité des personnes qui analysent des données recueillies au moyen de panels travaillent effectivement avec un seul panel qui, dans tous les cas connus, est celui qui porte sur leur propre pays.

---

[1] Prof. Dr. Dr.h.c. G. Schaber, président du CEPS/INSTEAD; Université de Liège, Belgique; Clark University, Massachusetts. G. Schmaus, chercheur principal au CEPS/INSTEAD; Luxembourg. M. G.G. Wagner, Institut allemand pour la recherche économique (DIW, Berlin); depuis septembre 1992, professeur à l'Université de la Ruhr, Bochum, Allemagne.

Règle générale, jusqu'à aujourd'hui on n'a traité que des données et des problèmes relatifs à un seul pays. On possède peu de connaissances à propos des différences et des similitudes entre les pays.

Les chercheurs en sciences sociales devront se tourner plus régulièrement vers des utilisations transnationales et vraiment comparatives des données recueillies au moyen de panels, afin de mieux connaître les divers systèmes nationaux qui régissent la fiscalité, la sécurité et la protection sociales ou le marché du travail, afin de connaître la façon dont ces systèmes fonctionnent et comment ils exercent un effet sur les divers groupes et sur les diverses catégories de personnes qui forment l'ensemble de la population.

Certaines équipes de recherche se sont déjà groupées afin d'effectuer collectivement des comparaisons internationales bien définies basées sur les données recueillies au moyen de divers panels nationaux, ces données étant traitées selon des normes généralement acceptées. Dans un cas, un groupe de travail, dont le coordonnateur est Greg Duncan (voir Duncan, G.T. et coll. 1991) qui travaille à la PSID, a été constitué. Ce groupe comprend des chercheurs du Canada, de la France, de l'Allemagne, de l'Irlande, du Luxembourg, des Pays-Bas, de la Suède et des États-Unis.

Dans un deuxième cas, des groupes de travail qui s'intéressent à des sujets tels que le marché du travail, la mobilité et la pauvreté ont été créés dans le cadre du Network on Household Panels de la ESF. Mais chaque ensemble de données correspondant a été traité séparément.

Les méthodes mentionnées ci-après sont plus exigeantes: une équipe dirigée par Richard Burkhauser et Tim Smeeding à la University of Syracuse aux É.-U. commence maintenant à grouper l'ensemble des données recueillies au moyen du panel allemand (voir Burkhauser, R.V.) ainsi que les données obtenues dans le cadre de la PSID aux États-Unis.

Avec le projet COPA nous tenterons de regrouper, au Luxembourg, des données recueillies au moyen de plusieurs panels européens ainsi que celles obtenues dans le cadre de la PSID aux É.-U.

2. LE PROBLÈME LIÉ AUX DONNÉES RECUEILLIES AU MOYEN D'UN PANEL

Il est difficile d'effectuer des recherches portant sur plus d'une nation à l'aide d'ensembles de données recueillies au moyen de panels nationaux parce que chacun des ensembles de données nationaux est structuré de façon différente et qu'il utilise une présentation qui lui est propre. En résumé, la situation est qu'on ne retrouve:

- aucun nom de variable commun,
- aucune présentation commune,
- aucun logiciel commun,
- aucune gestion à l'aide d'un système de base de données identique,
- aucun système de stockage commun, par ex., sous forme de fichiers système SPSSX/SAS.

Il n'existe pas de base de données centrale dans laquelle on pourrait stocker les divers ensembles de données nationaux. De plus, l'utilisation de tout logiciel de gestion de base de données crée des problèmes parce que la majorité des chercheurs ne sont pas prêts à se familiariser avec les complexités de différents systèmes de banque de données. Ils désirent travailler avec des progiciels statistiques qu'ils connaissent.

Actuellement, seules des équipes auxquelles participent activement des personnes travaillant aux divers panels nationaux peuvent effectuer des études comparatives internationales sur les données recueillies au moyen de panels.

Les chercheurs isolés ne peuvent progresser dans des analyses comparatives sans aide ou sans collaborer étroitement avec les équipes travaillant à chaque panel national.

Sans banque de données centrale ou commune, il est pratiquement impossible de s'attaquer **SYSTÉMATIQUEMENT** aux tâches à entreprendre pour normaliser chacune des variables de chacun des panels, afin d'établir en détail les définitions et les concepts requis pour effectuer des analyses harmonisées.

3. LA MÉTHODE COPA (COMPARABILITÉ DE PANELS)

Afin de surmonter ces problèmes, le CEPS/INSTEAD crée, au Luxembourg, d'abord en collaboration avec le DIW Berlin (Deutsches Institut für Wirtschaftsforschung - Institut allemand pour la recherche économique), une base de données comparative internationale renfermant des données recueillies au moyen de divers panels nationaux de ménages.

Le but du projet COPA est d'élaborer des instruments pour analyser, programmer et simuler des politiques socio-économiques. Ce projet vise à faciliter les recherches comparatives transnationales sur des **QUESTIONS DE POLITIQUE** comme l'activité, la distribution du revenu, la pauvreté, les problèmes des personnes âgées et ainsi de suite.

Au niveau le plus bas, dans le cadre du projet COPA, on élaborera progressivement des archives de données recueillies au moyen de panels de ménages existant en Europe et aux É.-U. Pour la COPA, on commencera par regrouper les données recueillies au moyen de panels dans le cadre de la PSID (voir Hill, M.S. 1992) (É.-U.), du SOEP (voir Wagmer, G.G. et coll. 1991) (Allemagne) et du PSELL (voir Hausman, P. 1987; Schmaus, G. 1987) (Luxembourg). Les nouveaux panels de ménages qui commencent à être utilisés en Europe seront ajoutés quand les ensembles de données seront disponibles.

Au deuxième niveau, qui est plus important, le projet COPA accroîtra la valeur des données originales recueillies au moyen d'un panel en rendant ces données **COMPATIBLES** et **COMPARABLES**. Cela signifie que la base de données de la COPA renfermera des variables harmonisées et normalisées au niveau transversal **ET** au niveau longitudinal, avec des noms de variables identiques correspondant à un plan commun établi pour définir et recoder les variables. La méthode comparative stricte utilisée fait que ce projet est unique[2] actuellement.

Voici les caractéristiques de la base de données COPA:

- accès à des variables harmonisées découlant de données recueillies au moyen de panels,
- accès aux variables de la LIS,
- possibilité d'accéder aux variables originales,
- noms de variables normalisés,
- présentation commune,
- logiciel commun,
- stockage dans une structure de base de données **relationnelle**,
- stockage sous forme de fichiers système SPSSX,
- possibilité de sortie de données brutes.

Au cours d'une deuxième étape, on ajoutera à la base de données un système de documentation (**MÉTA-BANQUE DE DONNÉES**) et, nous l'espérons, au cours d'une troisième étape on ajoutera une **BASE DE DONNÉES INSTITUTIONNELLES**.

MÉTA-BANQUE DE DONNÉES: On prévoit intégrer tous les renseignements nécessaires sur les variables originales et normalisées dans le système de documentation (sur ordinateur personnel) que le CEPS/INSTEAD a élaboré pour son propre panel de ménages. De la documentation additionnelle à propos des variables comparables nouvellement créées sera préparée, sous forme exploitable par une machine et sur support papier. Mais, au cours de notre première étape, nous ne pouvons que recueillir les manuels originaux de l'utilisateur de chacune des études par panel et les mettre à la disposition des utilisateurs des données recueillies pour le projet COPA.

---

[2] Par exemple, le projet de la University of Syracuse («Cross national Studies in Aging») regroupe toutes les données recueillies au moyen du panel de la SOEP en Allemagne ainsi que toutes les données recueillies à l'aide de la PSID aux É.-U. sur le même ordinateur et dans un environnement logiciel commun. Mais les deux bases de données ne sont pas harmonisées au niveau des variables.

BASE DE DONNÉES INSTITUTIONNELLES[3]: L'interprétation des résultats de la recherche transnationale effectuée à l'aide d'enquêtes par panel exige des renseignements adéquats sur le système de sécurité sociale, le système fiscal, le système d'éducation, etc. des pays. Quand nous créerons la base de micro-données, nous devrons élaborer la base de données institutionnelles de façon à ce que cette dernière soit étroitement liée à la base de micro-données. Dans ce domaine, des techniques très pratiques ont été élaborées et des documents très utiles ont été recueillis dans le cadre du projet de la LIS.

La base de données COPA devrait renfermer **LE PLUS POSSIBLE DE VARIABLES COMPARABLES**. Pour chaque panel, on utilise un ensemble de questions qui sont identiques d'un cycle à l'autre. Ces **QUESTIONS DE BASE** correspondent aux premières variables devant être normalisées. En faisant une sélection parmi les questionnaires des enquêtes par panel des divers pays pour trouver les questions de base que renferment chacun de ces questionnaires, nous pourrons constituer la première liste de variables de base:

LISTE DE VARIABLES DE BASE:

- variables démographiques,
- variables sur le revenu,
- variables sur la population active,
- variables sur le chômage,
- variables sur l'éducation,
- variables sur le logement,
- variables chronologiques.

Afin d'inclure la PSID dans le projet COPA, l'ensemble des questions de base pour toutes les études par panel incluses est plutôt limité parce que la PSID n'est pas réellement une étude par panel portant sur des **personnes** mais plutôt sur des chefs de ménage. La PSID renferme beaucoup de renseignements, obtenus par personne interposée, sur le conjoint du répondant, mais très peu à propos des autres adultes membres des ménages interviewés. Les enquêtes par panel européennes recueillent beaucoup plus de renseignements à propos des adultes membres des ménages qui ne sont ni le chef du ménage, ni le conjoint de ce dernier.

Une DEUXIÈME LISTE renfermera les variables relatives aux **ANTÉCÉDENTS DES PERSONNES** avant que ces dernières ne fassent partie de l'étude réalisée à l'aide d'un panel. Les sujets ci-après sont disponibles dans la majorité des fichiers et peuvent être harmonisés:

- antécédents familiaux,
- antécédents scolaires,
- antécédents professionnels,
- antécédents matrimoniaux,
- antécédents en matière de fécondité.

Pour toute enquête nationale réalisée à l'aide d'un panel, on dispose ou on peut disposer, en plus des sujets et questions de base, d'éléments particuliers qui ne sont pas inclus dans de nombreuses autres enquêtes réalisées à l'aide d'un panel ou qui ne sont inclus que dans un ou un nombre limité de cycles d'une même enquête. De tels éléments se prêtent très mal à une harmonisation et ils ne seront stockés que sous leur forme originale.

EN RÈGLE GÉNÉRALE, les **FICHIERS RÉSULTATS DE LA COPA** devraient renfermer toutes les variables qui peuvent être normalisées. **DE PLUS**, l'utilisateur de ces fichiers pourra **accéder** aux **VARIABLES ORIGINALES** des études par panel qui n'ont pas été rendues comparables pour une raison ou une autre. Cette procédure permet aux chercheurs d'accéder simultanément aux variables originales et aux variables harmonisées.

Dans le cadre du projet COPA, nous créerons des fichiers au **NIVEAU DES MÉNAGES** et au **NIVEAU DES PARTICULIERS**. Chaque fichier renfermera des variables pour un an et pour un ensemble de données

---

[3] En collaboration avec la LIS et le projet ASEG de l'Université de Francfort (directeur, prof. Richard Hauser).

recueillies au moyen d'un panel. Des identificateurs additionnels assureront qu'il est possible d'effectuer des **APPARIEMENTS** et des **REGROUPEMENTS** entre les fichiers individuels.

Tous les fichiers ne seront pas conservés dans un système de gestion de base de données. Au CEPS, les fichiers COPA sont stockés sous forme de fichiers SPSS. Il est très facile d'exporter ces fichiers à destination du progiciel SAS.

Bien que cela aille à l'encontre des idées reçues par les informaticiens, on peut créer une structure de données relationnelles sans utiliser un système de gestion de base de données complet comme les produits faisant appel au SQL tels que ORACLE et d'autres logiciels comme INGRES. Des progiciels statistiques (voir Schmaus, G. 1992; Witte, J. 1992) bien connus comme SPSS et SAS permettent aussi de stocker et d'accéder, de façon relationnelle, à des données recueillies au moyen d'un panel.

De plus, les progiciels présentent l'avantage qu'un chercheur qui connaît «son» progiciel n'a pas à se familiariser avec les complexités d'un système de gestion de base de données. Cela permet aux statisticiens de créer des fichiers de travail normalisés pour les chercheurs et d'établir des fichiers spécifiques pour des analyses très particulières.

Nous prévoyons normaliser les variables, les structures des fichiers et le système d'accès de façon à ce que l'analyse des données recueillies dans le cadre de différentes études par panel, dans un contexte transnational et longitudinal, soit possible avec un **MINIMUM DE MODIFICATIONS AUX PROGRAMMES** rédigés pour un pays. Cela sera le cas, au moins, pour les totalisations et les analyses normalisées. Des analyses plus complexes ne pourraient probablement pas être normalisées de cette façon, mais elles seront étayées de façon efficace par l'organisation des données.

Dans les premières phases de ce projet, toutes les variables harmonisées requises ne sont pas déjà disponibles. Par conséquent, il se peut qu'un chercheur doive créer certaines variables harmonisées à partir des fichiers d'archives de données non normalisées et apparier ces variables avec celles provenant de la base de données harmonisées recueillies au moyen de panels. Ce travail ne sera pas difficile parce que des identificateurs uniques permettront d'apparier les deux genres de fichiers.

4. LA TOUTE PREMIÈRE COPA - EXEMPLE

La dynamique du marché du travail constitue un bon exemple pour démontrer les **possibilités** des fichiers COPA ainsi que les **limites** imposées par les différents concepts utilisés dans les études par panel.

Les variables principales de la COPA disponibles actuellement nous permettent de nous attaquer vigoureusement à des questions portant sur le marché du travail et sur la dynamique du revenu ainsi qu'à des variables démographiques. Nous ne pouvons faire beaucoup plus actuellement à cause du nombre restreint de variables disponibles au niveau des particuliers dans la PSID.

Pour les analyses comparatives entreprises dans le cadre du projet COPA nous devrons, une fois que les données de la PSID seront incluses, utiliser un plus petit nombre de variables que ce n'est le cas sans ces données (puisqu'à ce sujet, la PSID ne porte que sur les chefs de ménage et les conjoints de ces derniers)[4].

Une des questions fondamentales relatives à la dynamique du marché du travail aux États-Unis et en Europe est l'importance des «emplois marginaux». Il s'agit d'emplois instables et mal rémunérés qui comptent très peu d'heures de travail. Par exemple, en Allemagne, tous les emplois comportant moins de vingt heures de travail par semaine sont considérés des emplois «marginaux», parce que le système de sécurité sociale ne s'applique pas à ces emplois. C'est une critique courante, du point de vue européen, de dire que la «machine des emplois»

[4] Comme nous l'avons mentionné plus haut, la PSID fait appel à un panel basé sur les réponses des chefs de ménage. Pour les ménages composés de deux personnes et plus, il y a beaucoup de renseignements (obtenus par personne interposée) à propos du conjoint du chef mais très peu à propos des autres membres du ménage, alors que dans les enquêtes par panel européennes tous les adultes membres du ménage sont interviewés.

américaine fonctionne au moyen d'emplois marginaux qui rendent impossible, aux personnes qui occupent de tels emplois, d'atteindre une position de bien-être raisonnable. De plus, ces emplois pourraient constituer une raison importante du ralentissement de la productivité aux États-Unis. Par contre, les emplois marginaux sont des emplois caractéristiques pour les conjoints qui ne sont pas intéressés à occuper un emploi à plein temps. Les emplois marginaux permettent de hausser le niveau de bien-être de ce groupe particulier.

Une comparaison des États-Unis, du Luxembourg et de la République fédérale d'Allemagne peut être utile parce que les structures globales du marché du travail y sont très différentes. Le marché du travail des États-Unis est moins réglementé (comparativement à la situation en Europe), le taux de chômage y est moyen et l'offre de main-d'oeuvre féminine élevée. En Allemagne, le marché du travail est strictement réglementé, le taux de chômage est élevé (comparativement à d'autres pays à salaire élevé) et l'offre de main-d'oeuvre féminine est faible. Sur le plan économique, le Luxembourg semble, dans les années 80, un pays très riche avec une réglementation du travail stricte, un plein emploi (qui crée des emplois occupés à plus de 60% par des navetteurs transfrontaliers) et une faible offre de main-d'oeuvre féminine.

Seules les données recueillies au moyen de panels permettent au chercheur de s'attaquer à la question de savoir si les emplois marginaux sont des emplois permanents pour les chefs de ménage ou des emplois occupés par les conjoints de ces derniers pendant une courte période. De plus, les données recueillies au moyen de panels permettent d'effectuer des recherches afin d'estimer des fonctions d'offre de main-d'oeuvre très complexes. Pour le moment, nous pouvons présenter certains résultats descriptifs préliminaires. Ces derniers montrent qu'il est possible d'effectuer une analyse comparative des données recueillies au moyen de panels, mais ils démontrent aussi que, même pour des chiffres descriptifs simples, beaucoup de problèmes se présentent quand on veut rendre des variables comparables.

Afin d'analyser la dynamique du marché du travail pour certaines catégories d'emploi, nous définissons une variable d'état pour le travail:

- personne inactive: 0 heure d'activité
- personne marginalement occupée: 1-19 heures de travail hebdomadaire
- personne occupée à temps partiel: 20-29 heures de travail hebdomadaire
- personne occupée à plein temps: 30 heures de travail hebdomadaire et plus

À l'aide des données recueillies au moyen du panel allemand, par exemple, nous savons qu'environ 50% des personnes qui occupent un emploi marginal considèrent qu'elles ne font pas partie de la population active. Cela signifie qu'elles répondent «chômeur (chômeuse)». On doit aussi poser une question à propos des «deuxièmes emplois» («Nebenerwerbstätigkeit») pour saisir leurs emplois marginaux. Dans la PSID, les intervieweurs doivent retourner à la section sur l'emploi si le répondant occupe un deuxième emploi. Il n'est donc pas nécessaire de grouper un premier et un deuxième emploi au cours de l'analyse parce que ce renseignement figure déjà dans l'enregistrement de la PSID. Dans le panel socio-économique luxembourgeois («PSELL»), il n'y a pas de question à propos d'un emploi marginal en plus des questions portant sur la situation relative à l'emploi principal. Cela peut causer un artefact. Mais, jusqu'ici, nous n'avons pas trouvé de meilleure solution pour rendre les trois enquêtes par panel compatibles.

Afin d'éviter des effets particuliers du processus de transition à la retraite dans nos résultats longitudinaux, notre analyse porte seulement sur les personnes de 16 à 50 ans.

Même un examen des résultats transversaux pour les femmes nous surprend: pour les trois pays, c'est aux États-Unis que la part de l'activité à plein temps des femmes est la plus élevée, la part pour le Luxembourg et pour l'Allemagne est beaucoup plus faible. Le taux de travail à temps partiel est le plus élevé en Allemagne et il ne l'est pas autant aux É.-U. et au Luxembourg. C'est en Allemagne que la part des emplois marginaux est la plus élevée, elle est très faible aux É.-U. et au Luxembourg. Le chômage constitue un problème pour l'Allemagne et les États-Unis seulement. À cause de la faible activité et du chômage très peu élevé au Luxembourg, c'est dans ce pays que la part des femmes inactives est la plus élevée.

Un examen des **résultats longitudinaux** donne l'impression globale que la dynamique du marché du travail est fort semblable dans les trois pays. Mais il est remarquable qu'aux É.-U., la probabilité qu'un emploi marginal

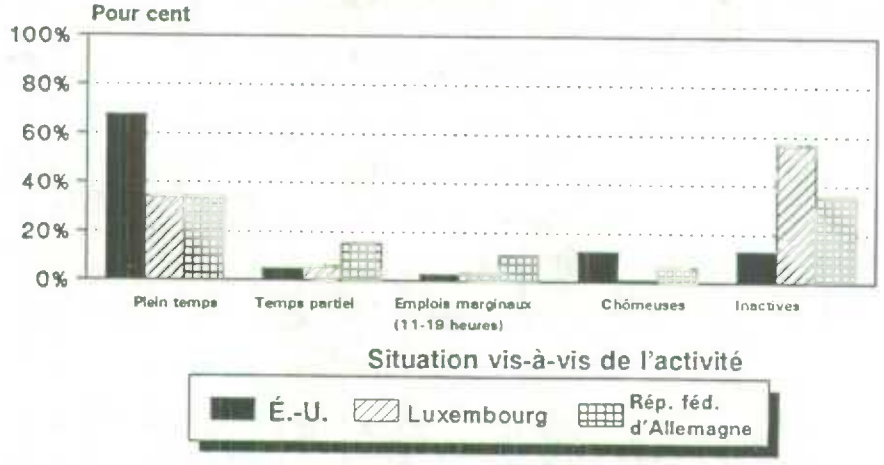
mènera à un emploi à plein temps est élevée. La République fédérale d'Allemagne présente la probabilité la plus élevée de rester dans un emploi marginal. C'est au Luxembourg que la probabilité de progresser pour obtenir un emploi régulier à temps partiel est la plus élevée.

Ces résultats ne correspondent pas à nos idées préconçues à propos des marchés du travail en Europe et aux É.-U. Ils stimulent donc d'autres recherches. Cela constitue une première étape encourageante pour la COPA.

Tableau 1.

Activité

Femmes, âgées de 16 à 50 ans, chefs de ménage et conjointes
É.-U., Luxembourg et République fédérale d'Allemagne

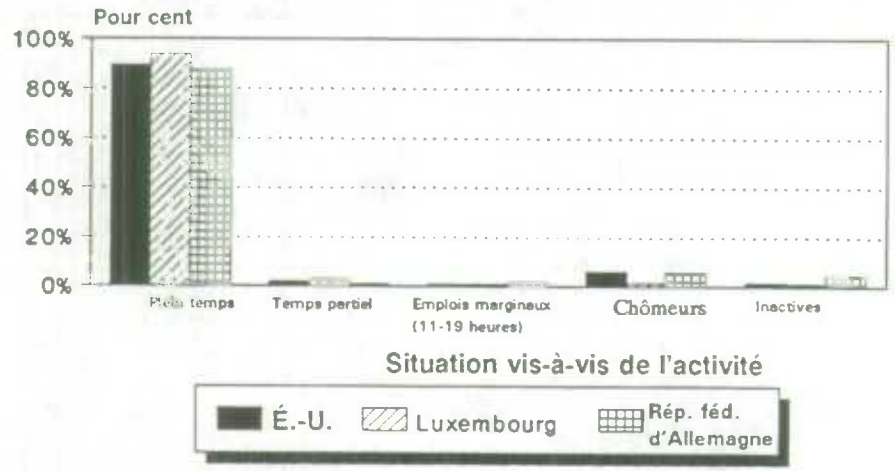


Sources : Fichiers COPA de la PSID (1983), du PSELL (1985) et du SOEP (1985)

Tableau 2.

Activité

Hommes, âgés de 16 à 50 ans, chefs de ménage et conjoints É.-U., Luxembourg et République fédérale d'Allemagne

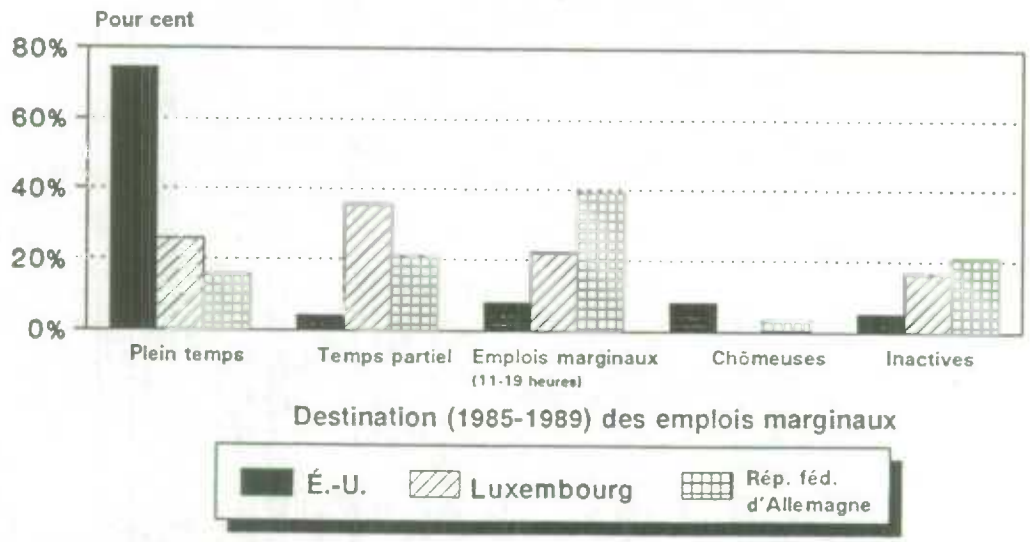


Sources: Fichiers COPA de la PSID (1983), du PSELL (1985) et du SOEP (1985)

Tableau 3.

Dynamique des emplois marginaux

Femmes, âgées de 16 à 50 ans, chefs de ménage et conjointes É.-U., Luxembourg et République fédérale d'Allemagne



Sources: Fichiers COPA de la PSID (1983), du PSELL (1985) et du SOEP (1985)

BIBLIOGRAPHIE

Burkhauser, R.V. An introduction to the German Socio-economic panel for English speaking researchers. Cross-National Studies in Aging, Program Project Paper No. 1. All-University Gerontology Center, Maxwell School of Citizenship and Public Affairs, Syracuse University New York.

Duncan, G.J., Gustafsson, B., Hauser, R., Schmaus, G., Laren, D., Messinger, H., Muffels, R., Nolan, B., et Ray, J.-C. (1991). Poverty dynamics in eight countries, (mimeo). Survey Research Center Ann Arbor, MI, U.S.A. (octobre).

Hausman, P. (1987). Niveaux de vie et de bien-être économique des ménages en 1985: principaux résultats. Document de travail PSELL, numéro 4. CEPS/INSTEAD, Walferdange, Luxembourg.

Hill, M.S. (1992). The Panel Study of Income Dynamics - A Users's Guide Newbury Park, London, New Delhi.

Schmaus, G. (1987). Organization of the database for the Luxembourger household panel (input, storage and analysis). Document de travail PSELL numéro 17a. CEPS/INSTEAD, Walferdange, Luxembourg.

Schmaus, G. (1992). Storage and retrieval of data from the Luxembourg. Household Panel (PSELL). Document de travail ESF, numéro 9. CEPS/INSTEAD and University of Essex.

Smeeding, T.M., et Schmaus, G. (1988). LIS Information Guide (Révisé, novembre). Document de travail LIS numéro 7, CEPS/INSTEAD, Walferdange.

Smeeding, T.M., et Schmaus, G. (1990). The LIS Database: Technical and Methodological Aspects. Dans (Éds.) T.M. Smeeding, M. O'Higgins et L. Rainwater. Poverty, Inequality and Income distribution in Comparative Perspective Hemel Hempstead Herfordshire.

Wagner, G.G. et coll. (1991). The socio-economic panel (SOEP) of Germany - methods of production and management of longitudinal data. DIW Discussion Paper Number 31a, Berlin.

Witte, J. (1992). Data management and analysis of the German socio-economic panel using SPSS, DIW documentation. Berlin.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

ANALYSE DE DONNÉES D'ESSAI LONGITUDINALES SUR LES ENTREPRISES AVEC VARIABLES NOMINALES ORDONNÉES

G. Arminger[1]

RÉSUMÉ

La plupart des données d'essai longitudinales sur les entreprises proviennent de questionnaires. C'est pourquoi un grand nombre des variables dépendantes sont soit des variables non métriques, par exemple dichotomiques, soit des variables à distribution censurée ou encore des variables nominales ordonnées. Dans cette communication, nous élargissons de deux manières le modèle de Heckman (1981) pour données longitudinales dichotomiques. Premièrement, nous transformons le modèle de seuil dichotomique en des modèles de seuil de type probit pour les variables nominales classées selon un ordre numérique, les variables à distribution censurée ou les variables nominales ordonnées. Deuxièmement, nous transformons le modèle à équation unique en un modèle à système d'équations et en un modèle factoriel d'évaluation. Nous accordons une attention particulière aux difficultés propres aux panels, comme les états initiaux et les contraintes auxquelles est assujettie la matrice des covariances de l'erreur. À titre d'exemple, nous spécifions et estimons un modèle de décisions simultanées concernant les prix, la production et les stocks pour des données d'essai sur les entreprises allemandes. Nous nous servons du programme MECOSA pour estimer les paramètres du modèle.

MOTS CLÉS: Modèle de seuil dichotomique; coefficient de corrélation polychorique; distribution censurée.

1. INTRODUCTION

L'analyse de données d'essai sur les entreprises consiste essentiellement en une analyse de données d'enquête non métriques recueillies auprès d'un échantillon d'entreprises. Afin de suivre l'évolution des entreprises dans le temps, on interview durant un certain nombre de périodes consécutives les unités qui participent à une enquête-entreprises et on obtient ainsi un ensemble de données longitudinales. Dans la section qui suit, nous étendons les principes d'interprétation exposés dans les ouvrages précurseurs de Heckman (1981a, 1981b) sur la spécification et l'estimation de variables de résultat dichotomiques aux variables nominales ordonnées et aux variables dépendantes à distribution censurée ainsi qu'aux systèmes d'équations simultanées de variables dépendantes non métriques. On estime les paramètres de ces modèles en supposant que les termes d'erreur des modèles de seuil suivent une distribution normale multidimensionnelle et en se servant de coefficients de corrélation polychoriques et polysériaux conditionnels dans les modèles de structure de moyennes et de covariances pour des variables dépendantes non métriques. Ces modèles et ces méthodes d'estimation ont été proposés à l'origine par Muthén (1984), puis perfectionnés par Küsters (1987) et Schepers et Arminger (1992). Nous accordons une attention particulière aux problèmes de l'hétérogénéité non observée, des états initiaux et de la définition de l'échelle. Nous limitons l'analyse et l'exemplification des modèles et des méthodes au cas le plus courant (dépendance à l'égard de l'état et hétérogénéité non observée aléatoire). À titre d'exemple, nous analysons la variable trichotomique de production pour un panel à quatre cycles constitué de 656 entreprises tirées de l'échantillon de l'enquête que mène périodiquement auprès des entreprises allemandes l'institut IFO de Munich.

[1] G. Arminger, Département des sciences économiques (FB 6), Bergische Universität, Gaußstr. 20, D-5600 Wuppertal, Allemagne.

2. SPÉCIFICATION DU MODÈLE

2.1 Le modèle de Heckman pour variables dichotomiques

Heckman (1981a, chap. 3.3) considère le modèle suivant pour une variable non observée y_{it}^*, $i = 1,...,n$, $t = 1,...,T$ où i désigne l'unité et t, une suite de points équidistants dans le temps:

$$y_{it}^* = \mu_{it} + \epsilon_{it}^* \tag{1}$$

$$\mu_{it} = x_{it}\beta + \sum_{j=1}^{\infty} \gamma_{t-j,t} y_{i,t-j} + \sum_{j=1}^{\infty} \lambda_{j,t-j} \prod_{l=1}^{j} y_{i,t-l} + \sum_{k=1}^{K} \delta_k y_{i,t-k}^* \tag{2}$$

$$\epsilon_i^* = (\epsilon_{i1}, ..., \epsilon_{iT}^*)', \quad \epsilon_i^* \sim N(0, \Sigma), \tag{3}$$

$$y_{it} = \begin{cases} 1 \text{ if } y_{it}^* > 0 \\ 0 \text{ if } y_{it}^* \leq 0 \end{cases}. \tag{4}$$

La variable non observée y_{it}^* est considérée comme une variable utilitaire qui est liée à la variable dépendante observée y_{it} par l'intermédiaire d'un modèle de seuil dichotomique avec comme seuil 0. Les valeurs y_{it}, y_{it}^* et x_{it} sont groupées respectivement dans les vecteurs $T \times 1$ y_i et y_i^* et dans le vecteur $R \times 1$ $x_i = (x_{i1},...,x_{iT})'$. Les variables aléatoires $\{y_i, x_i\}$ sont indépendantes et identiquement distribuées, ce qui équivaut à un échantillon aléatoire simple d'une population. La spécification du modèle consiste à déterminer la structure de la composante systématique, μ_{it}, et de la composante stochastique, ϵ_{it}^*, du modèle. On trouvera dans Heckman (1981a) et dans Hamerle et Ronning (1993) une analyse des diverses étapes de la spécification du modèle pour y_{it}^*. Nous ne reprenons ici que les principaux éléments de la spécification.

La première composante de μ_{it} représente la variation engendrée par les variables explicatives x_{it} qui varient probablement dans le temps. Le vecteur de paramètres β est temporellement constant dans ce modèle; il peut toutefois être transformé en vecteurs de paramètres β_t, $t = 1,...,T$, qui varient dans le temps.

La deuxième composante de μ_{it} traduit l'influence des états antérieurs de la variable dépendante observée $y_{i,t-j}$, $j \geq 1$, ce que l'on appelle la dépendance vraie à l'égard de l'état. Si $\gamma_{t-1,t} = \gamma_1 \neq 0$ et $\gamma_{t-j,t} = 0$ pour tous $j > 1$ et $t = 1,...,T$, nous avons un modèle markovien simple. Notons que pour pouvoir tenir compte des états antérieurs d'une variable, il faut connaître les états initiaux y_{i0}, $y_{i,-1}$,... selon la spécification des paramètres $\gamma_{t-j,t}$. Si les états initiaux sont connus et non stochastiques, on peut les inclure dans les vecteurs x_{it} comme des variables explicatives supplémentaires. Si ces états sont eux-mêmes des résultats du processus qui génère y_{it}^*, on doit tenir compte de leur distribution, comme il est mentionné dans Heckman (1981b) et dans la section 3 de cette communication. Notons que les effets des états antérieurs $y_{i,t-j}$ peuvent varier à chaque période. Ce phénomène est traduit par $\gamma_{t-j,t}$. Dans la plupart des applications, $\gamma_{t-j,t}$ équivaut à γ_{t-j} et presque tous les paramètres sont posés égal à 0.

La troisième composante de μ_{it} explique la dépendance de y_{it}^* à l'égard de la durée de l'état $y_{i,t} = 1$. Notons ici aussi que les effets de la durée peuvent être différents à chaque période. Ces effets sont paramétrisés en $\lambda_{j,t-j}$. Si $\lambda_{j,t-j} = \lambda$, nous avons le cas le plus simple, c'est-à-dire un effet de durée linéaire.

La quatrième composante de μ_{it} explique la dépendance de y_{it}^* à l'égard des valeurs antérieures $y_{i,t-j}^*$ des variables endogènes non observées. Les modèles qui renferment cette composante sont appelés modèles avec formation des habitudes ou persistance des habitudes. Le principe de base de ces modèles est que y_{it}^* dépend non pas de l'état antérieur réel de la variable observée mais de l'état ou de l'habitude antérieurs associés à $y_{i,t-j}^*$ plutôt qu'à $y_{i,t-j}$. Si les variables utilitaires initiales $y_{i,0}^*, y_{i,-1}^*, ..., y_{i,-(K-1)}^*$ sont connues et non stochastiques, elles

peuvent être incluses dans la liste des variables explicatives, sinon il faut poser des hypothèses sur leur distribution.

Passons maintenant à la spécification du terme d'erreur ϵ_{it}^*. Dans cette communication, nous supposons qu'il n'existe aucune corrélation entre les termes d'erreur et les variables explicatives passées, présentes et futures (forte exogénéité). La méthode que proposent Keane et Runkle (1992) pour traiter les cas de faible exogénéité ne peut être appliquée directement aux variables dépendantes limitées. Le terme d'erreur ϵ_{it}^* se décompose habituellement comme suit:

$$\epsilon_{it}^* = \alpha_i + \epsilon_{it}, \tag{5}$$

où α_i désigne le terme d'erreur qui varie selon les individus mais demeure fixe dans le temps et qui peut être considéré comme de l'hétérogénéité non observée, comme dans le cas des données métriques (voir Hsiao 1986). Les valeurs de α_i peuvent être considérées comme des effets fixes pour chaque i ou comme des effets aléatoires distribués selon $\alpha_i \sim N(0,\sigma_\alpha^2)$. Dans le premier cas, α_i est un paramètre propre à l'individu. Si y_{it}^* est défini par le modèle $y_{it}^* = x_{it}\beta + \alpha_i + \epsilon_{it}^*$ et que les valeurs correspondantes sont observées comme dans le cas des données métriques, on peut supprimer α_i en calculant les différences premières $y_{it}^* - y_{i,t-1}^* = (x_{it} - x_{i,t-1})\beta + \epsilon_{i,t}^* - \epsilon_{i,t-1}^*$. Cette technique ne s'applique pas dans le cas des modèles non métriques comme le modèle probit. Lorsqu'il s'agit d'un modèle logit dichotomique, on peut supprimer les α_i en se servant d'une statistique exhaustive comme élément de condition, comme le montrent Hsiao (1986) et Hamerle et Ronning (1993). Si α_i est une variable aléatoire, on la suppose non corrélée avec z_{it} et ϵ_{it}. Si ϵ_{it} a une variance constante σ_ϵ^2 et qu'elle est non liée, ϵ_i^* a la structure de covariance caractéristique suivante:

$$V(\epsilon_i^*) = \begin{bmatrix} \sigma_\alpha^2 + \sigma_\epsilon^2 & & & \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\epsilon^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_\epsilon^2 \end{bmatrix} = \sigma_\alpha^2 11' + \sigma_\epsilon^2 I,$$

où 1 est un vecteur $T \times 1$ formé de uns et I est la matrice unité $T \times T$. D'une manière plus générale, on peut supposer une structure en série ou une structure factorielle pour $V(\epsilon_i^*)$. Heckman (1981a) et Arminger (1992) donnent plus de détails à ce sujet. Heckman (1981a) analyse l'estimation de ces modèles par la méthode du maximum de vraisemblance (MV) en supposant que ϵ_i^* est distribué normalement. Il suppose en outre que seuls des résultats dichotomiques sont observés. Comme $\epsilon_i^* \sim N(0, \sum)$, on peut estimer le vecteur de paramètres des variables explicatives à l'aide d'un modèle probit unidimensionnel pour chaque y_{it}, en prenant soin d'indiquer la restriction habituelle $\sigma_{tt} = 1, t = 1,..., T$, ce qui signifie que β ne peut être estimé qu'au multiple près. Notons que cette restriction n'est pas nécessaire pour tous les cycles de panel. Si β est le même pour tous les cycles, il suffit d'appliquer des restrictions à la variance de ϵ_{it}^* pour un seul cycle, le premier habituellement (voir Arminger 1987).

2.2 Application à des modèles de seuil généraux

Madalla (1987) présente une analyse approfondie de la question et étend les modèles de Heckman à des modèles logit et tobit à effets fixes ainsi qu'à des modèles probit à effets aléatoires. Nous allons maintenant étendre systématiquement les modèles dichotomiques de Heckman (1981a) à des variables dépendantes à distribution censurée, à des variables classées numériquement et à des variables nominales ordonnées, ainsi qu'à des systèmes d'équations simultanées dans lesquels on trouve comme variables dépendantes un mélange de variables métriques et de variables limitées. Nous ne considérons que les modèles à effets aléatoires. Comme Heckman (1981a), nous examinons le cas où les états initiaux sont connus et non stochastiques ou, si l'on veut, le cas où on détient de l'information sur la distribution des états initiaux de telle sorte qu'il est possible d'inclure les états initiaux avec les variables dépendantes dans un modèle ou d'inclure directement les états initiaux dans les variables explicatives. En l'occurrence, le vecteur $T \times 1$ y_i^* peut s'écrire

$$y_i^* = \gamma + \Pi x_i + \epsilon_i^* , \tag{6}$$

où ϵ_i^* suit une distribution normale de moyenne 0 et de variance $\epsilon_i^* \sim N(0, \sum)$. Le vecteur $R \times 1$ x_i contient les variables explicatives $x_{it}, t = 1,...,T$. Le vecteur $T \times 1$ γ est le vecteur des constantes de régression. La matrice $T x R$ Π est la matrice des coefficients de régression. Si $\mu_{it} = x_{it}\beta_t$, Π est structurée de la façon suivante:

$$\Pi + \begin{pmatrix} \beta_1 & 0 & \dots & 0 \\ 0 & \beta_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \beta_T \end{pmatrix} . \tag{7}$$

Lorsqu'il s'agit de paramètres fixes dans le temps, le vecteur β_t est remplacé par β.

La spécification de Σ par l'intermédiaire d'un modèle pour hétérogénéité non observée et corrélation sériale ainsi que la spécification de $y_i^* = \gamma + \Pi x_i + \epsilon_i^*$ décrite ci-dessus donnent une structure de moyennes et de covariances conditionnelles pour le vecteur de variables latentes y_i^*, $y_i^* \sim N(\gamma + \Pi x_i, \sum)$.

Le modèle ainsi élargi autorise non seulement le modèle de seuil dichotomique de l'équation (4), mais aussi n'importe lequel des modèles de seuil suivants qui applique y_{it}^* sur la variable observée y_{it} (voir Schepers, Arminger et Küsters 1991). Pour des raisons de commodité, nous omettons l'indice $i = 1,...,n$.

- y_t est une variable métrique (relation d'identité)

$$y_t = y_t^* . \tag{8}$$

- y_t est une variable nominale ordonnée avec des seuils inconnus $\tau_{t,1} < \tau_{t,2} < ... < \tau_{t,K_t}$ et des catégories $y_t = 1,...,K_t + 1$ (relation probit ordinale, McKelvey et Zavoina 1975).

$$\begin{gathered} y_t = k \iff y_t^* \in [\tau_{t,k-1}, \tau_{t,k}) \quad \text{où} \\ [\tau_{t,0}, \tau_{t,1}) = (-\infty, \tau_{t,1}) \quad \text{and} \quad \tau_{t,K_t+1} = +\infty . \end{gathered} \tag{9}$$

  Notons que pour des raisons d'identification, le seuil $\tau_{t,1}$ est posé égal à 0 et la variance du terme d'erreur de forme réduite σ_t^2 est posée égale à 1. Les paramètres en β_t ne sont définis qu'au multiple près. Si on considère des modèles d'équations simultanées ou qu'on analyse simultanément deux cycles de panel ou plus, on ne peut tester en règle générale que les hypothèses de la proportionnalité des coefficients de régression des diverses équations. Les hypothèses de l'égalité des coefficients de régression des diverses équations ne peuvent être testées que si l'on pose des hypothèses additionnelles, qui sont parfois non vérifiables (Sobel et Arminger 1992).

- Les variables classées numériquement peuvent être traitées de la même manière que les variables nominales ordonnées mentionnées ci-dessus, à la différence que les limites de classe tiennent lieu cette fois de seuils connus (Stewart 1983). On n'a pas à poser de restrictions concernant l'identification.

- y_t est une variable à distribution censurée unilatéralement, avec un seuil $\tau_{t,1}$ connu a priori (relation tobit, Tobin 1958).

$$y_t = \begin{cases} y_t^* & \text{si } y_t^* > \tau_{t,1} \\ \tau_{t,1} & \text{si } y_t^* \leq \tau_{t,1} \end{cases}. \tag{10}$$

- y_t est une variable à distribution censurée bilatéralement, avec des seuils $\tau_{t,1} < \tau_{t,2}$ connus a priori (relation probit à deux limites, Rosett et Nelson 1975).

$$y_t = \begin{cases} \tau_{t,1} & \text{si } y_t^* \leq \tau_{t,1} \\ y_t^* & \text{si } \tau_{t,1} < y_t^* < \tau_{t,2} \\ \tau_{t,2} & \text{si } \tau_{t,2} \leq y_t^* \end{cases}. \tag{11}$$

En ce qui concerne les modèles de seuil généraux, il faut envisager de faire plusieurs modifications aux modèles dichotomiques pour données de panel. Premièrement, il faut modifier la dépendance à l'égard de l'état pour les variables dépendantes non dichotomiques. En ce qui à trait aux variables dépendantes à distribution censurée, il devrait exister une variable fictive $d_{i,t-j}$ qui prend la valeur 1 si la variable $y^*_{i,t-j}$ est observée et la valeur 0 si $y_{i,t-j}$ correspond à la valeur de seuil. Pour ce est des variables classées numériquement et des variables nominales ordonnées, il faut définir un vecteur $K \times 1$ de variables fictives $d^{(k)}_{i,t-j}$. La variable fictive $d^{(k)}_{i,t-j}$ est égale à 1 si $y^*_{i,t-j}$ appartient à la catégorie k, $k = 2,...,K+1$, et est égale à 0 dans le cas contraire.

En deuxième lieu, il faut déterminer très soigneusement les restrictions relatives à l'identification qui doivent s'appliquer aux variances $\sigma_t^2, t = 1,...,T$ des ϵ^*_{it} et aux seuils inclus dans le vecteur $\tau^{(i)}$ pour chaque cycle de panel t. À notre avis, on devrait poser les vecteurs de seuils $\tau^{(i)}$ égaux pour tous les cycles de panels, sinon les catégories des variables nominales ordonnées n'auront censément pas la même signification d'une période à l'autre. Cette restriction suppose automatiquement qu'il n'est pas nécessaire d'assujettir les variances σ_t^2 à des conditions, sauf en ce qui concerne le premier cycle, et que ces variances peuvent varier d'un cycle à l'autre. En ce qui a trait aux variables dépendantes à distribution censurée, aucune restriction n'est requise pour σ_t^2.

Un autre aspect de la question a trait à la spécification de modèles pour un système de variables. Au lieu d'envisager une seule variable dans T cycles, on peut envisager un vecteur y_{it}^* de H variables dépendantes dans le temps. Chaque élément de y_{it}^* est désigné par y^*_{ith}, $h = 1,...,H$. Le vecteur de variables dépendantes y_i^* de l'observation i est donc un vecteur $H \cdot T \times 1$ de variables dépendantes observées à T périodes. Chaque variable latente y^*_{ith} observée à chaque période est alors appliquée sur l'observation y_{ith} au moyen d'un modèle de seuil comme celui défini ci-dessus (on peut utiliser des seuils différents pour chaque élément). Dans ces circonstances, la matrice de covariances, Σ, de ϵ_i^* contient non seulement la structure de covariance sériale conditionnelle pour chaque variable y^*_{ih}, mais aussi la structure de covariance conditionnelle des variables pour toutes les périodes. Arminger et Ronning (1991) donnent un exemple d'un modèle de ce genre.

3. MÉTHODE D'ESTIMATION

3.1 Coefficients de corrélation polychoriques conditionnels

Dans le cas de structures générales de moyennes et de covariances, nous supposons qu'un vecteur $P \times 1$ y_i^* de variables dépendantes latentes suit une distribution normale multidimensionnelle de moyenne et de covariance conditionnelles:

$$\begin{aligned} E(y_i^* | x_i) &= \gamma(\vartheta) + \Pi(\vartheta) x_i, \\ V(y_i^* | x_i) &= \Sigma(\vartheta). \end{aligned} \tag{12}$$

Dans l'analyse de données de panel, P est égal à T si c'est une variable dépendante unidimensionnelle qui est analysée et est égal à $H \cdot T$ si c'est une variable dépendante multidimensionnelle qui fait l'objet de l'analyse; $\gamma(\vartheta)$ est un vecteur $P \times 1$ de constantes de régression et $\Pi(\vartheta)$ est une matrice $P \times R$ de coefficients de

régression de forme réduite; x_i est un vecteur $R \times 1$ de variables explicatives; $\sum(\vartheta)$ est la matrice de covariance $P \times P$ des erreurs de forme réduite; ϑ est le vecteur $\tilde{q} \times 1$ des paramètres structurels à estimer. Les paramètres de forme réduite $\gamma(\vartheta)$, $\Pi(\vartheta)$ et $\sum(\vartheta)$ sont des fonctions continûment différentiables d'un vecteur commun ϑ. Parmi les exemples les plus courants, mentionnons les systèmes d'équations simultanées

$$y_i^* = By_i^* + \Gamma x_i + \epsilon_i \quad \text{où} \quad \epsilon_i \sim N(0,\Omega), \tag{13}$$

avec les paramètres de forme réduite

$$\Pi(\vartheta) = (I - B)^{-1}\Gamma \quad \text{et} \quad \sum(\vartheta) = (I - B)^{-1}\Omega(I - B)^{-1'} \tag{14}$$

de même que l'analyse factorielle confirmatoire

$$y_i^* = \Lambda\eta_i + \epsilon_i \quad \text{où} \quad \eta_i \sim N(0,\Phi) \quad \text{et} \quad \epsilon_i \sim N(0,\Theta) \tag{15}$$

avec les paramètres de forme réduite

$$\Pi(\vartheta) = 0, \quad \sum(\vartheta) = \Lambda\Phi\Lambda' + \Theta. \tag{16}$$

Dans le premier exemple, ϑ consiste dans les matrices de paramètres structurels B, Γ et Ω. Dans le second exemple, ϑ est constitué de Λ, de Φ et de Θ.

Notons que la structure caractéristique des modèles longitudinaux, c.-à-d. $\mu_{it} = x_i\beta$, ne peut s'apparenter directement à la forme réduite de l'équation (13). Si on se sert du modèle défini en (13) avec, comme vecteur de variables explicatives, $x_i = (x_{i1},...,x_{iT})'$, les variables dépendantes y_{it}^* font l'objet d'une régression non seulement par rapport à x_{it}, mais aussi par rapport à toutes les autres variables $x_{is}, s \neq t$. Les restrictions qui s'appliquent à la forme réduite et qui concernent les paramètres doivent être introduites à la troisième étape de la procédure d'estimation.

L'estimation du vecteur de paramètres structurels sur la base du vecteur d'observations y_i se fait en trois étapes. Le contenu de cette sous-section repose sur l'ouvrage de Schepers, Arminger et Küsters (1991). Le calcul des estimations à l'aide du programme MECOSA est décrit dans Schepers et Arminger (1992).

1. À la première étape, on estime les paramètres de seuil τ, les coefficients de forme réduite γ et Π de l'équation de régression ainsi que la variance d'erreur de forme réduite σ_t^2 de l'équation t au moyen de la fonction du maximum de vraisemblance marginale. Il convient de souligner qu'à cette étape, on estime la structure de moyennes pour le cas où il n'y a pas de restrictions, contrairement à l'équation (13), où il y en a. Les paramètres à estimer dans l'équation t sont les seuils, désignés par le vecteur τ_t, la constante de régression, désignée par γ_t, les coefficients de régression, c.-à.-d. la t-ième ligne de Π, désignée par $\Pi_{t.}$, ainsi que la variance, désignée par σ_t^2.

2. La deuxième étape consiste à estimer les covariances des termes d'erreur contenus dans les équations de forme réduite. Notons qu'à cette étape, on estime les covariances sans que des restrictions s'appliquent aux paramètres. Comme les erreurs sont supposées distribuées normalement et qu'on a déjà obtenu à la première étape des estimateurs fortement convergents des coefficients de forme réduite, le problème se réduit à maximiser la fonction de vraisemblance logarithmique

$$l_{ij}(\sigma_{ij}) = \sum_{i=1}^{n} \ell n\, P(y_{it}, y_{ij} \mid x_i, \hat{\tau}_t, \hat{\gamma}_t, \hat{\Pi}_{t.}, \hat{\sigma}_t^2, \hat{\tau}_j, \hat{\gamma}_j, \hat{\Pi}_{j.}, \hat{\sigma}_j^2, \sigma_{ij}) \ , \tag{17}$$

où $P(y_{it},y_{ij}\,|\,x_i,\hat{\tau}_i,\hat{\gamma}_i,\hat{\Pi}_{i\cdot},\hat{\sigma}_i^2,\hat{\tau}_j,\hat{\gamma}_j,\hat{\Pi}_{j\cdot},\hat{\sigma}_j^2,\sigma_{ij})$ est la probabilité bidimensionnelle de y_{it} et y_{ij} étant donné x_i et les coefficients de forme réduite. Un exemple caractéristique de cette probabilité bidimensionnelle est le cas où y_i et y_j sont toutes deux des variables ordinales. La probabilité que $y_{it} = k$ et $y_{ij} = l$ est donnée par l'équation

$$P(y_{it} = k, y_{ij} = l\,|\,x_i) = \int_{\hat{\tau}_{i,k-1}}^{\hat{\tau}_{i,k}} \int_{\hat{\tau}_{j,l-1}}^{\hat{\tau}_{j,l}} \varphi(y_i^*, y_j^*\,|\,\hat{\mu}_{it}, \hat{\sigma}_i^2, \hat{\mu}_{ij}, \hat{\sigma}_j^2, \sigma_{ij})\, dy_j^*\, dy_i^* \tag{18}$$

où $\hat{\mu}_{it} = \hat{\gamma}_i + \hat{\Pi}_{i\cdot}x_i, \hat{\mu}_{ij} = \hat{\gamma}_j + \hat{\Pi}_{j\cdot}x_i$ et $\varphi(y_i^*, y_j^*\,|\,\mu_i,\sigma_i^2,\mu_j,\sigma_j^2,\sigma_{ij})$ est la fonction de densité normale bidimensionnelle. Notons que pour des variables ordinales, $\sigma_i^2 = \hat{\sigma}_j^2 = 1$. Par conséquent, σ_{ij} est un coefficient de corrélation qui est appelé coefficient de corrélation polychorique. La fonction de vraisemblance logarithmique $l_0(\sigma_{ij})$ doit être modifiée si on utilise des variables d'un autre genre. Il convient de noter que σ_{ij} désigne les covariances des termes d'erreur des équations de $y_t^*, t = 1,...,P$, étant donné x_i. Contrairement à la procédure dans LISREL 7, on ne suppose pas que les variables y_{it}^* et y_{ij}^* sont distribuées conjointement selon une loi normale. On suppose seulement que les erreurs sont distribuées normalement.

On rassemble ensuite les seuils estimés $\hat{\tau}_i$, les coefficients de forme réduite $\hat{\gamma}_i$ et $\hat{\Pi}_{i\cdot}$, les variances $\hat{\sigma}_i^2$ et les covariances $\hat{\sigma}_{ij}$ de toutes les équations dans un vecteur $\hat{\kappa}_n$ qui dépend de la taille d'échantillon n. En vue de la dernière étape de l'estimation, on calcule une estimation fortement convergente de la matrice des covariances asymptotiques, W, de $\hat{\kappa}_n$. Cette estimation est désignée par $\hat{W}_n$. Il est difficile de calculer la matrice des covariances asymptotiques parce que les estimations $\hat{\sigma}_{ij}$ calculées à la deuxième étape dépendent des coefficients estimés $\hat{\tau}_f,\hat{\gamma}_f,\hat{\Pi}_{f\cdot},\hat{\sigma}_f^2, f = i,j$ de la première étape. Les éléments de la matrice des covariances asymptotiques W sont présentés dans Küsters (1987). On calcule l'estimation $\hat{W}_n$ par le programme MECOSA en se servant des dérivées premières analytiques et des dérivées secondes numériques de la fonction de vraisemblance logarithmique des première et deuxième étapes.

3. À la troisième étape, on exprime le vecteur de seuils, κ, les coefficients de régression de forme réduite et la matrice de covariance de forme réduite comme une fonction des paramètres structurels qui nous intéressent, ceux-ci étant regroupés dans le vecteur de paramètres ϑ. On estime ensuite ce vecteur en minimisant la forme quadratique

$$Q_n(\vartheta) = (\hat{\kappa}_n - \kappa(\vartheta))'\hat{W}_n^{-1}(\hat{\kappa}_n - \kappa(\vartheta)), \tag{19}$$

qui représente une méthode de la distance minimum reposant sur la normalité asymptotique des estimateurs des coefficients de forme réduite. Le vecteur $\hat{\kappa}_n$ suit une distribution normale asymptotique avec espérance $\kappa(\vartheta)$ et matrice de covariance W. Comme $\hat{W}_n$ est une estimation fortement convergente de W, la forme quadratique $Q_n(\vartheta)$ suit une distribution de χ^2 centrale avec $p - \bar{q}$ degrés de liberté si le modèle est spécifié correctement et que la taille d'échantillon est suffisamment grande. Le nombre p correspond au nombre d'éléments de $\hat{\kappa}_n$ tandis que $\bar{q}$ est le nombre d'éléments de ϑ. Le calcul de $\hat{W}_n$ est une opération très lourde dans le cas des modèles qui comptent de nombreux paramètres. Pour pallier cet inconvénient, on peut se servir de la matrice de pondérations de la méthode des MCG de LISREL, qui peut reposer sur la matrice de covariance estimée de y_i^* et de x_i. On peut calculer cette dernière matrice à l'aide de la matrice estimée $\hat{\Sigma}$ des termes d'erreur, de la matrice $\hat{\Pi}$ des coefficients de régression de forme réduite et de la matrice de covariance estimée des variables explicatives. La fonction $Q_n(\vartheta)$ est minimisée au moyen de l'algorithme de Davidon-Fletcher-Powell avec des dérivées premières numériques.

Le programme MECOSA suit les étapes de l'estimation décrites ci-dessus. À la troisième étape, on exploite pleinement les fonctions de GAUSS pour être en mesure d'estimer des paramètres assujettis à des restrictions arbitraires. On peut définir le vecteur de paramètres $\kappa(\vartheta)$ au moyen du langage matriciel et de la fonction de procédures de GAUSS. Par conséquent, il est possible d'imposer des restrictions arbitraires à $\tau(\vartheta), \gamma(\vartheta), \Pi(\vartheta)$ et $\Sigma(\vartheta)$, ainsi qu'à l'identité $V(\epsilon_i^*) = \Sigma(\vartheta)$ dans l'analyse de données longitudinales.

À la première étape, MECOSA produit des estimations des paramètres de forme réduite $\Pi(\vartheta)$ en opérant une régression de toutes les variables dépendantes dans $y_i^* = (y_{i1}^*,...,y_{iT}^*)'$ par rapport à toutes les variables explicatives incluses dans x_i sans la restriction habituelle selon laquelle l'effet de x_{it} sur $y_{i,t+j}, j \geq 1$, doit être nul. Cette restriction sera introduite dans le programme à la troisième étape. Si le modèle comprend des

variables dépendantes décalées, on peut devoir appliquer des restrictions aux estimations de paramètres dès la première étape. Dans ces conditions, on peut ajuster le programme en conséquence en exécutant la première étape de MECOSA T fois pour chaque variable dépendante y_{it}^*, laquelle fait ensuite l'objet d'une régression par rapport à x_{it}. À la deuxième étape, on doit corriger les données d'entrée de lot habituelles de MECOSA pour tenir compte des T résultats d'estimation obtenus à la première étape. On calcule enfin les covariances ou les coefficients de corrélation polychoriques de la forme réduite compte tenu des restrictions de la première étape.

3.2 Le problème des états initiaux

En ce qui concerne l'estimation du modèle général de Heckman, on a supposé que, pour un modèle avec dépendance à l'égard de l'état ou avec persistance des habitudes, les états initiaux sont connus et non stochastiques ou qu'il est possible d'inclure l'information existante sur les états initiaux dans la liste de variables explicatives. Si ce n'est pas le cas, on doit supposer qu'un nouveau processus débute au premier cycle du panel ou bien qu'il n'existe aucune corrélation entre les termes d'erreur du processus pour les états initiaux et les termes d'erreur des T cycles du panel (voir Heckman 1981a,b). Notons que la première hypothèse vaut pour la déperdition d'effectifs dans un panel où la variable dépendante est définie comme le fait qu'une personne participe au cycle t étant donné qu'elle a participé au cycle 1. Tous participent au cycle 1. Par conséquent, on connaît les états initiaux et un nouveau processus débute au temps 1.

Heckman (1981b) considère ce cas particulier du problème des états initiaux:

$$\epsilon_{it}^* = \beta + \gamma y_{i,t-1} + \alpha_i + \epsilon_{it} \quad \text{où} \tag{20}$$

$$y_{it} = \begin{cases} 1 \text{ if } y_{it}^* > 0 \\ 0 \text{ if } y_{it}^* \le 0 \end{cases} .$$

Les effets aléatoires sont $\alpha_i \sim N(0,\sigma_\alpha^2), E(\alpha_i \epsilon_{it}) = 0, \epsilon_{it} \sim N(0,\sigma_\epsilon^2)$ où $\sigma_\epsilon^2 = 1$.

La seule valeur initiale dans ce modèle est y_{i0}. Qu'arrive-t-il si y_{i0} n'est pas connue et est non stochastique mais qu'elle est déterminée par le même processus que pour y_{it}, $t = 1,...,T$? Dans ces conditions, y_{i0} est dépendante de α_i et la fonction de vraisemblance conditionnelle d'échantillon de y_{it}, $t = 1,...,T$, étant donné y_{i0} est obtenue au moyen d'une intégration par rapport à la variable non observée α, qui peut être exprimée par $\alpha = \sigma_\alpha \eta$ où $\eta \sim N(0,1)$.

$$L(\beta_0,\gamma,\sigma_\alpha^2) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{t=1}^{T} (\Phi(\beta + \gamma y_{i,t-1} + \sigma_\alpha \eta)^{y_{it}} (1 - \Phi(\beta + \gamma y_{i,t-1} + \sigma_\alpha \eta))^{1-y_{it}} \times P(y_{i0} | \sigma_\alpha \eta) \varphi(\eta) d\eta \ , \tag{21}$$

où $\varphi(.)$ est la fonction de densité normale. Le terme $P(y_{i0} \mid \alpha)$ est la probabilité que la variable aléatoire y_{i0}^* soit > 0 ou bien ≤ 0. Heckman (1981b) étudie les diverses façons de spécifier $P(y_{i0} \mid \alpha)$ en posant des restrictions pour le processus qui génère les variables exogènes x_{it} et examine l'application de la méthode d'estimation de Kiefer et Wolfowitz (1956). De toutes manières, cette méthode d'estimation s'avère trop peu commode.

La première solution simple que propose Heckman devant ce problème d'estimation est d'estimer α_i comme paramètre par des méthodes du maximum de vraisemblance. Or, ces méthodes ne produisent que des estimations convergentes des paramètres structurels β et γ lorsque $T \to \infty$ (voir Neyman et Scott 1948 et Andersen 1973), ce que l'on ne peut supposer pour des données de panel. Les quelques résultats de la simulation de Monte Carlo que présente Heckman (1981b) pour $T = 8$ cycles de panel dans le modèle de l'équation (21) ne sont sûrement pas suffisants pour pouvoir considérer l'utilisation de la méthode du maximum de vraisemblance pour estimer les α_i de même que β et γ comme une solution générale.

La deuxième solution simple que propose Heckman (1981b) est de remplacer $P(y_{i0} \mid \alpha)$ par $F(x_{i0}\delta)$, où x_{i0} est un vecteur de variables explicatives pour y_{i0}^* et $F(.)$ est la fonction de répartition de la variable aléatoire

$$y_{i0}^* = x_{i0}\delta + \epsilon_{i0}^* . \tag{22}$$

Cette solution ponctuelle se veut simplement une tentative pour remplacer l'hétérogénéité non observée en y_{i0}^* par l'hétérogénéité observée dans x_{i0}. Rien n'empêche le terme d'erreur d'être corrélé avec α_i et ϵ_{it}, $t = 1,...,T$, sans restriction et ce, afin de refléter la corrélation sériale entre ϵ_{i0}^* et $\epsilon_{it}^* = \alpha_i + \epsilon_{it}$ induite par α_i. En pratique, cela revient à inclure y_{i0}^* dans le vecteur des variables dépendantes et à ajouter des paramètres dans la matrice de covariance complète de $\tilde{\epsilon}_i^* = (\epsilon_{i0}^*, \epsilon_{i1}^*, ..., \epsilon_{iT}^*)'$. Si $\epsilon_{it}^* = \alpha_i + \epsilon_{it}$ avec $V(\alpha_i) = \sigma_\alpha^2$ et $V(\epsilon_{it}) = \sigma_\epsilon^2$, la matrice de covariance complète s'écrit

$$V(\tilde{\epsilon}_i^*) = \begin{bmatrix} \sigma_{00}^2 & & & & \\ \sigma_{10} & \sigma_\alpha^2 + \sigma_\epsilon^2 & & & \\ \sigma_{20} & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\epsilon^2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \sigma_{T0} & \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_\epsilon^2 \end{bmatrix} . \tag{23}$$

La simulation de Monte Carlo de Heckman (1981b) montre que cette méthode produit plus efficacement que la méthode du maximum de vraisemblance des estimations convergentes de β et de γ. Si $F(.) = \Phi(.)$, on peut introduire directement la méthode ci-dessus dans le programme MECOSA en spécifiant les équations suivantes:

$$y_{i0}^* = x_{i0}\delta + \epsilon_{i0}^* . \tag{24}$$

$$y_{it}^* = x_{it}\beta + \epsilon_{it}^* . \tag{25}$$

La matrice de covariance conjointe de ϵ_{i0}^* et de ϵ_i^* peut être construite comme la matrice de l'équation (23).

La solution proposée ci-dessus pour résoudre le problème des états initiaux tient fondamentalement à la spécification d'un modèle pour les états initiaux. L'essentiel est de spécifier un modèle pour états initiaux qui permette d'établir les meilleures prévisions possibles. Ce modèle ne doit pas nécessairement être le même que pour y_{it}, $t = 1,..., T$. Rien n'empêche les termes d'erreur ϵ_{i0}^*, $\epsilon_{i,-1}^*$, $\epsilon_{i,-k}^*$ d'être corrélés avec ϵ_{it}^*. Si le modèle pour les états initiaux est mal spécifié, la solution simple ne produira pas de bons résultats. Si le modèle est correct, les estimateurs du maximum de vraisemblance de B, Γ et $V(\epsilon_i^*)$ seront convergents pour une valeur T fixe lorsque $n \to \infty$.

4. ANALYSE DE LA PRODUCTION À L'AIDE DE DONNÉES D'ESSAI SUR LES ENTREPRISES ALLEMANDES

Afin d'illustrer le modèle défini dans la section 2 et les méthodes d'estimation décrites dans la section 3, nous allons analyser des données d'essai commerciales de l'Institut IFO de Munich. Il s'agit plus précisément des données recueillies dans quatre cycles d'enquête (août et novembre 1987, février et mai 1988) auprès de 656 entreprises allemandes. Les variables et les codes utilisés dans l'analyse figurent dans le tableau 1.

Tableau 1: Questions et variables tirées de l'enquête de l'IFO menée auprès des entreprises.

Question	type de variable	définition mathém. de la variable	nom de la variable
Par rapport au mois précédent, la production de XY pour le marché intérieur a été plus élevée(3), la même(2), plus faible(1)?	ordinale	$\Delta y_t = y_t - y_{t-1}$	production O
À l'heure actuelle, le stock de produits finis de XY est égal à zéro, équivaut à moins de 0.5/4, 1/4,..., 6/4 de mois de production, équivaut à plus de 6/4 de mois de production?	métrique	$\ell n \frac{f_t}{f_{t-1}}$	stock de produits finis LFP
À l'heure actuelle, le stock de matières premières de XY est égal à zéro, équivaut à moins de 0.5/4, 1/4,..., 6/4 de mois de production, équivaut à plus de 6/4 de mois de production?	métrique	$\ell n \frac{r_t}{r_{t-1}}$	stock de matières premières LRP
À l'heure actuelle, le carnet de commandes de XY est assez bien rempli (par ex.: prolongement de la période de livraison)(3), est stable(2), laisse à désirer(1)?	ordinale	$a_t - a_t^*$	carnet de commandes AB
Si l'on fait abstraction des variations saisonnières, l'activité commerciale de XY dans les 6 prochains mois sera plutôt bonne(3), se maintiendra(2), sera plutôt mauvaise(1)?	ordinale	$d_{t,t-1} - d_t$	prévisions commerciales GL
Par rapport au mois précédent, la demande de produits de XY (au pays et à l'étranger) s'est accrue(3), s'est maintenue(2), a diminué(1)?	ordinale	$\Delta d_t = d_t - d_{t-1}$	demande au temps t D
Effectifs de l'entreprise	métrique	$\ell n\ k$	LNE
Activité de production au temps t-1 Valeur 1 si $\Delta y_{t-1} = 1$ et valeur 0 dans le cas contraire	fictive	u_t	production OL1
Activité de production au temps t-1 Valeur 1 si $\Delta y_{t-1} = 3$ et valeur 0 dans le cas contraire	fictive	v_t	production OL2

Comme l'indice inférieur dont sont affectées les variables représente le numéro du cycle de panel, les symboles O_0, LFP_0, LRP_0,... désignent les variables correspondantes au temps 0; O_1, LFP_1, LRP_1,... désignent les variables au temps $t = 1$. Les variables O (production), AB (carnet de commandes), GL (prévisions commerciales) et D (demande) sont évaluées seulement comme des variables nominales ordonnées, les variables LFP, LRP et LNE sont des variables métriques tandis que OL1 et OL2 sont des variables fictives.

Le modèle que nous spécifions ci-dessous est expliqué dans Arminger et Ronning (1991). Le processus que décrit ce modèle dépend du cycle 0; celui-ci est donc exclu de la modélisation.

$$\Delta y_t = \mu_t + \beta_{t1}(a_{t-1} - a^*_{t-1}) + \beta_{t2}(d_{t,t+1} - d_t) + \beta_{t3}s_t + \gamma_{t1}(\ell n\, k) + \gamma_{t2}\left[\ell n \frac{r_t}{r_{t-1}}\right]$$

$$+\gamma_{t3}\left[\ell n \frac{f_t}{f_{t-1}}\right] + \gamma_{t4}u_t + \gamma_{t5}v_t + \alpha + \epsilon_t, t = 1,2,3. \tag{26}$$

Le modèle ci-dessus explique la variation de la production au temps t en fonction du carnet de commandes au temps t - 1, des prévisions commerciales au temps t, du nombre d'employés, de la variation relative du stock de matières premières entre t - 1 et t, de la variation relative du stock de produits finis entre t - 1 et t, et des états de la variable production au temps t - 1. La variable ln k, c.-à-d. le nombre d'employés, ne varie pas dans le temps.

La variable s_t est définie par l'expression

$$s_t = [(d_t - d_{t-1}) - (d_{t-1,t} - d_{t-1})], \tag{27}$$

on peut la voir comme la variable de bouleversement ou d'effet de surprise. Si $(d_t - d_{t-1}) > (d_{t-1,t} - d_{t-1})$, l'effet est positif, c'est-à-dire que la demande est supérieure aux prévisions de la période antérieure, sinon l'effet est nul ou négatif. On peut estimer l'effet de s_t en posant β_{t3} comme le paramètre de la demande $(d_t - d_{t-1})$ et $-\beta_{t3}$ comme le paramètre des prévisions commerciales de la période antérieure, $(d_{t-1,t} - d_{t-1})$.

Notons que les variables Δy_t, $(a_{t-1} - a^*_{t-1})$, $(d_{t,t-1} - d_t)$ et $(d_t - d_{t-1})$ sont observées uniquement à une échelle ordinale, la règle d'observation qui est supposée s'appliquer étant la suivante:

$$O_t = \begin{cases} 1 \text{ si } \Delta y_t \leq \tau_1^{(1)} \\ 2 \text{ si } \tau_1 < \Delta y_t \leq \tau_2^{(1)} \\ 3 \text{ si } \Delta y_t > \tau_2^{(1)} \end{cases}. \tag{28}$$

Les règles d'observation pour AB_t, GL_t et D_t sont analogues. Pour des raisons d'identification, le premier seuil est posé égal à 0.

La variable aléatoire $\alpha \sim N(0,\sigma_\alpha^2)$ reproduit l'hétérogénéité non observée; $\epsilon_t \sim N(0,\sigma_t^2)$ est supposée non liée. Puisque AB_t, GL_t et D_t sont observées uniquement à une échelle ordinale, nous supposons que $(a_{t-1} - a^*_{t-1})$, $(d_{t,t-1} - d_t)$ et $(d_t - d_{t-1})$ sont des variables endogènes dont les moyennes dépendent des variables exogènes $\ell n\, k$, $\left(\ell n \frac{r_t}{r_{t-1}}\right)$, $\left(\ell n \frac{f_t}{f_{t-1}}\right)$, u_t et v_t et d'un vecteur multidimensionnel des erreurs normales qui n'est pas corrélé avec α ni avec ϵ_t. Le modèle pour toutes les variables endogènes dans le premier cycle s'écrit donc

$$\begin{pmatrix} a_0 - a_0^* \\ d_{2,1} - d_1 \\ d_1 - d_0 \\ d_{1,0} - d_0 \\ y_1 - y_0 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \beta_{51} & \beta_{52} & \beta_{53} & -\beta_{53} & 0 \end{pmatrix} \begin{pmatrix} a_0 - a_0^* \\ d_{2,1} - d_1 \\ d_1 - d_0 \\ d_{1,0} - d_0 \\ y_1 - y_0 \end{pmatrix}$$

$$+ \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \gamma_{15} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & \gamma_{25} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \gamma_{34} & \gamma_{35} \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} & \gamma_{45} \\ \gamma_{51} & \gamma_{52} & \gamma_{53} & \gamma_{54} & \gamma_{55} \end{pmatrix} \begin{pmatrix} \ell nk \\ \ell n\frac{r_1}{r_0} \\ \ell n\frac{f_1}{f_0} \\ u_1 \\ v_1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \alpha \end{pmatrix} + \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \\ \epsilon_5 \end{pmatrix} . \quad (29)$$

Les modèles pour les cycles 2 et 3 se construisent de la même façon. Notons que la variable GL est présente deux fois dans le premier cycle: GL_0 et GL_1. Dans le deuxième cycle, on doit tirer la valeur GL_1 du premier cycle. Le modèle en entier consiste donc en 13 équations. Toutefois, nous nous intéressons essentiellement aux équations 5, 9 et 13, c'est-à-dire à $y_1 - y_0$, $y_2 - y_1$ et $y_3 - y_2$. Le premier modèle ne tient compte d'aucune contrainte imposée par la proportionnalité des coefficients pour chaque cycle et l'hétérogénéité non observée. Les estimations de paramètres calculées à l'aide de MECOSA sont reproduites dans le tableau 2.

Tableau 2: Estimations de paramètres sans contrainte pour le modèle de production de l'IFO (valeurs t entre parenthèses).

variables explicatives	cycle 1	cycle 2	cycle 3
τ_1	0	0	0
τ_2	2.175 (28.989)	2.175 (28.989)	2.175 (28.989)
μ	0.604 (1.613)	1.147 (3.281)	0.599 (2.793)
$(a_{t-1} - a_{t-1}^*)$	0.439 (5.685)	0.305 (4.161)	0.293 (6.236)
$d_{t-1,t} - d_t$	0.053 (1.046)	0.089 (2.319)	0.125 (2.822)
s_t	0.413 (10.993)	0.312 (9.844)	0.385 (13.532)
$\ell n\, k$	-0.049 (-1.433)	-0.023 (-0.707)	0.015 (0.468)
$\ell n\frac{r_t}{r_{t-1}}$	0.097 (0.679)	-0.077 (-0.481)	0.086 (0.632)
$\ell n\frac{f_t}{f_{t-1}}$	0.238 (2.973)	-0.031 (-0.369)	-0.058 (-0.796)
u_t	-0.543 (-3.040)	-0.834 (-5.916)	-0.227 (-2.386)
v_t	0.129 (1.039)	-0.204 (-1.697)	0.429 (3.217)
R^2_{MZ}	0.089	0.184	0.116
covariances			
cycle 1	0.514		
cycle 2	0.127	0.715	
cycle 3	0.045	-0.010	0.616

Les pseudo-R^2 de McKelvey et Zavoina (1975) montrent qu'une faible partie seulement de la variance de la production est expliquée. La production augmente dans le deuxième cycle comparativement aux premier et

troisième cycles. D'après les valeurs *t*, les variables «carnet de commandes» (AB) et «effet de surprise» sont plus importantes que la variable «prévisions commerciales» (GL). Les entreprises réagissent principalement à des bouleversements récents. Si le bouleversement est positif, les entreprises accroissent leur production. Les variables «stock de matières premières» et «stock de produits finis» ont moins d'importance que la dépendance à l'égard de l'état dans la période précédente. Dans ce cas-ci, toutefois, seule la diminution de la production dans la période précédente importe. Les covariances des erreurs sont plutôt faibles, ce qui indique que l'hétérogénéité non observée est probablement négligeable.

Le tableau suivant donne les résultats de l'estimation sous contrainte des mêmes paramètres suivant l'hypothèse de la proportionnalité des coefficients de régression, à l'exception des constantes et de l'effet du nombre d'employés.

Tableau 3: Estimations de paramètres sous contrainte pour le modèle de production de l'IFO (valeurs *t* entre parenthèses).

variables explicatives	cycle 1	cycle 2	cycle 3
τ_1	0	0	0
τ_2	2.135 (30.379)	2.135 (30.379)	2.135 (30.379)
μ	0.842 (2.497)	1.203 (3.025)	0.736 (3.106)
$(a_{t-1} - a^*_{t-1})$	0.372 (8.547)	0.372 (8.547)	0.372 (8.547)
$d_{t-1,t} - d_t$	0.103 (3.539)	0.103 (3.539)	0.103 (3.539)
s_t	0.417 (13.305)	0.417 (13.305)	0.417 (13.305)
$\ell n\, k$	-0.064 (-1.976)	-0.032 (-0.775)	0.008 (0.230)
$\ell n \frac{r_t}{r_{t-1}}$	0.122 (1.472)	0.122 (1.472)	0.122 (1.472)
$\ell n \frac{f_t}{f_{t-1}}$	0.029 (0.634)	0.029 (0.634)	0.029 (0.634)
u_t	-0.548 (-7.380)	-0.548 (-7.380)	-0.548 (-7.380)
v_t	0.070 (0.959)	0.070 (0.959)	0.070 (0.959)
λ		0.749	0.914
σ_a^2	0.055 (1.412)		
valeur de χ^2	3.841	df	14

L'hypothèse de la proportionnalité n'est pas rejetée à un seuil de 0.05. D'après l'inverse du coefficient de proportionnalité λ, le deuxième cycle a une variance d'erreur plus élevée en termes absolus que les premier et troisième cycles. La variance de l'hétérogénéité s'avère non significative à un seuil de 0.05.

BIBLIOGRAPHIE

Andersen, E.B. (1973). *Conditional Inference and Models for Measuring*, Copenhagen: Mentalhygiejnisk Forsknings Institut.

Arminger, G. (1987). Misspecification, asymptotic stability and ordinal measurements in models for the analysis of panel data. *Sociological Methods and Research*, 15, 336-348.

Arminger, G., et Ronning, G. (1991). Ein Strukturmodell für Preis-, Produktions- und Lagerhaltungsentscheidungen von Firmen. *IFO-STUDIEN, Zeitschrift für empirische Wirtschaftsforschung*, 37, 229-254.

Arminger, G. (1992). Analyzing panel data with non-metric dependent variables: Probit models, generalized estimating equations, missing data and absorbing states. *Discussion Paper No. 59*, Deutsches Institut für Wirtschaftsforschung, Berlin.

Hamerle, A., et Ronning, G. (1993). Analysis of discrete panel data, à paraître dans G. Arminger, C.C. Clogg et M.E. Sobel (Éds.). *Handbook of Statistical Modeling for the Behavioral Sciences*, New York: Plenum.

Heckman, J.J. (1981a). Statistical models for discrete panel data, in C.F. Manski and D. McFadden (Eds.). *Structural Analysis of Discrete Data with Econometric Applications*, 114-178.

Heckman, J.J. (1981b). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete stochastic process, in C.F. Manski and D. McFadden (Eds.). *Structural Analysis of Discrete Data with Econometric Applications*, 179-195.

Hsiao, C. (1986). *Analysis of Panel Data*, Cambridge, Massachusetts: Cambridge University Press.

Keane, M.P., et Runkle, D.E. (1992). On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous. *Journal of Business & Economic Statistics*, 10, 1, 1-29.

Kiefer, J., et Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27, 887-906.

Küsters, U. (1987). *Hierarchische Mittelwert- und Kovarianzstrukturmodelle mit nichtmetrischen endogenen Variablen*, Heidelberg: Physica Verlag.

Maddala, G.S. (1987). Limited dependent variable models using panel data. *Journal of Human Resources*, XXII, 3, 307-336.

McKelvey, R.D., et Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103-120.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115-132.

Neyman, J., et Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16, 1-32.

Rosett, R.N., et Nelson, F.D. (1975). Estimation of the two-limit probit regression model. *Econometrica*, 43, 141-146.

Schepers, A., Arminger, G., et Küsters, U. (1991). The analysis of non-metric endogenous variables in latent variable models: the MECOSA Approach, in P. Gruber (Ed.). *Econometric Decision Models: New Methods of Modeling and Applications*, Springer Verlag, Heidelberg, 1991, 459-472.

Schepers, A., et Arminger, G. (1992). *MECOSA: A Program for the Analysis of General Mean- and Covariance Structures with Non-Metric Variables, User Guide*, SLI-AG, Züricher Str. 300, CH-8500 Frauenfeld, Switzerland.

Sobel, M., et Arminger, G. (1992). Modeling household fertility decisions: A nonlinear simultaneous probit model. *Journal of the American Statistical Association*, 87, 38-47.

Stewart, M.B. (1983). On least squares estimation when the dependent variable is grouped. *Review of Economic Studies L*, 737-753.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24-36.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

MODÉLISATION LOGISTIQUE DE DONNÉES D'ENQUÊTE LONGITUDINALES POUVANT COMPORTER UNE ERREUR DE MESURE

C.J. Skinner[1]

RÉSUMÉ

On examine un modèle logistique établissant un rapport entre une réponse binaire fournie à une occasion et la réponse fournie la fois précédente ainsi que d'autres variables auxiliaires. L'erreur de mesure liée à la réponse binaire peut introduire un biais dans les estimations. Des méthodes d'estimation comportant un rajustement qui tient compte de l'erreur de mesure sont proposées pour différents modèles de mesure. Ces méthodes sont illustrées à l'aide de données de l'enquête américaine «Panel Study of Income Dynamics», la réponse examinée étant l'occupation ou non par le répondant d'un emploi syndiqué.

MOTS CLÉS: Flux brut; transition; longitudinal; erreur de mesure.

1. INTRODUCTION

Dans l'analyse des données d'enquête longitudinales, il est souvent intéressant d'estimer un *taux de transition*, c.-à-d. la proportion des unités de la population se trouvant dans un état à une occasion qui, la fois suivante, se trouvent dans un état différent. Par exemple, les analystes du marché du travail peuvent être intéressés à la matrice 3 x 3 des transitions entre les états «personne occupée», «chômeur» et «inactif». À des fins d'analyse, il est souvent intéressant d'étudier comment les taux varient entre différents sous-groupes de la population. Par exemple, dans le cas des différents états d'activité, il peut être intéressant d'étudier le lien entre les taux de transition et le sexe, l'âge et la région.

Les taux de transition sont des proportions et, par conséquent, peuvent être estimés de la façon normale à partir de données d'enquête. Toutefois, les éléments non diagonaux des matrices de transition représentent souvent des nombres peu élevés, de sorte qu'à mesure qu'augmente le nombre de sous-groupes, les tailles d'échantillon sur lesquelles se fondent certaines estimations peuvent devenir faibles et être la source d'erreurs d'échantillonnage élevées. Dans ce cas, une modélisation du rapport entre les taux et les variables auxiliaires définissant les sous-groupes est souhaitable. À la section 2, nous décrivons un modèle logistique pour la représentation du lien entre les probabilités de transition et les variables auxiliaires dans une situation comportant deux états, et l'obtention de réponses à deux occasions. Des modèles semblables ont été utilisés aussi bien dans des applications biostatistiques (par ex. Korn et Whittemore 1979; Muenz et Rubinstein 1985) que dans des applications économétriques (par ex. Hsiao 1986, sect. 7.4; Maddala 1987).

Lorsqu'on estime des flux bruts à partir de données d'enquête, un important problème qui se pose est celui de l'erreur de mesure. Des erreurs aléatoires dans les états mesurés peuvent introduire d'importants biais par excès dans les estimateurs courants des proportions de répondants passant d'un état à l'autre. Un certain nombre d'estimateurs de remplacement utilisant des données de réinterviews ont été proposés pour réduire ce biais (Meyer 1988). Le but de la présente communication est d'étendre ce travail à l'estimation des modèles logistiques mentionnés ci-dessus.

[1] C.J. Skinner, Department of Social Statistics, University of Southampton, SO9 5NH, R.-U.

2. LE MODÈLE

Considérons une population finie de taille N, qui demeure fixe entre les deux occasions $t = 1, 2$ et qui est décomposée en I cellules de tailles $N_1, ..., N_I$ ($\sum N_i = N$) selon les niveaux d'un ou de plusieurs facteurs définis à $t = 1$. Soit y_t une variable indicatrice binaire représentant les deux états et soit N_{ijk} le nombre d'unités de la cellule i pour lesquelles $y_1 = j$ et $y_2 = k$ ($i = 1, ..., I; j = 0, 1; k = 0, 1$). Soit $N_{ij.} = N_{ij0} + N_{ij1}$; on a alors

$$\sum_{j=0}^{1} \sum_{k=0}^{1} N_{ijk} = \sum_{j=0}^{1} N_{ij.} = N_i, \; i = 1, ..., I.$$

Supposons que les valeurs de la population finie soient engendrées par un modèle de telle façon que les N_{ijk} suivent une distribution multinomiale avec paramètres N et ϕ_{ijk} ($i = 1, ..., I; j = 0, 1; k = 0, 1$). En particulier, on a $E(N_{ijk}) = N\phi_{ijk}$. Posons

$$\pi_{ij} = \phi_{ij1} / \phi_{ij.},$$

où

$$\phi_{ij.} = \phi_{ij0} + \phi_{ij1}, \; i = 1, ..., I; \; j = 0, 1.$$

Alors, π_{ij} représente la probabilité de transition du modèle ou le flux brut dans la cellule i entre l'état j à $t = 1$ et l'état 1 à $t = 2$. Nous nous donnons comme objectif d'étudier la dépendance de π_{ij} à l'égard de i et j, et nous considérons le modèle logistique suivant pour π_{ij}:

$$\pi_{ij} = F(x_{ij}\beta), \; i = 1, ..., I; \; j = 0, 1, \tag{1}$$

où

$$F(t) = e^t / (1 + e^t),$$

les x_{ij} sont des vecteurs $1 \times s$ de constantes connues et β est un vecteur $s \times 1$ de paramètres inconnus. Notons que (1) peut aussi être exprimée sous la forme suivante

$$\log[\pi_{ij} / (1 - \pi_{ij})] = x_{ij}\beta. \tag{2}$$

À des fins d'illustration, certains cas spéciaux de ce modèle sont présentés ci-dessous. Cependant, des éléments de notation doivent d'abord être introduits. Pour une série de vecteurs $1 \times k$, a_{ij} ($i = 1, ..., I; j = 0, 1$), soit $[a_{ij}]$ la matrice $2I \times k$ ayant comme lignes $a_{10}, a_{11}, a_{20}, a_{21}, ..., a_{I0}, a_{I1}$. Soit $X = [x_{ij}]$, $\ell = [\ell_{ij}]$, $\pi = [\pi_{ij}]$, $\phi = [\phi_{ij}]$, $f(\beta) = [f_{ij}(\beta)]$, où $\ell_{ij} = \log[\pi_{ij} / (1 - \pi_{ij})]$, $f_{ij}(\beta) = F(x_{ij}\beta)$.

Alors, (1) peut être reformulée ainsi

$$\pi = f(\beta) \tag{3}$$

et (2) peut être reformulée ainsi

$$\ell = X\beta. \tag{4}$$

Exemples de modèles

(i) **Taux de transition constants**

Soit $s = 2$, $x_{ij} = (1 j)$ et $\beta = (\beta_1 \beta_2)'$. Alors $\pi_{i0} = F(\beta_1)$ et $\pi_{i1} = F(\beta_1 + \beta_2)$ pour tous les i.

(ii) **Modèle additif**

Soit $s = r + 2$, $x_{ij} = (1 z_i j)$ et $\beta = (\beta_1 \beta_2' \beta_3)'$, où z_i est un vecteur $1 \times r$ de constantes connues obtenu d'après les niveaux des facteurs définissant les I cellules et β_2 est un vecteur $r \times 1$ de paramètres inconnus. Par exemple, les cellules peuvent être le résultat du croisement de I/2 groupes d'âge par 2 sexes et z_i peut être $(a_i a_i^2 s_i)$ où a_i est le point milieu du groupe d'âge et s_i est une variable fictive représentant le sexe pour

la cellule i. La valeur du rapport $P(y_2 = 1/y_1 = 1)/(P(y_2 = 1/y_1 = 0)$ est $\exp(\beta_3)$, qui est constant d'une cellule à l'autre.

(iii) **Modèles distincts pour des états précédents différents**

Soit $s = 2r + 2$, $x_{ij} = (1\ z_i\ j\ jz_i)$ et $\beta = (\beta_1\ \beta_2'\ \beta_3\ \beta_4')'$, où z_i est défini comme en (ii). Contrairement à l'exemple (ii), ce modèle permet une interaction entre y_1 et la cellule i. Les taux de transition sont maintenant $\pi_{i0} = F(\beta_1 + z_i\beta_2)$ et $\pi_{i1} = F[(\beta_1 + \beta_3) + z_i(\beta_2 + \beta_4)]$.

(iv) **Modèle saturé**

Soit $s = 2I$ et supposons que X soit non singulière. Il existe alors une relation biunivoque entre π et β puisque (4) peut être inversé, ce qui donne $\beta = X^{-1}\ell$.

En général, nous prenons β comme vecteur des paramètres à l'étude. Comme il est indiqué ci-dessus, β est bien défini uniquement si le modèle (1) est valable. En pratique, toutefois, il peut demeurer intéressant d'ajuster un modèle, comme le modèle des effets principaux en (ii), même si ce modèle n'est valable qu'approximativement. En vertu uniquement de l'hypothèse d'une distribution multinomiale et d'une spécification de x_{ij}, mais sans nécessairement que (1) soit valable, nous définissons β comme étant la solution de

$$\sum_i \sum_j x_{ij}\,\phi_{i.}\,(f_{ij}(\beta) - \pi_{ij}) = 0. \tag{5}$$

3. ESTIMATION

Soit $\hat{N}_{ijk}$ un estimateur de N_{ijk} qui peut exiger une pondération ou d'autres rajustements propres à l'enquête. Considérons un cadre asymptotique dans lequel N et la taille de l'échantillon n augmentent, mais où I et les ϕ_{ijk} sont fixes. Soient

$$\hat{N}_{ij.} = \sum_k \hat{N}_{ijk},\ w_{ij.} = \hat{N}_{ij}/\hat{N},\ p_{ij} = \hat{N}_{ij1}/N_{ij},\ \hat{N} = \sum_i \sum_j \hat{N}_{ij.}$$

$$w = [w_{ij.}],\ p = [p_{ij}].$$

Nous supposons alors que $(p'\,w')'$ est convergent pour $(\pi'\,\phi')'$ et que la distribution asymptotique quand $n \to \infty$ de $\sqrt{n}\,[(p'\,w')' - (\pi'\,\phi')']$ est normale, avec vecteur de moyennes nul et matrice de covariances

$$V_{(p,w)} = \begin{bmatrix} V_p & C_{pw} \\ C_{pw}' & V_w \end{bmatrix}. \tag{6}$$

Si w et p sont donnés, β peut être estimé par la solution $\hat{\beta}$ des équations (5), $\phi_{ij.}$ et π_{ij} étant remplacés par $w_{ij.}$ et p_{ij} respectivement.

Nous avons maintenant besoin d'une notation additionnelle. Pour une série de scalaires $a_{ij}\ (i = 1,...,I;\ j = 0,1)$, supposons que $\mathrm{diag}[a_{ij}]$ dénote la matrice diagonale $2I \times 2I$ ayant les éléments diagonaux $a_{10}, a_{11}, ..., a_{I0}, a_{I1}$. Soit

$$D(w) = \mathrm{diag}[w_{ij.}],\ D(\phi) = \mathrm{diag}\,[\phi_{ij.}],$$

$$D(\epsilon) = \mathrm{diag}[\epsilon_{ij}],\ \Delta = \mathrm{diag}[\phi_{ij.} f_{ij}(\beta)\,\{1 - f_{ij}(\beta)\}],$$

où $\epsilon_{ij} = f_{ij}(\beta) - \pi_{ij}$ est l'erreur d'approximation du modèle pour la cellule i et $y_1 = j$. Dans ce cas, $\hat{\beta}$ résoud les équations d'estimation

$$X'D(w)f(\hat{\beta}) = X'D(w)p\,, \tag{7}$$

(voir Roberts et coll. 1987, équation 2.3). La matrice de covariances asymptotique de $\hat{\beta}$ est donnée par

$$V(\hat{\beta}) = n^{-1}(X'\Delta X)^{-1}\, X'\sum X(X'\Delta X)^{-1}\,, \tag{8}$$

où

$$\sum = [D(\phi)\ D(\epsilon)]\ V_{(p,w)}\ [D(\phi)\ D(\epsilon)]'\,. \tag{9}$$

Si le modèle logistique est valable, on a $D(\epsilon) = 0$ $\sum$ se ramène alors à $D(\phi)V_p D(\phi)$ et $V(\hat{\beta})$ est réduite à une expression analogue à l'équation (2.4) de Roberts et coll. (1984).

Si l'on a des estimateurs de V_p, V_w et C_{pw}, $V(\hat{\beta})$ peut être estimée en substituant w, p, et $\hat{\beta}$ à ϕ, π, et β respectivement dans Δ et $\sum$.

Exemple: échantillonnage aléatoire simple

Supposons que n unités d'une population soient prélevées par échantillonnage aléatoire simple, et que $n_{ij.}$ et n_{ijl}, les quantités analogues à $N_{ij.}$ et N_{ijl}, soient observées $(i = 1, ..., I;\ j = 0, 1)$. Alors, les équations du maximum de vraisemblance sont, par analogie avec (5):

$$\sum_i \sum_j x_{ij}\,[n_{ij.} f_{ij}(\beta) - n_{ijl}] = 0\,,$$

ce qui équivaut à prendre $w_{ij.} = n_{ij.}/n$, $p_{ij} = n_{ijl}/n_{ij.}$ dans (7). Alors, p et w sont asymptotiquement non corrélés $(C_{wp} = 0)$,

$$V_p = \mathrm{diag}[\pi_{ij}(1 - \pi_{ij})\phi_{ij.}^{-1}]\,,\quad V_w = D(\phi) - \phi\phi'$$

et, d'après (9),

$$\sum = \mathrm{diag}[\pi_{ij}(1 - \pi_{ij})\phi_{ij.}] + D(\epsilon)\,[D(\phi) - \phi\phi']\,D(\epsilon)\,.$$

Si le modèle logistique est valable, alors $\sum = \Delta$ et $V(\hat{\beta}) = n^{-1}(X'\Delta X)^{-1}$.

4. EFFET DE L'ERREUR DE MESURE

Supposons maintenant que nous n'observions pas $\hat{N}_{ijk}$, mais seulement $\hat{N}^*_{ijk}$, un estimateur de N^*_{ijk}, le nombre d'unités de la cellule i pour lesquelles $y_1^* = j$ et $y_2^* = k$, où y_1^* et y_2^* sont des versions de y_1 et y_2, respectivement, mesurées avec une erreur. Nous supposons que les N^*_{ijk} suivent une distribution multinomiale avec paramètres N et ϕ^*_{ijk}, et nous définissons $\hat{N}^*_{ij}, \phi^*_{ij}, w^*_{ij}, p^*_{ij}, \pi^*_{ij}, \phi^*, w^*, p^*, \pi^*, D(w^*)$ et $D(\phi^*)$, de la même façon que leurs versions sans astérisques.

Soit $\hat{\beta}^*$ la solution de l'estimation des équations (7) après remplacement de w et p par w^* et p^* respectivement. Alors, en supposant que w^* et p^* soient convergents pour ϕ^* et π^* respectivement, $\hat{\beta}^*$ sera convergent pour la solution β^* des équations:

$$X'D(\phi^*)f(\beta^*) = X'D(\phi^*)\pi^*\,. \tag{10}$$

En général, β^* ne sera pas égal à β à moins que $\phi^* = \phi$ et $\pi^* = \pi$. Par conséquent, l'erreur de mesure introduit un biais même pour de grands échantillons.

5. RAJUSTEMENT TENANT COMPTE DE L'ERREUR DE MESURE

La nature du rajustement tenant compte de l'erreur de mesure dépendra de la formulation du modèle de l'erreur de mesure, laquelle dépendra à son tour, en pratique, de la nature et de l'étendue des données de validation disponibles. Nous supposons d'abord qu'il n'y a pas d'erreur de classification des cellules, de telle sorte que $\phi^*_{i..} = \phi_{i..}$, où $\phi_{i..} = \phi_{i0.} + \phi_{i1.}$, $\phi^*_{i..} = \phi^*_{i0.} + \phi^*_{i1.}$.

Nous supposons ensuite que seules des données de validation transversales sont disponibles, auquel cas il est naturel, conformément à l'approche de Abowd et Zellner (1985), de supposer des *erreurs de mesure indépendantes à l'intérieur des cellules*:

$$\phi^*_{ijk} = \sum_{\ell=0}^{1} \sum_{m=0}^{1} \theta^1_{ij\ell}\ \theta^2_{ikm}\ \phi_{i\ell m}\,, \tag{11}$$

où $\theta^t_{ijk} = pr(y_t^* = j \mid y_t = k,$ cellule $i)$ est la probabilité de classer de façon erronée l'état k comme étant l'état j dans la cellule i au temps t. Sans données de validation longitudinales, il est difficile de savoir comment formuler un modèle visant des erreurs dépendantes, bien qu'une analyse de sensibilité visant des cas où il n'y a pas indépendance soit possible (Rao et Singh 1991).

Si $\phi^t(i)$, $\phi(i)$ et $\phi^*(i)$ dénotent les matrices 2×2 ayant comme $jk^{ième}$ éléments ϕ^t_{ijk}, ϕ_{ijk} et ϕ^*_{ijk} respectivement, (11) peut être reformulée ainsi

$$\phi^*(i) = \theta^1(i)\ \phi(i)\ \theta^2(i)'\,.$$

Si les estimateurs $\hat{\theta}^t(i)$ des $\theta^t(i)$ sont disponibles par suite d'études de validation, un estimateur rajusté de $\phi(i)$ est donné par

$$\hat{\phi}(i) = \hat{\theta}^1(i)^{-1}\ \hat{\phi}^*(i)\ [\hat{\theta}^2(i)']^{-1}\,, \tag{12}$$

où le $jk^{ième}$ élément de $\hat{\phi}^*(i)$ est $\hat{N}^*_{ijk} / \hat{N}^*$. Un estimateur rajusté de β est alors obtenu par la résolution de (5), en remplaçant $\phi_{ij.}$ et π_{ij} par $\hat{\phi}_{ij}$ et $\hat{\pi}_{ij} = \hat{\phi}_{ij1} / \hat{\phi}_{ij.}$ respectivement. Si l'on suppose la convergence de $\hat{\phi}^*_{ijk}$ et $\hat{\theta}^t_{ijk}$ pour ϕ^*_{ijk} et θ^t_{ijk} respectivement, l'estimateur rajusté sera convergent et sa matrice de covariances asymptotique sera semblable à celle définie en (8) et (9), où $V_{(p,w)}$ est remplacée par la matrice de covariances asymptotique du vecteur de $\hat{\pi}_{ij}$ et $\hat{\phi}_{ij.}$. Cette matrice peut être estimée par la méthode δ, pourvu que des estimations de la matrice de covariances des $\hat{\theta}^t_{ijk}$ soient disponibles.

Un problème que pose cette méthode est que les valeurs de $\hat{\phi}_{ijk}$ découlant de (12) peuvent se situer hors de l'intervalle [0,1], ce qui peut se produire souvent puisque les $\hat{\phi}^*(i)$ sont susceptibles d'afficher un degré appréciable de variabilité d'échantillonnage; c'est précisément pour cela, en effet, qu'un modèle logistique a été choisi initialement. On est ainsi amené soit à imposer une procédure d'inférence avec contraintes, soit à envisager une formulation plus étroite du modèle de l'erreur de mesure. Une telle hypothèse plus restrictive, conformément à ce que proposaient Chua et Fuller (1987), consiste à supposer des *erreurs de mesure non biaisées*:

$$\phi^*_{ij.} = \phi_{ij.}\,,\ \phi^*_{i,k} = \phi_{i,k} \quad i=1,\ ..,\ I,\ j=0,1,\ k=0,1.$$

Soit

$$\alpha^t_i = pr(y_t^* = 1 \mid y_t = 0,\ \text{cell } i) / pr(y_t^* = 1 \mid \text{cell } i) \tag{13}$$

une mesure de l'«ampleur» de l'erreur de mesure au temps t dans la cellule i. Une conséquence de la condition d'absence de biais est que le côté droit de (13) demeure inchangé si 1 est remplacé par 0 et vice-versa. Si l'on pose $\pi_{ij}^{*} = \phi_{ij1}^{*} / \phi_{ij.}^{*}$, il découle de l'hypothèse selon laquelle les erreurs de mesure sont à la fois indépendantes et non biaisées que

$$\pi_{ij}^{*} = (1-\alpha_i^1)\,(1-\alpha_i^2)\,\pi_{ij} + [1-(1-\alpha_i^1)\,(1-\alpha_i^2)]\,\phi_{i.1}/\phi_{i..}\,.$$

Si l'on a un estimateur $\tilde{\gamma}_i$ de $\gamma_i = [(1-\alpha_i^1)\,(1-\alpha_i^2)]^{-1}$, on peut estimer π_{ij} par

$$\tilde{\pi}_{ij} = \tilde{\gamma}_i p_{ij}^{*} - (\tilde{\gamma}_i - 1) p_i^{*}\,, \tag{14}$$

où $p_{ij}^{*} = \hat{\phi}_{ij1}/\hat{\phi}_{ij.}^{*}$ et $p_i^{*} = (w_{i0}^{*} p_{i0}^{*} + w_{i1}^{*} p_{i1}^{*})/(w_{i0}^{*} + w_{i1}^{*})$.

On obtient alors un estimateur rajusté de β en résolvant (5) après avoir remplacé ϕ_{ij1} et π_{ij} par $\phi_{ij.}^{*}$ et $\tilde{\pi}_{ij}$ respectivement, c'est-à-dire

$$X'D(w^{*})f(\tilde{\beta}) = X'D(w^{*})\tilde{\pi}\,,$$

où $\tilde{\pi} = [\tilde{\pi}_{ij}]$. Nous définissons $\gamma = (\gamma_1, ..., \gamma_I)'$, $\tilde{\gamma} = (\tilde{\gamma}_1, ..., \tilde{\gamma}_I)'$, et nous supposons que $n^{1/2}(\tilde{\gamma}-\gamma) \to N[0, V_\gamma]$ quand $n \to \infty$ et que $\tilde{\gamma}$ est asymptotiquement indépendant de $(p^{*\prime}, w^{*\prime})$. Puisque w^{*} et $\tilde{\pi}$ sont convergents pour $\phi^{*} = \phi$ et π respectivement, $\tilde{\beta}$ est convergent pour β. La matrice de covariances asymptotique de $\tilde{\beta}$ est donnée par l'expression (8), avec $V_{(p,w)}$ en (9) remplacée par la matrice de covariances asymptotique (normalisée) de $(\tilde{\pi}', w^{*\prime})'$. Si l'on a des estimateurs convergents de $V(p^{*})$, $V(w^{*})$ et $V(\tilde{\gamma})$, un estimateur convergent de $V(\tilde{\beta})$ peut être obtenu comme auparavant.

6. UN EXEMPLE

Le présent exemple se fonde sur les données de l'enquête américaine «Panel Study of Income Dynamics» (PSID). Le tableau 1 présente une classification croisée de la variable:

y_t^{*} = 1 si la personne occupe un emploi syndiqué
= 0 sinon,

pour les deux années $t=1$ (1983) et $t=2$ (1987), dans le cas des hommes faisant partie de l'échantillon autopondéré du «Survey Research Centre» (Hill 1992, p. 9) qui, pour les deux années, travaillaient au moment de l'enquête mais n'étaient pas des travailleurs autonomes ni des employés du gouvernement.

Tableau 1: Nombres d'unités de l'échantillon pour des variables observées.

		y_2^{*} = 0	y_2^{*} = 1
y_1^{*}	0	684	33
	1	43	191

Deux facteurs susceptibles d'influer sur les transitions entre états sont examinés. Le premier, l'âge, est divisé en quatre catégories, 18-29, 30-34, 35-44, 45+, à peu près de tailles égales pour l'échantillon étudié. Le deuxième facteur répartit les secteurs d'emploi en deux catégories, grosso modo selon leur tendance à inclure des emplois syndiqués ou non. La première catégorie, moins susceptible d'inclure des emplois syndiqués, comprend les emplois professionnels, ceux des secteurs de la gestion et des ventes, ainsi que les emplois dans les exploitations agricoles. La deuxième catégorie, plus susceptible d'inclure des emplois syndiqués, comprend les emplois manuels et de bureau. Ces deux facteurs définissent ensemble I = 8 cellules.

À des fins de simplicité, nous ne tenons pas compte ici de la complexité du plan d'échantillonnage et nous supposons un échantillonnage aléatoire simple. L'ajustement de divers modèles logistiques avec y_2^* comme réponse et l'examen des statistiques chi-carré du rapport des vraisemblances incitent à formuler un modèle avec

$$x_{ij} = (1 \;\; j \;\; \text{age(2)} \;\; \text{age(3)} \;\; \text{age(4)} \;\; \text{work} \;\; \text{j.age(2)} \;\; \text{j.age(3)} \;\; \text{j.age(4)}),$$

où j est la valeur de y_1^*, age(2) - age(4) sont des indicateurs binaires représentant le facteur âge et «work» est un indicateur binaire du deuxième facteur. Ainsi, le modèle comprend une interaction entre l'âge et y_1^*, qui reflète le fait qu'avec l'augmentation de l'âge, il y a une mobilité décroissante soit de $y_1^* = 0$ à $y_2^* = 1$, soit de $y_1^* = 1$ à $y_2^* = 0$. Par contre, il semble y avoir peu d'indications d'une interaction entre y_1^* et le deuxième facteur ou entre les deux facteurs. Les estimations des paramètres et les erreurs-types sont données à la première colonne du tableau 2.

Tableau 2: Estimations des paramètres pour le modèle logistique.

	Erreur de mesure non prise en compte			Rajustement pour tenir compte de l'erreur de mesure	
Variable auxiliaire	Coefficient estimé	Erreur-type fondée sur le modèle	Erreur-type robuste	Coefficient estimé	Erreur-type
Constante	-2.81	0.34	0.33	-2.75	0.33
y_1	3.13	0.44	0.44	3.61	0.48
age(2)	-0.69	0.47	0.47	-1.11	0.49
age(3)	-1.02	0.53	0.53	-1.74	0.54
age(4)	-0.93	0.53	0.53	-2.26	0.55
y_1.age(2)	0.80	0.65	0.65	1.19	0.71
y_1.age(3)	1.92	0.75	0.74	2.74	0.80
y_1.age(4)	2.54	0.76	0.76	4.76	0.84
work	0.63	0.30	0.29	0.32	0.29

Une source d'information sur l'erreur de mesure réside dans l'étude de validation de la PSID (Hill 1992, p. 29). Cette étude a consisté à comparer les réponses de la PSID avec les dossiers de l'employeur pour un échantillon de travailleurs d'une grande entreprise. Une classification croisée des réponses de l'enquête et des réponses validées, pour la variable des réponses de 1987, est présentée au tableau 3.

Tableau 3: Classification croisée des réponses validées et des réponses de l'enquête, d'après l'étude de validation.

		Enquête y_2^*	
		0	1
Réponse validés y_2	0	140	8
	1	2	302

Si l'on suppose que les matrices des erreurs de classification de l'étude de validation et celles de l'ensemble de la population sont les mêmes, et que les erreurs sont indépendantes avec une distribution identique dans le temps, les nombres observés du tableau 1 sont rajustés selon l'approche exprimée en (12) pour donner la partie gauche du tableau 4.

Tableau 4: Nombres rajustés selon différents modèles de mesure.

Matrices d'erreurs de classification communes

		y_2 = 0	y_2 = 1
y_1	0	764	- 8
	1	3	192

Erreurs non biaisées, α commun

		y_2 = 0	y_2 = 1
		690	27
		37	197

En vertu de ce modèle de mesure, il semble que pratiquement toutes les transitions observées puissent s'expliquer par l'erreur de mesure, et donc qu'il ne donne rien de continuer d'ajuster un modèle logistique. À titre de modèle de mesure de remplacement, supposons maintenant que les erreurs de mesure soient non seulement indépendantes et identiquement distribuées dans le temps, mais aussi non biaisées tant dans l'ensemble de la population que dans l'étude de validation. La valeur estimée de α dans (13) pour l'étude de validation est $\hat{\alpha} = 0.02$. Si l'on suppose que cette valeur (plutôt que les matrices entières d'erreurs de classification) est la même dans l'ensemble de la population et dans l'étude de validation, et qu'elle a une distribution identique dans le temps, la matrice des nombres rajustés, selon l'approche exprimée en (14), est donnée par la partie droite du tableau 4. Ce rajustement est très différent du premier, et ne produit qu'un rajustement modeste du tableau 1. La différence s'explique par le fait que la distribution marginale de y_2^* employée dans l'étude de validation diffère de celle de l'ensemble de la population. Par conséquent, l'hypothèse des erreurs de mesure non biaisées, avec α commun aux deux populations, se traduit par des matrices d'erreurs de classification très différentes.

Si l'on étend le rajustement effectué en vertu du modèle d'erreurs non biaisées au modèle logistique selon l'approche exposée à la section 5, et si l'on suppose un α commun pour toutes les cellules, on obtient les estimations rajustées du tableau 2. Notons que même si le rajustement apparaît de faible ampleur avec $\hat{\alpha}$ seulement égal à 0.02, l'effet sur les coefficients de l'âge et de l'interaction entre l'âge et y_1 est très marqué. Les erreurs-types rajustées tiennent compte de l'erreur d'estimation de α, et il est rassurant de constater qu'elles ne sont pas beaucoup plus élevées que les erreurs-types originales.

REMERCIEMENTS

La présente recherche a bénéficié de la subvention R000 23 2522 du Economic and Social Research Council. L'auteur est reconnaissant envers Keith Humphreys pour l'exécution du travail de calcul sur lequel s'appuie la section 6. Les données ont été rendues disponibles par l'entremise du service d'archivage des données du ESRC, par le Inter-University Consortium for Political and Social Research, Ann Arbor, Michigan. Les données ont été initialement recueillies par J.N. Morgan. Ni le chercheur original, ni le Consortium, ni le service d'archivage n'assument quelque responsabilité que ce soit à l'égard de l'analyse présentée ici.

BIBLIOGRAPHIE

Abowd, H.M., et Zellner, A. (1985). Estimating gross flows. *Journal of Business and Economic Statistics*, 3, 254-283.

Chua, T.C., et Fuller, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

Fuller, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

Hill, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, Sage.

Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Korn, E.L., et Whittemore, A.S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* 35, 795-802.

Maddala, G.S. (1987). Limited dependent variable models using panel data. *Journal of Human Resources*, 22, 307-338.

Meyer, B.D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 385-390.

Muenz, L.R., et Rubinstein, L.V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics*, 41, 91-101.

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

Singh, A.C., et Rao, J.N.K. (1991). Classification error adjustments for gross flow estimates. Technical Report No. 183, Laboratory for Research in Statistics and Probability, Carleton University.

SESSION 10

Questions liées à la qualité

Recueil du Symposium 1992 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

LA QUALITÉ DES DONNÉES ET L'ENQUÊTE LONGITUDINALE DE L'OPCS

I. Macdonald Davies[1]

RÉSUMÉ

L'enquête longitudinale de l'OPCS ne comporte pas de collecte directe de données. Il s'agit d'une enquête par couplage d'enregistrements portant sur de nombreux événements traités de façon courante par l'OPCS, y compris les données des recensements de 1971, 1981 et 1991. En tout temps, cette enquête vise environ 500 000 personnes habitant l'Angleterre et le pays de Galles, soit 1 % de la population.

Le couplage s'effectue d'après les renseignements fournis (nom, sexe, date de naissance) à l'égard de la plupart des événements. Dans cette communication, on étudie les questions de qualité soulevées par ce genre de couplage, les défis que comporte le fait de travailler avec un échantillon dynamique et l'aperçu qu'offre l'enquête longitudinale sur la qualité des sources de données constituantes.

MOTS CLÉS: Longitudinal; couplage d'enregistrements; qualité des données.

1. INTRODUCTION

1.1 L'enquête longitudinale

L'enquête longitudinale (EL) consiste en une collecte d'information par couplage d'enregistrements effectuée par l'Office of Population Censuses and Surveys (OPCS). En tout temps, cette enquête vise environ 500 000 habitants de l'Angleterre et du pays de Galles, soit 1 % de la population. À l'heure actuelle, elle porte sur un échantillon d'enregistrements des recensements de 1971 et de 1981, ainsi que sur des événements liés à l'état civil (naissances chez les femmes appartenant à l'échantillon, décès, cas de cancer, veuvage, etc.) (OPCS 1989). L'enquête tient compte uniquement de renseignements que l'OPCS recueille de façon courante à l'égard de toute la population. On n'effectue aucune collecte directe de données et on n'utilise pas les données de d'autres ministères.

1.2 Appartenance à l'échantillon

L'appartenance à l'échantillon EL est déterminée par le jour et le mois de naissance. Toute personne née l'un des quatre jours fixés pour chaque année appartient à l'échantillon. La fraction de sondage est donc de 4/365, soit de 1.1 %. L'échantillon EL initial a été prélevé à partir des données du recensement de 1971. Pendant la période intercensitaire, on a mis l'échantillon à jour en y ajoutant les personnes nées aux dates pertinentes et les immigrants nés aux mêmes dates et inscrits auprès du Service national de santé (National Health Service). On a retiré de l'échantillon les personnes décédées et les émigrés. À partir de tous les enregistrements du recensement de 1981, on a prélevé un autre échantillon composé de personnes nées aux dates pertinentes; ces enregistrements ont été couplés aux données du fichier EL. La mise à jour de l'échantillon s'est poursuivie pendant toute la décennie. À partir des données du recensement de 1991, on a de nouveau prélevé un échantillon composé de toutes les personnes nées aux dates pertinentes. Le couplage de ces enregistrements aux données du fichier EL est en cours. Le couplage d'enregistrements présente donc une sorte de profil de la vie de chaque individu visé par l'enquête, en fonction des événements enregistrés et connus de l'OPCS.

---

[1] I. Macdonald Davies, Office of Population Censuses and Surveys, Health Statistics, St. Catherine's House, 10 Kingsway, Londres WC2B 6JP.

1.3 Registre central du service national de santé (NHSCR)

Le registre central du service national de santé (NHSCR) est un élément essentiel au déroulement de l'enquête longitudinale. Ce registre est mis à jour au fur et à mesure par l'ajout des naissances et des immigrants qui s'inscrivent auprès d'omnipraticiens. Les membres de la population peuvent être «signalés» dans le registre; les renseignements sur les décès, les cas de cancer ou les sorties du champ de l'enquête (surtout l'émigration) deviennent ainsi accessibles à la recherche.

1.4 Organisation du couplage et consignation des données

Chaque individu visé par l'enquête EL reçoit un numéro EL dès qu'il s'ajoute à l'échantillon. Ce numéro sert à coupler les enregistrements constituants aux données du fichier EL. Il est inscrit au registre principal du NHSCR, dans lequel on consigne uniquement le numéro EL, le nom, la date de naissance et le sexe. Les enregistrements statistiques qui composent le fichier EL sont conservés à part; on y consigne le numéro EL, mais pas le nom. Dans le cas des individus visés par l'enquête et signalés dans le registre, le NHSCR dispose d'un mécanisme permettant d'ajouter le numéro EL à l'échantillon pertinent, établi à la suite du traitement courant des données du recensement et des enregistrements d'événements. Ce numéro EL sert ensuite à coupler aux données du fichier EL la partie statistique de l'enregistrement des données du recensement ou d'un événement. Ce système est nécessaire car, au moment du traitement courant des données du recensement et des événements, les noms ne sont pas entrés dans l'ordinateur. Le NHSCR fournit également des renseignements sur les ajouts attribuables à l'immigration et sur les suppressions (dues surtout à l'émigration). Enfin, le NHSCR enregistre les décès et les cas de cancer survenus chez les individus visés par l'enquête et signalés dans le registre. Ces personnes sont comparées avec celles qui déclarent une date de naissance au moment de l'enregistrement d'un décès ou d'un cas de cancer dans le cadre du traitement courant de ces données par l'OPCS. Idéalement, les deux ensembles d'enregistrements devraient être identiques. Toutefois, certaines personnes ne déclarent pas leur date de naissance de façon cohérente. La contre-vérification permet donc de réduire au minimum l'omission d'événements relatifs aux individus visés par l'enquête.

2. QUALITÉ DES DONNÉES

2.1 Introduction

Le fichier EL offre l'avantage de confirmer une information à partir de sources multiples, ce qui n'élimine pas un certain doute lorsque les sources diffèrent entre elles. La qualité de l'information dépend de celle des sources d'apport, mais peut lui être supérieure grâce aux vérifications supplémentaires. La pertinence de l'échantillon et l'efficacité du couplage interne aux données du fichier EL ont également une incidence sur la qualité. On abordera ici les indicateurs de qualité suivants: taux de repérage, fractions de sondage, taux de couplage et représentativité des enregistrements couplés. Quelques exemples seront également donnés.

2.2 Sources des données constituantes

En Angleterre et au pays de Galles, les habitants sont tenus par la loi de remplir un formulaire de recensement et on déploie de grands efforts pour recenser toute la population. La couverture est presque complète (OPCS 1983, OPCS 1988, OPCS 1992). Les renseignements sur les naissances et les décès sont fournis par les services administratifs chargés d'enregistrer ces données. Si l'enregistrement est exigé par la loi, il est également motivé par des raisons pratiques (ex.: nécessité d'obtenir un certificat de décès pour inhumer ou incinérer un défunt), ce qui permet d'assurer que l'enregistrement est effectivement complet. Les renseignements sur l'immigration et l'émigration sont fournis par le NHSCR, dans la mesure où les personnes visées se sont inscrites auprès d'un médecin (dans le cas de l'immigration) et où le NHSCR est avisé qu'elles ont quitté le pays (dans le cas de l'émigration). Les données relatives à l'émigration sont particulièrement lacunaires, dans une proportion pouvant aller jusqu'à 50 % (OPCS 1988). La dernière catégorie de données recueillies de façon courante dans le cadre de l'enquête longitudinale concerne les cas de cancer. Établies sur une base facultative, ces données sont tirées des registres d'hôpitaux et compilées par l'OPCS pour l'ensemble du pays. La couverture s'est améliorée depuis qu'on a commencé à recueillir ces données en 1971, mais on continue de constater des écarts entre les régions (Swerdlow 1986).

2.3 Repérage au NHSCR

Il est essentiel de «signaler» au registre du NHSCR (repérerage) l'inscription d'un individu visé par l'enquête, pour assurer un degré de couplage élevé des données ultérieures. Près de 98 % des individus appartenant à l'échantillon initial du recensement de 1971 sont maintenant repérés au NHSCR. La plupart ont été repérés dès le début, mais 5 400 l'ont été lorsqu'on a commencé à disposer de renseignements supplémentaires à la suite du recensement de 1981. Toutes les personnes qui s'ajoutent à l'échantillon en naissant ou en immigrant sont automatiquement signalées dans le registre, les nouvelles naissances et les nouveaux immigrants faisant partie du registre principal. Dans le cas du recensement de 1981, plus de 98 % des individus appartenant à l'échantillon ont été repérés. Les taux de repérage diffèrent selon les sous-groupes de la population. Fox et Goldblatt (1982) ont signalé que ces taux étaient légèrement inférieurs chez les femmes (ce qui s'explique par le changement de nom à la suite d'un mariage ou d'un divorce). Les taux de non-repérage les plus élevés concernaient les personnes nées au Pakistan et à Hong Kong. Chez les personnes nées dans ces pays, le nom ne constitue pas un moyen de repérage aussi efficace que chez les personnes nées au Royaume-Uni. Il arrive également que la date de naissance ne soit pas déclarée de façon uniforme.

2.4 Fractions de sondage prévues lors du recensement de 1971

La fraction de sondage prévue à l'égard de l'enquête longitudinale est de 1.1 %. Fox et Goldblatt (1982) ont étudié les fractions de sondage réelles relatives au recensement de 1971. En tenant compte uniquement des individus repérés au NHSCR, ils ont relevé une fraction de sondage réelle de 1.06 dans le cas des hommes et de 1.05 dans le cas des femmes. La variation des fractions de sondage selon l'âge et l'état matrimonial n'était pas plus importante que celles prévues dans le cas d'un échantillon aléatoire de cette taille. Les fractions de sondage étaient faibles dans le cas des personnes ne faisant pas partie d'un ménage privé. Ce groupe comprenait probablement un fort pourcentage de cas difficiles à repérer et de personnes dont la date de naissance n'était pas enregistrée. Le tableau 1 indique les fractions de sondage selon le pays de naissance et le sexe. Dans bon nombre de cas, les fractions de sondage observées diffèrent nettement des fractions prévues, compte tenu de l'ensemble des fractions de sondage et des taux de non-repérage. Dans le cas des personnes nées dans le nouveau Commonwealth asiatique, les taux de non-repérage élevés portaient à croire que ces groupes seraient sous-représentés dans l'échantillon final. Pourtant, ils étaient nettement surreprésentés, une proportion relativement élevée de personnes nées dans ces pays ayant déclaré leur date de naissance sur le formulaire de recensement de 1971.

Tableau 1: Échantillon EL du recensement de 1971: Fractions de sondage observées et prévues selon le sexe et le pays de naissance.

	Hommes		Femmes	
	Fractions observées	Fractions prévues	Fractions observées	Fractions prévues
Angleterre et pays de Galles	1.06	1.07	1.05	1.06
République d'Irlande	1.05	0.96	1.08	0.97
Ancien Commonwealth	0.98	0.98	0.98	0.98
Nouveau Commonwealth asiatique*	1.37	0.88	1.21	0.88
Europe	1.16	0.99	1.15	0.97

* Comprend le nouveau Commonwealth océanien.

2.5 Fractions de sondage prévues pour les événements

Babb et Hattersley (1992) ont étudié les données relatives aux femmes visées par l'enquête qui étaient nées en 1950 et après, et dont les enfants étaient nés après le recensement de 1971. Ces naissances ont été couplées aux données de l'enquête d'après l'enregistrement de la date de naissance de la mère, déclarée au moment de l'enregistrement de la naissance. On a amélioré la qualité des données fournies entre 1971 et 1981 en utilisant celles du recensement de 1981 pour repérer les naissances «omises» initialement, en général parce que la date de naissance de la mère était incorrecte. Babb et Hattersley ont constaté que les fractions de sondage relatives aux femmes visées par l'enquête étaient acceptables (allant de 1.0 à 1.2). Les fractions de sondage relatives aux naissances chez les femmes visées par l'enquête accusaient une variation plus importante: faibles dans le cas des naissances chez les femmes plus âgées, et particulièrement élevées à l'égard des naissances chez les jeunes filles de 15 à 17 ans.

2.6 Taux de couplage: d'un recensement à l'autre

Les taux de couplage permettent de mesurer l'efficacité obtenue en ajoutant de nouveaux renseignements aux données de l'enquête longitudinale. Prenons, par exemple, le couplage entre les recensements de 1971 et de 1981. À partir des données du recensement de 1971, on a prélevé un échantillon composé de toutes les personnes nées aux dates retenues pour les fins de l'enquête. On a prélevé un échantillon équivalent à partir des données du recensement de 1981. Logiquement, l'échantillon de 1971, mis à jour en fonction des ajouts et des suppressions effectués entre 1971 et 1981, devrait être identique à l'échantillon de 1981. Or, ce n'est pas le cas! Il peut manquer un enregistrement, par exemple si la personne n'est pas enregistrée lors du recensement ou si l'enregistrement informatique ne comprend pas la date de naissance. Une personne peut s'ajouter à l'échantillon ou en sortir sans qu'un enregistrement en fasse état (en raison surtout d'un manque d'uniformité dans la déclaration de la date de naissance). Les taux de couplage de données tirées d'un recensement peuvent être mesurés de deux façons: par couplage aval (qu'est-il arrivé aux personnes appartenant à l'échantillon de 1971?) et par couplage amont (toutes les personnes appartenant à l'échantillon de 1981 appartenaient-elles à l'échantillon auparavant?). En tenant compte uniquement des personnes repérées au NHSCR, le taux de couplage aval était de 91 %. Le tableau 2 indique que sur les 513 000 personnes ayant participé au recensement de 1971, 59 000 étaient décédées et 6 000 avaient émigré. On aurait donc dû enregistrer 448 000 personnes à la suite du recensement de 1981; en fait, on n'en a relevé que 408 000. Les 40 000 autres n'ont pu être couplées pour diverses raisons et en particulier à cause du manque d'uniformité dans la déclaration de la date de naissance (37 %) et du non-dénombrement d'un ménage ou d'une personne à son adresse habituelle (38 %) (OPCS 1988). Les taux de couplage aval variaient selon les sous-groupes du recensement de 1971. Les taux de couplage relatifs aux personnes âgées (75 ans et plus en 1971) étaient inférieurs à la moyenne avec un taux de 86 %, probablement à cause du non-couplage des décès. Dans le cas des personnes très âgées (90 ans et plus en 1971), on savait qu'en 1981, 96 % d'entre elles étaient décédées. On n'a pu dénombrer que la moitié des autres, ce qui donnait un taux de couplage de 49 %. Les 2 % manquants correspondaient probablement à des personnes décédées dont le décès n'avait pas été couplé à l'échantillon, sans doute à cause d'une erreur dans la date de naissance. Le taux de couplage amont des recensements de 1971 et de 1981 était de 93 %: sur les 528 000 personnes appartenant à l'échantillon de 1981, 414 000 avaient été enregistrées au recensement de 1971. Durant cette décennie, 64 000 personnes s'étaient ajoutées à l'échantillon en naissant et 14 000 en immigrant. Les 36 000 restantes ne s'étaient ajoutées à l'échantillon EL que lors du recensement de 1981, car elles n'avaient pas déclaré de date de naissance au moment de l'enregistrement d'une naissance (1 000), de l'immigration (4 000) ou du recensement de 1971 (31 000).

Tableau 2: Échantillons EL des recensements de 1971 et de 1981: couplage aval et couplage amont.

Couplage aval		Couplage amont	
	Nombre		**Nombre**
Échantillon du recensement de 1971[1]	513 000	Échantillon du recensement de 1981[2]	528 000
Personnes décédées avant le recensement de 1981	59 000	Personnes nées après le recensement de 1971	64 000
Personnes émigrées avant le recensement de 1981	6 000	Personnes immigrées après le recensement de 1971[3]	14 000
		Naissances ajoutées lors du recensement de 1981	1 000
		Immigrants ajoutés lors du recensement de 1981	4 000
Personnes admissibles lors du recensement de 1981	448 000	Personnes omises lors du recensement de 1971	445 000
Personnes enregistrées lors du recensement de 1981	408 000	Personnes enregistrées lors du recensement de 1971	414 000
Taux de couplage aval	91 %	Taux de couplage amont	93 %

[1] Personnes repérées au NHSCR avant le recensement de 1981.
[2] Personnes repérées au NHSCR.
[3] Personnes immigrées en Angleterre et au pays de Galles, y compris celles qui habitaient l'Écosse en 1971.

2.7 Couplage des événements à l'échantillon EL existant

Le taux de couplage des événements à l'échantillon EL est très élevé. Le mécanisme de contre-vérification des décès et des cas de cancer devrait assurer un couplage presque complet des décès (Fox et Goldblatt 1982) et des cas de cancer (Leon 1988). Dans le cas des événements couplés uniquement d'après la date de naissance déclarée au moment de l'enregistrement (naissances chez les femmes visées par l'enquête et veuvage), le taux de succès est inférieur, le couplage dépendant entièrement de l'exactitude de la date de naissance enregistrée. Babb et Hattersley (1992) ont relevé un taux de couplage de 86 % à l'égard des naissances chez les femmes visées par l'enquête de 1981 à 1988. En utilisant les données du recensement pour dénombrer les naissances «omises», on obtient un taux de couplage plus élevé: 94 % des naissances chez les femmes visées par l'enquête entre 1971 et 1981 ont été couplées à l'échantillon EL. De façon plus générale, l'enquête longitudinale bénéficie d'une nouvelle validation tous les dix ans grâce au tirage d'un échantillon pertinent à partir des données du recensement. On peut ainsi évaluer le nombre d'ajouts (naissances et immigration) et de suppressions (décès et émigration) qui ont été omis. L'enquête longitudinale permet également de vérifier le couplage des événements survenus dans l'intervalle (naissances chez les femmes visées par l'enquête et veuvage).

2.8 Perspectives d'avenir

L'informatisation récente du registre du NHSCR a déjà permis d'ajouter au fichier EL un nombre important d'événements omis. Et la marge d'erreur sera davantage réduite, grâce à l'informatisation plus poussée d'une grande partie du traitement des données relatives aux événements. L'accès aux données du recensement de 1991 permettra d'effectuer une meilleure vérification de la qualité. L'OPCS collabore avec la City University, à Londres, pour établir un rapport technique qui traitera de tous les aspects de la qualité des données de l'enquête jusqu'au recensement de 1991 inclusivement. L'enquête longitudinale sera transférée sur un nouveau système informatique qui facilitera une contre-vérification plus détaillée des enregistrements. La qualité globale des données de l'enquête continuera donc de s'améliorer.

3. UN ÉCHANTILLON EL DYNAMIQUE

L'échantillon EL évolue continuellement. Des personnes s'y ajoutent en naissant ou en immigrant, ou encore en déclarant pour la première fois leur date de naissance sur un formulaire de recensement. D'autres en sortent de façon définitive (décès ou émigration permanente) ou temporaire (émigration et retour ultérieur). Le manque d'uniformité dans l'enregistrement de la date de naissance pose un problème particulier. Si la date de naissance est enregistrée au moment de la naissance, de l'enregistrement auprès du NHS (dans le cas d'un immigrant) ou du recensement, la personne devient membre de l'échantillon sans égard à la date de naissance déclarée à un autre moment (autres recensements, enregistrement d'un cas de cancer ou d'un décès). La base de données EL est structurée de façon à permettre aux chercheurs de prélever le sous-échantillon propre à leur domaine d'intérêt. Il est possible, par exemple, de choisir uniquement les membres qui sont repérés au NHSCR (ce qui est fortement recommandé), tous les membres qui sont présents lors d'un recensement donné, uniquement les membres présents lors de deux ou trois recensements, pour tenir compte de l'évolution de l'échantillon afin de choisir des membres à mesure qu'ils entrent dans le champ de l'enquête. Comme l'explique Hattersley (1992), chaque type de sous-échantillon présente cependant des avantages et des inconvénients.

4. ÉVALUATION DE LA QUALITÉ DES SOURCES DE DONNÉES CONSTITUANTES

4.1 Introduction

L'enquête longitudinale permet d'obtenir des renseignements de meilleure qualité pour les fins d'une recherche fondée sur les résultats. L'exemple le plus généralement reconnu est celui de la mortalité professionnelle (Fox et Goldblatt). Traditionnellement, on avait recours aux données sur le travail et aux données socio-économiques recueillies au moment de l'enregistrement d'un décès et on les comparait aux chiffres démographiques tirés du recensement. L'enquête longitudinale fournit des renseignements tirés du recensement à l'égard de personnes décédées ultérieurement. Elle élimine donc le biais dû à l'utilisation de données provenant de différentes sources. En disposant des données du recensement sur tous les individus visés par l'enquête, on peut également examiner les résultats (décès, veuvage, fertilité, etc.) en fonction de toute l'information socio-démographique recueillie lors du recensement. Bien qu'on cherche avant tout à effectuer une recherche sur un sujet précis, on peut également se rendre compte de la qualité des données constituantes.

4.2 Uniformité de l'enregistrement lors des recensements de 1971 et de 1981

Certaines variables liées à une personne sont immuables et doivent demeurer les mêmes dans tous les enregistrements relatifs à un individu visé par l'enquête. En outre, certains changements s'avèrent impossibles. En comparant les recensements de 1971 et de 1981, OPCS (1988) a relevé des incohérences dans les enregistrements: 0.3 % pour le sexe, 0.4 % pour le pays de naissance (plus de 5 % dans le cas de certains pays autres que ceux du Royaume-Uni) et 3.4 % pour la date de naissance (Figure 8). En outre, 0.4 % des personnes ont enregistré une suite impossible de changements d'état matrimonial (ex.: personne mariée, veuve ou divorcée en 1971 et célibataire en 1981).

4.3 Double dénombrement lors du recensement de 1981

À la suite d'un recensement, une personne peut figurer sur deux formulaires: l'un a trait à son domicile habituel et l'autre à l'endroit où elle a été dénombrée (si elle était absente de son domicile habituel). Comme il concerne des personnes, le fichier EL réunit les deux enregistrements du recensement à l'égard d'un individu visé par l'enquête. Lors du recensement de 1981, 525 000 individus visés par l'enquête ont été dénombrés à leur adresse habituelle. En outre, 4 300 personnes ont été dénombrées ailleurs qu'à leur adresse habituelle **et** inscrites sur le formulaire relatif à leur domicile habituel. Enfin, 6 500 autres personnes ont été dénombrées ailleurs qu'à leur adresse habituelle et inscrites uniquement sur le formulaire pertinent. On peut ainsi évaluer le double dénombrement afin d'interpréter les données relatives à différentes bases de recensement démographique. Il est également possible d'évaluer l'uniformité des renseignements obtenus dans les deux cas.

5. RÉSUMÉ

L'enquête longitudinale de l'OPCS utilise le couplage d'enregistrements de données recueillies de façon courante pour rendre ces données encore plus complètes. Elle constitue une ressource nationale pour l'étude de l'évolution socio-démographique et de la variation socio-démographique des taux des événements. L'échantillon est représentatif et la qualité du couplage d'enregistrements est très élevée. L'utilisation du fichier EL représente un certain défi puisqu'il évolue continuellement et qu'un manque d'uniformité dans l'enregistrement de la date de naissance peut entraîner l'omission de données. L'enquête longitudinale permet non seulement d'effectuer un vaste éventail de recherches, mais aussi d'évaluer la qualité des sources de données constituantes.

BIBLIOGRAPHIE

Babb, P., et Hattersley, L. (1992). An examination of the quality of OPCS Longitudinal Study data for use in fertility analyses. LS User Guide Number 10, London: Social Statistics Research Unit, City University.

Britton, M., et Birch, F. (1985). 1981 Census post-enumeration survey. London: HMSO.

Hattersley, L. (1992). Selecting samples for analysis for the LS. Longitudinal study newsletter no. 7, London: OPCS.

Fox, J., et Goldblatt, P.O. (1982). 1971-1975 Longitudinal study: Socio-demographic mortality differentials. LS series no. 1, London: HMSO.

Leon, D.A. (1988). 1971-75 Longitudinal study: Social distribution of cancer, LS series no. 3. London: HMSO.

OPCS (1983). Census 1971 General Report, Part 3 Statistical assessment. London: HMSO.

OPCS (1988). Census 1971-1981 The Longitudinal study: Linked census data, England and Wales. London: HMSO.

OPCS (1989). Longitudinal Study Newsletter no. 1. London: OPCS.

OPCS (1992). Provisional mid-1991 population estimates for England and Wales and constituent local and health authorities based on 1991 Census results. OPCS Monitor, PP1 92/1, London: OPCS.

Swerdlow, A.J. (1986). Cancer registration in England and Wales: Some aspects relevant to interpretation of the data. *Journal of the Royal Statistical Society Series A*, 149, 146-160. London.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

ÉTUDE DE L'ERREUR NON DUE À L'ÉCHANTILLONNAGE DANS UNE ENQUÊTE LONGITUDINALE PORTANT SUR DES CONTRIBUABLES

S. Hostetter[1]

RÉSUMÉ

Pour répondre aux besoins de données croissants et variés des chercheurs en politique fiscale, l'Internal Revenue Service (IRS) des É.-U. a intégré des études longitudinales à ses échantillons statistiques de déclarants fiscaux. Dans la présente communication, on examine la plus imposante de ces études par panel – fondée sur un panel de 90 000 familles de contribuables avec 1987 comme année de référence. Nous faisons état d'observations anecdotiques et empiriques d'erreurs non dues à l'échantillonnage et nous décrivons les méthodes de correction ou de compensation utilisées par l'IRS pour améliorer les données à des fins d'estimation et de modélisation des politiques.

MOTS CLÉS: Statistiques fiscales; erreur non due à l'échantillonnage; politique fiscale.

1. INTRODUCTION

La présente communication décrit le processus que nous utilisons, à l'IRS, pour corriger les données. Cet examen du processus est un complément à certains travaux plus quantitatifs qui font aussi l'objet d'une présentation à cette conférence (Czajka et Schirm 1992). L'Internal Revenue Service (IRS) des É.-U. a constitué un nouveau panel d'importance, comprenant 90 000 familles de contribuables. Une analyse approfondie effectuée récemment, portant sur les erreurs de déclaration des contribuables et la façon dont l'IRS établit les liens pour créer les familles de contribuables de ce panel, fournit des renseignements utiles sur l'erreur non due à l'échantillonnage.

En 1987, le Bureau d'analyse fiscale du Département du Trésor américain a demandé à la Division des statistiques du revenu de l'IRS d'amorcer un remaniement en profondeur de notre échantillon de déclarations de revenus des particuliers, afin de l'améliorer pour qu'il puisse permettre une modélisation plus précise des effets des recommandations en matière de politique fiscale (Hostetter et O'Conor à paraître). Trois tâches nous étaient confiées:

- **Concevoir et mettre en application une unité familiale** de contribuables, de façon que le comité «Joint Committee on Taxation» du Département du Trésor et du Congrès puisse modéliser l'effet de modifications des lois fiscales sur les unités familiales économiques (Nelson 1986);
- **Redéfinir la stratification de l'échantillonnage transversal**, afin de renforcer l'échantillon de composantes du revenu revêtant une importance pour la politique fiscale, et d'obtenir une meilleure couverture de certains groupes démographiques (Hostetter et coll. 1990; Schirm et Czajka 1991); et
- **Concevoir et mettre en oeuvre un panel de déclarations de revenus des particuliers**, pour mesurer l'effet de la politique fiscale sur le comportement de contribuables individuels dans le temps, plutôt que de suivre l'évolution de grands groupes.

[1] S. Hostetter, Internal Revenue Service et Joint Committee on Taxation, Room 1015 Longworth HOB, Washington, D.C., É.-U. 20515.

Nos efforts relatifs aux deux premiers objectifs ont été décrits dans des communications précédentes, et nous apprécions l'occasion qui nous est offerte, dans le cadre du Symposium sur les enquêtes longitudinales, de faire état de notre travail concernant la conception et la mise en oeuvre du panel de déclarations de revenus de l'IRS, tout en étant à l'écoute de l'information et des opinions de nos collègues. La présente communication comprendra une brève description de quelques panels importants utilisés dans le passé, et mentionnera certaines contraintes générales liées à l'emploi de panels. Toutefois, l'essentiel de la présentation consistera à décrire la méthode employée par l'IRS et analyser les résultats préliminaires suite à la révision d'environ 331 000 dossiers de déclarations de revenus, couplés aussi bien en unités du panel qu'en familles de contribuables. En conclusion, nous examinerons nos projets en ce qui concerne le maintien et l'utilisation de ce panel, ainsi que les plans de l'IRS en vue d'améliorer la qualité des dossiers administratifs.

2. PANELS ANTÉRIEURS ET LIMITES DES ENQUÊTES PAR PANEL

2.1 Panels antérieurs

La formation du panel de l'IRS s'est inspirée de nombreux travaux antérieurs en matière d'utilisation de panels. Le premier panel d'importance -- le Continuous Work History Sample (CWHS) -- a été créé vers la fin des années 1930 par la Social Security Administration (Buckler et Smith 1980), et les premières recherches fondées sur ces données ont été présentées dans les années 1970 (Ruggles et Ruggles 1974). Ces données du CWHS revêtent une importance particulière dans l'élaboration du panel de l'IRS; en effet, l'IRS inclut depuis 1979 un panel de 20 000 numéros de sécurité sociale (NSS) du CWHS dans son échantillon annuel des particuliers, et ces mêmes 20 000 particuliers sont aussi inclus dans le présent panel. Plusieurs autres panels antérieurs ont également servi de base aux méthodes de conception et d'élaboration appliquées au panel de l'IRS. En voici quelques-uns parmi les plus importants:

- **L'étude de «Panel Study of Income Dynamics» (Duncan et coll. 1984),**
- **Le fichier longitudinal canadien de 10 pour cent (Hoskins et Yazdani 1985)**, et
- **L'enquête «Survey of Income and Program Participation» (Kasprzyk et Frankel 1985)**.

Parmi les autres travaux ayant servi de référence aux efforts actuels, notons l'intégration de sources administratives à des données d'enquête (Scheuren 1985; Scheuren 1975), et l'élaboration d'un fichier à usage public (David 1989).

2.2 Limites des données de panels

De toute évidence, les panels mentionnés ci-dessus constituent une abondante source d'information, notamment en ce qui a trait aux cycles de revenus et de richesse ou aux comportements individuels attribuables à la politique fiscale ou à l'évolution de la situation économique. Toutefois, nous avons constaté qu'il y avait des inconvénients et certaines limites associés à ces gains, en particulier lorsque la matière de départ est formée de dossiers administratifs. Par exemple:

- Un panel est un bon outil pour comparer la situation avant et après des **modifications des lois fiscales**, mais pour certains usages des données, des changements importants causent une rupture dans la continuité des données, ce qui peut être une source de problèmes.
- Des **ressources** considérables sont nécessaires pour mettre sur pied et tenir à jour un panel, surtout pour garder la trace de tous les membres du panel. Il est difficile de garder un personnel compétent, ayant reçu toute la formation voulue, au service d'un même projet pendant de nombreuses années.
- Les questions de **pondération**, pour une série de panels, sont très complexes. En général, il faudra utiliser à la fois des poids fondés sur le plan et des poids fondés sur le modèle. Parfois, on peut atténuer les difficultés de pondération en accordant plus d'importance à la stratification initiale, faite au moment de la sélection du panel (Czajka et Schirm 1992).

- La création et le maintien d'un vaste panel sont **coûteux**, si le projet est de longue durée ou s'il doit fournir de l'information sur des caractéristiques variées.

3. DESCRIPTION DU PANEL DE L'IRS

Le panel de la **Division des statistiques du revenu (Statistics of Income Division -- SOI)**, compte environ 90 000 familles de contribuables. Il a été créé pour l'année fiscale 1987, par la sélection de *90 000 déclarations de personnes non à charge (déclarations parentes)* et des déclarations de toute *personne à charge inscrite dans la déclaration parente*. L'échantillon transversal de la SOI de 1987, déjà prélevé, a été désigné comme l'échantillon du panel, après quelques légères modifications seulement (Czajka et Walker 1989). L'échantillon transversal de la SOI comprend des déclarations d'«années antérieures» (déclarations soumises la même année civile que celles de 1987, mais visant une période d'imposition antérieure). De façon générale, ces déclarations ont des caractéristiques différentes de celles soumises à temps, et elles sont jugées représentatives des déclarations qui seront soumises en retard pour l'année d'imposition courante. Ainsi, la composition initiale du panel a été déterminée d'après le plan d'échantillonnage régissant le prélèvement de l'échantillon annuel de base de déclarations des particuliers, lequel a été conçu pour répondre aux besoins variés d'une vaste clientèle, dont aucun n'incorporait de concepts longitudinaux. Cette décision apparaissait à ce moment comme la plus pratique, et puisque nous ne savions pas à l'époque tout ce que nous savons aujourd'hui, il est peu probable qu'un panel sélectionné autrement en 1987 aurait pu, mieux que celui-là, répondre à l'ensemble de nos besoins.

Aux États-Unis, la Social Security Administration attribue à presque tous les particuliers un numéro qui les identifie de façon unique, semblable au NAS (numéro d'assurance sociale) au Canada. Le numéro attribué aux É.-U. porte le nom approprié de numéro de sécurité sociale (NSS). Nous avons créé un fichier comprenant tous les NSS indiqués dans les 90 000 déclarations -- soit plus de 200 000. Le NSS sert à identifier les particuliers faisant partie du panel. En raison des imperfections du système, l'utilisation du NSS fait que certains particuliers qui devraient être membres du panel en sont exclus, et que d'autres qui ne devraient pas en être membres y sont inclus. À des fins de traitement, ce fichier identifie les membres du panel réellement «détenteurs d'une carte». En raison d'un élargissement de l'échantillon dû aux changements de la structure familiale, le panel de la SOI comptait environ 135 000 déclarations pour 1990, mais celles-ci représentent toujours 90 000 unités de panel et seront pondérées en conséquence.

4. PLANIFICATION DE LA RÉVISION DU PANEL

4.1 Objectifs de la révision du panel

Le Projet de révision du panel comportait quatre grands objectifs:

- **Définir la composition du panel,**
- **«Éliminer les erreurs» dans la base de données du panel pour les trois premières années,**
- **Établir des familles de contribuables selon des liens adéquats,**
- **Amasser de l'information pour construire des modèles pouvant servir à des révisions futures du panel.**

La méthode utilisée consistait à passer en revue les liens du panel, les unités du panel, les liens des unités familiales et les dossiers individuels pour les trois années, de façon informatisée si possible, et manuellement lorsque les unités ne répondaient pas aux critères d'appariement rigides appliqués au départ. À l'automne 1991 (quatre ans après la mise sur pied du panel), le personnel avait créé une base de données contenant toutes les déclarations du panel sélectionnées pour 1987, 1988 et 1989.

Cinq ans après l'établissement du panel (à l'automne 1992), la SOI avait terminé la vérification de 331 000 déclarations, y compris l'examen manuel de plus de 150 000 déclarations visant les trois premières années. Une telle durée est essentielle au calendrier de correction et d'amélioration, car on ne peut examiner intelligemment les liens entre les membres du panel, et les activités de ces derniers, si l'on ne dispose pas des données d'au

moins trois ans. Contrairement aux données d'une seule année, les données de trois ans offrent assez d'information pour permettre de faire la différence entre un changement et une erreur. Par conséquent, le personnel de la SOI n'a pu établir les liens des familles de façon *fiable* qu'une fois terminé ce travail d'élimination des erreurs.

4.2 Définition d'unité familiale de contribuables

La famille de contribuables est définie comme celle formée du(des) contribuable(s) non à charge figurant dans la déclaration de revenus d'un particulier et de toutes les personnes à charge inscrites par le(s) contribuable(s) non à charge, pour une *année particulière*. Le contribuable non à charge peut être soit la première personne inscrite -- le déclarant principal sur une déclaration de personne seule, ou les déclarants principal et secondaire (conjoint) sur une déclaration commune. Dans plus de 90 % des déclarations communes, l'homme est inscrit comme déclarant principal. Pour établir la famille de contribuables, la SOI a couplé toutes les déclarations de revenus de personnes à charge à la déclaration de revenus sur laquelle elles ont été déclarées à charge -- la déclaration parente. La plupart de ces personnes à charge sont des enfants qui sont tenus de remplir une déclaration malgré le fait que leurs parents réclament une exemption à leur égard.

La famille de contribuables est établie administrativement d'après l'information des déclarations de revenus plutôt que par la collecte de données d'enquête et des contacts avec les répondants, méthodes normalement utilisées pour établir les «ménages», l'unité utilisée plus couramment pour réunir des caractéristiques sur les revenus et la démographie. L'utilité des familles de contribuables dans l'analyse de la politique fiscale tient au fait qu'une telle famille représente une unité économique partageant le même revenu commun. Nous nous rendons compte que, comme les ménages ou d'autres définitions de famille largement utilisées, cette unité comporte sans doute des imperfections, mais c'est celle dont nous disposons, à l'IRS.

La SOI construit des familles de contribuables aussi bien pour son échantillon transversal que pour son panel, mais c'est au panel que nous nous attarderons dans la présente communication. Les familles de contribuables comportent des liens valables pour une seule année, contrairement aux unités du panel, auxquelles les membres appartiennent pour la vie (ou plutôt la durée d'existence du panel). Les familles de contribuables peuvent demeurer constantes, mais elles peuvent aussi changer. Par exemple, si un membre jusque-là déclaré comme personne à charge, une fois qu'il atteint l'âge adulte, quitte la maison et n'est plus inscrit comme personne à charge, il deviendra lui-même une unité familiale distincte s'il soumet une déclaration. Une partie du travail de révision consistait à déterminer si un changement était attribuable à une véritable modification de la dynamique familiale, qu'il s'agisse d'une situation courante, comme dans notre exemple, ou d'une situation inhabituelle, ou encore si le changement était dû au fait que la personne n'était pas la même, auquel cas une correction s'imposait. Environ 70 % des déclarations sont soumises par des contribuables mariés; ce sont presque toutes des déclarations communes, et seules quelques-unes d'entre elles proviennent de «mariés soumettant une déclaration distincte». Les 30 % qui restent proviennent de personnes seules, et environ 20 % de ces dernières sont des déclarations de «chef de ménage». Les contribuables soumettant des déclarations de «chef de ménage» sont des personnes seules qui doivent avoir une personne à charge répondant à certaines conditions. Les contribuables soumettant des déclarations de personnes seules peuvent aussi avoir des personnes à charge.

4.3 Problèmes reliés aux NSS des personnes à charge

Si le panel a été établi en 1987, c'est que pour la première fois cette année-là, les contribuables devaient indiquer le NSS des personnes à charge, de sorte que l'information des déclarations de revenus permettait désormais d'établir des liens entre les personnes pour former des familles. Auparavant, la SOI n'aurait pas pu coupler les déclarations des personnes à charge à la déclaration parente. La loi a été introduite progressivement sur une période de trois ans, avec comme exigence finale la déclaration du NSS pour les enfants d'un an ou plus. Fait peu étonnant, les contribuables ont été lents à se plier à cette exigence, trouvant toute une panoplie de méthodes pour s'y soustraire. Dans notre révision, nous avons considéré comme personne à charge un membre du panel si une exemption avait été demandée à son égard en 1987, peu importe qu'un NSS valide ait été indiqué ou non. Toutefois, nous n'avons pas pu inclure de telles personnes à charge dans le fichier avant que leur NSS exact soit indiqué dans la déclaration d'une année ultérieure.

Certains contribuables ont attribué leur propre NSS à leurs personnes à charge, certains ont omis de l'inscrire et d'autres ont emprunté le NSS d'autres familles ou d'amis. Dans un cas, une contribuable a utilisé le NSS d'un ex-conjoint pour une de ses personnes à charge. Dans sa déclaration d'une année ultérieure, l'ex-conjoint et sa famille ont été intégrés à la famille du panel original en raison de ce NSS commun et, surprise, nous avons constaté que l'ex-conjoint réclamait une exemption à l'égard des deux mêmes enfants. Bien que cet exemple précis représente une situation rare, toute une variété de cas semblables ont été fréquemment relevés.

4.4 Couplage initial des déclarations du panel

Le NSS est l'élément constant qui a permis de créer et de réviser les liens et la structure des familles, et d'en déterminer les modifications. La base de données du panel que le personnel de la SOI a établie en 1991 était notre premier fichier avec liens familiaux disponible pour révision. Le premier effort, comme c'est le cas le plus souvent dans un nouveau travail, a été difficile parce que nous ne savions pas quel niveau d'erreur il fallait prévoir à des fins de planification, nous ignorions quelles méthodes de révision se révéleraient les plus fructueuses et nous ne savions pas comment les erreurs pourraient être regroupées en catégories pour permettre une analyse appropriée. Dans un environnement de production comme celui de l'IRS, ces difficultés sont particulièrement aiguës.

Dans cette première intervention, nous avons choisi de ne pas faire d'hypothèses concernant la validité des liens des familles ou du panel -- et d'accorder la préférence à la méthode de couplage la plus simple et la plus apte à englober le *maximum* de déclarations. Étant donné qu'aux États-Unis les personnes à charge peuvent soumettre une déclaration distincte tout en étant l'objet d'une demande d'exemption sur la déclaration de leurs parents, le processus de couplage initial a consisté à relier *à la déclaration parente* toute déclaration soumise par des personnes à charge inscrites sur la déclaration parente. Avant d'avoir l'occasion de faire une révision, nous ne voulions pas retrancher du panel, ou omettre de coupler, toute déclaration dont le NSS principal ou secondaire était celui d'une personne à charge faisant l'objet d'une exemption sur une déclaration parente du panel.

4.5 Comportement des contribuables à l'égard de l'information à déclarer

Notre capacité de construire des familles de contribuables exactes par la sélection et le couplage de déclarations était évidemment liée aux facteurs humains agissant -- pas toujours de façon positive -- sur le comportement des contribuables face à l'information déclarée. Le comportement face à la déclaration de revenus représente une part importante de notre erreur non due à l'échantillonnage. Comment procéder pour attribuer chaque personne au panel, à l'unité du panel et à la famille de contribuables appropriés? C'est une tâche ***très difficile***, et nous avons élaboré trois fichiers particuliers pour faciliter le suivi et la gestion des variations de ces caractéristiques.

La création et la tenue à jour de familles de contribuables sur un certain nombre d'années sont en outre compliquées par le fait que les contribuables commettent ce que nous avons appelé des «family matching sins» («entorses aux liens familiaux»), dont sept cas seulement sont présentés ci-après.

- **Mariage** - un changement d'état matrimonial exige que la situation soit décelée et révisée. Le pire cas survient si le nouveau conjoint est un autre membre du panel, car il faut alors aussi corriger la pondération,
- **Divorce** - la conséquence est qu'il existe deux familles dans la même unité du panel,
- **Remariage** - l'effet est d'introduire un non-membre (un «VISITEUR») dans le panel, et peut-être aussi en même temps des visiteurs à charge (un visiteur étant toute personne figurant dans une déclaration du panel sans être membre du panel),
- **Inscription de personnes à charge** - (enfants ou parents) qui ne sont pas membres du panel, et qui s'ajoutent donc aux visiteurs,
- **Divorce entre un membre et un visiteur** - le visiteur doit alors être retranché de la sélection active et, par la suite, on doit faire de même pour tout visiteur à charge qui n'est pas réclamé par le membre du panel,

- **Partage du NSS du déclarant**, et
- **Indication d'un NSS inexact**.

Seuls les deux derniers cas représentent des erreurs. Les autres rendent nécessaire une vérification ou une correction. Ce processus sera permanent et bénéficiera de l'expérience de révision décrite dans cette communication, qui nous a beaucoup appris sur les problèmes causés par des éléments humains comme ceux qui ont été mentionnés.

5. RÉVISION DU PANEL

5.1 Contribuables individuels: définition et «retouches»

Après avoir établi des liens entre les déclarations pour établir des familles de contribuables «préliminaires», il nous restait à vérifier que ces liens étaient raisonnables et créaient réellement des familles de contribuables. Pour ce processus d'élimination d'erreurs, l'élément le plus important à vérifier ou à modifier était, de toute évidence, le NSS. Les données sur les **personnes** ont été révisées et corrigées d'après l'information des **déclarations de revenus**. Même si des erreurs possibles ayant trait aux personnes ou aux déclarations de revenus étaient repérées en vue d'une révision, dès que nous examinions un élément du panel, nous regardions l'unité au complet, c'est-à-dire *L'ENSEMBLE des déclarations pour L'ENSEMBLE des trois années*. La correction des NSS était le but principal, mais le processus de révision et de correction visait aussi l'identification des unités du panel et des unités familiales. Pour que les corrections soient plus claires, des codes ont été ajoutés afin de préciser quelle personne en particulier, dans la déclaration, était visée par une correction. En outre, nous avons attribué des codes d'état décrivant le genre d'intervention -- par exemple suppression ou modification -- et des codes de raison décrivant la cause ou les circonstances de l'erreur. Ces codes additionnels serviront dans l'avenir à l'élaboration de modèles de révision informatisés.

Les six zones de données auxquelles nous avons apporté des corrections sont les suivantes:

- **Le numéro d'identification du panel**,
- **Le numéro d'identification de la famille**,
- **Le NSS du contribuable principal**,
- **Le NSS du contribuable secondaire**,
- **Les NSS des personnes à charge (jusqu'à un maximum de 10)**,
- **Le code d'état du déclarant au moment de l'entrée dans le panel**.

Certains éléments de données ayant trait au revenu et à l'impôt ont été inclus dans les documents de révision à des fins d'information, mais aucun n'a été modifié. Nous n'avons pas tenté, non plus, de corriger l'information pour la rendre conforme au code d'imposition. Notre objectif était plutôt de recueillir l'information sur les contribuables, sans signaler ni traiter les erreurs. En d'autres termes, nous voulions éliminer et mesurer l'erreur non due à l'échantillonnage liée à l'identification des particuliers et des familles de contribuables.

5.2 Documents d'information ayant servi à la vérification et à la correction

Nous disposions d'une vaste gamme de sources d'information pertinentes au moment de notre révision. D'après les données des déclarations, et grâce au fichier comptable principal de l'IRS, nous avons extrait et utilisé les renseignements suivants:

- **Nom complet du (des) contribuable(s)**,
- **Adresse du (des) contribuable(s)**,
- **NSS du déclarant principal**,

- **NSS du déclarant secondaire,**
- **NSS de toutes les personnes à charge,**
- **Catégorie de déclarant (état matrimonial),**
- **Indicateur de catégorie de personne à charge,**
- **Nombre et type des exemptions,**
- **Noms de toutes les personnes à charge,**
- **Certaines données sur le revenu et l'impôt.**

En utilisant comme source l'information sur la sélection de l'échantillon de la SOI et sur le processus initial de formation du panel, nous avons extrait les données suivantes:

- **Codes de définition des strates de l'échantillon,**
- **Code de lien familial douteux,**
- **Contrôle de nom d'après la déclaration de revenus,**
- **Année fiscale visée par la déclaration,**
- **Numéro d'identification de la famille,**
- **Numéro d'identification du panel.**

Enfin, nous avons obtenu de la Social Security Administration (SSA) trois éléments de données très importants:

- **Contrôle de nom (quatre premières lettres de chaque nom de famille inscrit à la Sécurité sociale pour le numéro de sécurité sociale concerné) pour *tous les NSS indiqués*,**
- **Date de naissance pour *tous les NSS indiqués*,** et
- **Date de décès pour *tous les NSS indiqués*.**

Ces dernières données étaient cruciales pour l'exactitude de notre révision. Par exemple, si le contrôle de nom du contribuable principal (noté par l'IRS d'après le nom inscrit sur la déclaration) ne correspondait pas au contrôle de nom de la SSA pour le NSS indiqué, il y avait de fortes chances que la déclaration entière soit incorrectement représentée par le NSS de base -- celui du contribuable principal. Souvent, nous relevions une autre déclaration, pour la même année ou l'année suivante, ayant le même NSS, mais provenant d'une autre personne (Steffick 1992).

5.3 Organisation des unités du panel en vue de la révision

Après avoir établi les liens initiaux et effectué le travail d'expérimentation, le personnel de la SOI a amélioré les définitions des groupes de révision initiaux et en a créé plusieurs nouveaux. Les groupes de révision appartenaient à l'une ou l'autre de deux catégories -- «exempt d'erreur» et «douteux». Les unités des groupes de la catégorie «exempt d'erreur» étaient déterminées de façon automatisée, l'exactitude des définitions étant vérifiée à l'aide d'une révision manuelle d'un échantillon d'unités. Les unités des groupes «exempts d'erreur» répondaient à des critères stricts; elles avaient les caractéristiques suivantes:

- **Aucune donnée non concordante,**
- **Aucun changement,**
- **Zéro, une ou deux personnes à charge,** et
- **La déclaration de l'année de base ne visait pas une année antérieure.**

On considérait qu'il y avait données non concordantes dans les cas suivants: contrôles de noms pour les NSS ne correspondant pas aux noms inscrits sur la déclaration; noms de contribuables secondaires ou de personnes à

charge ne correspondant pas au nom du contribuable principal; ou codes postaux ZIP des personnes à charge ne correspondant pas à celui des parents. Par «aucun changement», on entendait **aucune modification**, d'une année à l'autre, de la catégorie de déclarant, des NSS, des personnes à charge ou du code ZIP. Au départ, les déclarations relatives à des années antérieures -- celles incluant l'année fiscale 1986 ou une année précédente -- n'étaient pas incluses dans ce groupe «exempt d'erreur», peu importe leurs autres caractéristiques. Après avoir fait un certain nombre de révisions manuelles de déclarations sélectionnées pour l'année de référence mais couvrant une année antérieure, nous avons éliminé cette restriction, de sorte que si toutes les autres caractéristiques demeuraient constantes, les déclarations visant des années antérieures pouvaient être considérées «exemptes d'erreur». Cette décision a permis de réduire de 1,517 le nombre de déclarations exigeant une révision manuelle.

Une fois terminée et vérifiée la révision informatique, il restait 153,153 déclarations, sur un total de 330,956, à réviser manuellement. Les critères énoncés ci-dessus nous avaient donc permis de désigner 54 % des déclarations comme «exemptes d'erreur» après la vérification informatique.

5.4 Formation des groupes «douteux» pour la révision manuelle

Bien que nous ne sachions trop à quels types d'erreurs nous attendre, nous avons décidé, afin de bien gérer la révision et de disposer des rapports des données sommaires sur les divers types d'erreurs, de regrouper les déclarations à réviser selon certaines catégories d'erreurs. Ces groupes de révision englobaient différentes situations d'erreur, mais n'en excluaient pas d'autres, de sorte qu'une déclaration donnée n'était pas nécessairement incluse de façon unique dans un groupe de révision. Par conséquent, l'ordre dans lequel les groupes étaient traités faisait une différence. Plus précisément, les quelques groupes que nous souhaitions voir englober toutes les unités visées -- pour nous permettre d'évaluer la fréquence ou l'ampleur d'un type d'erreur particulier -- devaient être choisis en premier. Dans tous les cas, toutes les erreurs possibles dans les données d'une unité du panel étaient examinées dès que l'unité était attribuée à un groupe. La liste ci-dessous fournit certaines des descriptions abrégées utilisées dans la formation des groupes de contrôle ayant servi à la révision manuelle. Les définitions et l'ordre de priorité étaient encore sujets à changement (et, en fait, certaines modifications ont été faites) au cours de la première phase de production.

- **Un code d'état relatif à une personne à charge avait été attribué à la déclaration de l'année de base,**
- **Déclarations d'années précédentes qui n'étaient pas «exemptes d'erreur»,**
- **Déclarations pour lesquelles les contribuables principal et secondaire représentaient des unités du panel différentes,**
- **Déclarations comportant des visiteurs (personnes autres que des membres du panel),**
- **Déclarations couplées pour cause de personne à charge, mais ayant un code d'état visant une personne non à charge,**
- **Contrôle de nom sur la déclaration ne correspondant pas au contrôle de nom de la sécurité sociale,**
- **Personne mariée soumettant une déclaration distincte à l'intérieur d'une même unité du panel,**
- **NSS secondaire inconsistant d'une année à l'autre.**

Nous avons trouvé utile de regrouper ainsi les déclarations par type d'erreurs, non seulement pour gérer la progression du projet, mais aussi pour former le personnel, car cette méthode permettait de consolider rapidement les notions apprises. L'uniformité de la révision visant les cas semblables s'est également révélée utile **à la compréhension et à la documentation des conditions d'erreur systématique**. Au cours des réunions d'évaluation et de planification, nous avons pu faire la distinction entre les erreurs systématiques et les erreurs exceptionnelles. Nous avons pu, par conséquent, établir des méthodes convenant aux conditions systématiques, et repérer et traiter adéquatement les situations d'exception.

5.5 Recherches additionnelles pour des cas exceptionnels

Malgré l'abondance des données dont disposait le personnel effectuant la révision, certaines erreurs ne pouvaient être corrigées d'après l'information contenue dans l'ensemble des déclarations des trois années. Par exemple,

nos données ne nous permettaient pas de corriger le dossier d'un contribuable inclus dans le panel d'après sa déclaration de 1987, sur laquelle figurait un NSS erroné. C'était la personne elle-même, avec les caractéristiques fiscales sélectionnées, que nous voulions inclure, mais si nous n'avions pas son NSS, nous ne pouvions la trouver de nouveau. Si les données des trois années ne contenaient aucune information nous permettant de corriger le NSS, le personnel faisait une interrogation par nom dans le fichier comptable principal de l'IRS afin de déterminer le NSS exact. Même si cette étape additionnelle était longue et coûteuse, nous la considérions justifiée dans le cadre de ce premier travail d'élimination d'erreurs. Si nous pensions ainsi, c'était en partie parce que nous percevions cette élimination d'erreurs initiale comme un travail **à la fois de production et de recherche.** Le travail de production consistait évidemment à éliminer les erreurs dans le fichier, comme nous venons de le décrire. Le travail de recherche consistait à:

- **Utiliser différentes méthodes de révision,**
- **Documenter nos méthodes,**
- **Documenter nos résultats,**
- **Évaluer ces méthodes.**

L'évaluation sera un processus permanent dans le cadre duquel nous étudierons les méthodes sous l'angle de leur rendement, de leur exactitude technique et de leur efficience.

6. RÉSULTATS INITIAUX DE LA RÉVISION DU PANEL

La figure 1 présente une répartition des taux d'erreur pour chaque catégorie de NSS. Fait surprenant, les NSS de personnes à charge comportaient le même taux d'erreur que les NSS secondaires. En outre, un plus grand pourcentage d'entre eux ont été corrigés que ce ne fut le cas pour les NSS secondaires. Ce résultat tient peut-être au fait que l'exactitude des NSS de personnes à charge s'est améliorée constamment, au fur et à mesure que les contribuables se conformaient à la nouvelle exigence. En revanche, quand une erreur touchant le contribuable secondaire survient, elle est souvent reproduite d'une année à l'autre par retranscription des données de l'année précédente. Une correction de NSS était effectuée uniquement lorsque nous étions sûrs -- d'après le contrôle de nom, l'âge, et le nom et l'adresse inscrits sur la déclaration -- de connaître le NSS exact. Les NSS secondaires affichent le «pourcentage de non-corrections» le plus élevé, ce qui correspond à des cas où il nous était impossible de corriger ce qui apparaissait comme un NSS inexact. Le pourcentage d'erreurs total, tant pour les NSS secondaires que pour les NSS de personnes à charge, était de 3.5 %. Comme prévu, l'erreur et le «pourcentage de non-corrections» pour les contribuables principaux étaient minimes. Connaissant le processus administratif mis en oeuvre par l'IRS pour examiner, rejeter et corriger les NSS des contribuables principaux, nous avions présumé qu'il y aurait très peu d'erreurs touchant ce groupe au moment de l'extraction des déclarations du fichier principal par la SOI. Nous savons que l'IRS s'assure que le NSS principal est un NSS valide et que le contrôle de nom correspond à ce NSS. Le traitement fait par l'IRS permet non seulement d'éliminer les erreurs de déclaration telles qu'une mauvaise transposition de caractères, mais il décèle et empêche également la plupart des erreurs d'entrée au clavier.

Figure 1: Taux d'erreur par type de NSS.

Type de NSS	Pourcentage d'erreurs	Pourcentage de corrections	Pourcentage de non-corrections
Principal	0.18	0.03	0.15
Secondaire	3.50	1.10	2.40
Personne à charge	3.50	1.70	1.80

Figure 2: Erreurs de NSS selon le nombre de chiffres inexacts.

Type de NSS	Pourcentage selon le type d'erreur		
	1-2 chiffres	3-4 chiffres	5 chiffres et plus
Principal	33.7	1.3	56.0
Secondaire	76.7	2.5	20.8
Personne à charge	81.4	4.3	14.3

La figure 2 indique la nature des erreurs observées dans les NSS. Dans le cas des NSS principaux, ce sont les erreurs de cinq chiffres ou plus, considérées comme reflétant l'inscription d'un mauvais NSS, qui ont été les plus fréquentes. En revanche, les NSS de déclarants secondaires et de personnes à charge sont ceux qui affichent le pourcentage le plus élevé d'erreurs connues de 1 ou 2 chiffres. Il s'agissait très probablement d'erreurs de transcription du contribuable ou du préposé à l'entrée des données au clavier. Cela est conforme à ce que nous savons des normes de traitement de l'IRS. Ces résultats montrent que nous avons une quantité non négligeable d'erreurs dans les NSS de déclarants secondaires et de personnes à charge. Toutefois, ces estimations sont préliminaires, et la fréquence des erreurs -- notamment celles touchant les NSS de déclarants secondaires et de personnes à charge -- observée dans le panel ne peut permettre de tirer des conclusions quant à la population totale des déclarations de revenus. Comme il est signalé dans la dernière section de cette communication, l'IRS entend améliorer la qualité de ces deux catégories de NSS.

7. LE PANEL CONSOLIDÉ DE LA SOI

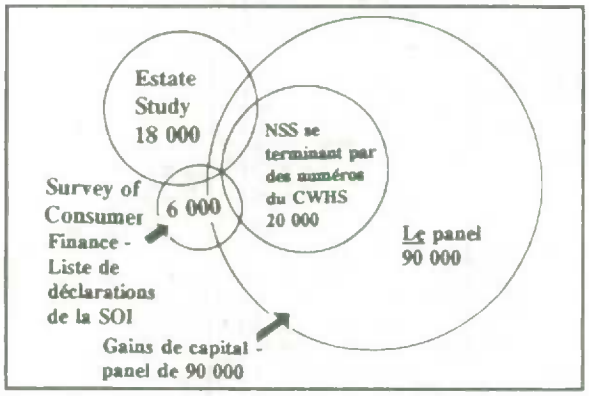
Voilà qui termine la description du processus initial de révision du panel. Le panel de base que nous venons de décrire, toutefois, n'est qu'un point de départ. Notre plan est d'élaborer un **panel consolidé** qui recoupera le panel de base, afin de mettre plusieurs sources à contribution pour l'analyse des politiques. Autrement dit, en superposant plusieurs ensembles de données longitudinales relatives aux contribuables individuels, nous pourrons, à un moindre coût, offrir des données plus riches à l'appui de chacune des études particulières. (Voir la figure 3.) L'examen de l'erreur non due à l'échantillonnage dans la présente communication s'est limité au panel de contribuables *de base*, mais il devrait être considéré dans le contexte du «panel consolidé», car nous prévoyons faire un travail similaire pour d'autres panels dans l'avenir (Hostetter 1992). Les quatre autres panels importants qui seront inclus dans la conception, la gestion et le traitement des panels de la SOI sont les suivants:

- Comme nous l'avons mentionné précédemment, un échantillon de 20 000 NSS -- **le CWHS** -- est inclus dans le panel. Des familles de contribuables ont donc été établies pour ce groupe. Ce groupe offre un recoupement direct avec l'échantillon transversal, ce qui se révélera utile dans des études comparatives utilisant des séries chronologiques.

- L'enquête **Survey of Consumer Finances** est menée tous les trois ans auprès des ménages par le Federal Reserve Board, en collaboration avec la SOI. Elle est utilisé par le Federal Reserve Board, le Congrès, le Département du Trésor et des chercheurs pour étudier une vaste gamme de caractéristiques financières des ménages (Kennickell et Woodburn 1992).

- Mettant à profit les résultats d'une étude pilote sur 18 000 personnes décédées et bénéficiaires figurant sur les déclarations de succession de 1989, la SOI entreprendra en 1993 une étude appelée **Estate Collation Study**, qui comportera un recoupement partiel des bénéficiaires et de leurs familles de contribuables (Johnson et Woodburn 1992).

Figure 3: Composantes du panel consolidé.



- À compter de 1993, la SOI incorporera des **données sur les gains de capital** pour le panel entier de 90 000 unités. Les unités du panel et, à l'intérieur de ces dernières, les familles de contribuables, demeureront les mêmes; la différence proviendra de l'information additionnelle -- les codes d'actif et les renseignements sur les transactions pour tous les gains et pertes de capital signalés dans les déclarations du panel (Holik, Hostetter et Labate 1989).

8. PROJETS FUTURS

Nous en arrivons maintenant à nos projets pour l'avenir -- tant à la SOI qu'au groupe de traitement de l'IRS. Au cours de 1992 et de 1993, la SOI améliorera les méthodes d'élimination des erreurs dans les panels, en utilisant des modèles fondés sur des données quinquennales et l'information sur les erreurs de déclaration des contribuables recueillie au cours du processus décrit dans cette communication. L'exactitude et la rapidité seront améliorées par l'intégration du processus d'élimination des erreurs au processus de production interactif de la SOI, d'abord à la fin et, en 1993, au début du processus de traitement statistique de la SOI.

En outre, en 1993, la SOI perfectionnera son étude des correspondances entre les déclarations de renseignements et les déclarations de revenus, qui s'appuie sur le panel décrit dans cette communication, et qui vise à apporter, dans toute la mesure du possible, des corrections aux NSS. L'IRS a déjà commencé à vérifier les NSS secondaires dans le cadre de son traitement courant du fichier principal, et vérifiera les NSS des deux premières personnes à charge en 1993. Ainsi, le taux d'erreur dans les correspondances entre NSS, établies tant pour améliorer la qualité du panel que pour accroître la capacité de l'IRS de faire d'importantes contributions aux estimations visant la population américaine, se rapprochera du faible taux d'erreur actuellement observé dans le cas des NSS principaux -- 0.2 %.

BIBLIOGRAPHIE

Buckler, W., et Smith, C. (1980). The continuous work history sample (CWHS): Description and contents. *Economic and Demographic Statistics*, Social Security Administration.

Czajka, J.L., et Schirm, A.L. (1992). Selection and maintenance of a highly stratified panel sample. *Recueil du Symposium 92 de Statistique Canada: Conception et analyse des enquêtes longitudinales*, Ottawa (Ontario) Canada.

Czajka, J.L., et Walker, B. (1989). Combining panel and cross-sectional selection in an annual sample of tax returns. *American Statistical Association 1989 Proceedings of the Section on Survey Methods*.

David, M.H. (1989). Managing panel data for scientific analysis: The role of relational database management systems. *The American Statistical Association International Symposium on Panel Surveys*, John Wiley & Sons.

Duncan, G., et coll. (1984). The role of panel studies in a world of scarce resources. *The Collection and Analysis of Economic and Consumer Behavior Data* (S. Sudman et M.A. Spaeth, Éds.), Bureau of Economic and Business Research, Champaign, IL.

Holik, D., Hostetter, S., et Labate, J. (1989). Sales of capital assets. *American Statistical Association 1989 Proceedings of the Section on Survey Research Methods*.

Hoskins, E., et Yazdani, M. (1985). Some longitudinal methodologies and issues relevant to the modelling and analysis of tax policies and programs. *Multinational Tax Modelling Symposium Proceedings*, Revenu Canada Impôt.

Hostetter, S. (1992). Managing multiple uses of panels. *American Statistical Association 1992 Proceedings of the Section on Social Statistics*.

Hostetter, S., et O'Conor, K. (à venir). Satisfying the need of income policy modelers while preserving the reliability of descriptive statistics. *Statistics of Income Methods and Results -- From Data to Information: 1991-1992*, Internal Revenue Service.

Hostetter, S., et coll. (1990). Choosing the appropriate income classifier for economic tax modeling. *American Statistical Association 1990 Proceedings of the Section on Survey Research Methods*.

Johnson, B., et Woodburn, L. (1992). The underlying methodology of the estate multiplier technique: Recent improvements and estimates for 1989. Discussion présentée au «1992 Joint Statistical Meetings», Boston, MA.

Kasprzyk, D., et Frankel, D. (éds.) (1985). *Survey of Income and Program Participation and Related Longitudinal Surveys: 1984*, Bureau of the Census.

Kennickell, A.B., et Woodburn, L. (1992). Methodological issues in the estimation of household net worth: Results from the 1989 survey of consumer finances. *American Statistical Association 1992 Proceedings of the Survey Research Section*.

Nelson, S.C. (1986). Family economic income and other income concepts used in analyzing tax reform. *Compendium of Tax Research, 1986*, Department of Treasury, Office of Tax Analysis.

Ruggles, N.D., et Ruggles, R. (1974). The anatomy of earnings behavior. *The Distribution of Economic Wellbeing*, (F. Thomas Juster, Éd.), Cambridge, MA, Ballinger.

Scheuren, F. (1985). Methodological issues in linkage of multiple data bases. *Record Linkage Techniques - 1985*, Internal Revenue Service.

Scheuren, F. (1975). ORS management of the HEW income security survey -- some administrative issues. Document de travail, Office of Research and Statistics, Social Security Administration.

Schirm, A.L., et Czajka, J.L. (1991). Alternative designs for a cross-sectional sample of individual tax returns: The old and the new. *American Statistical Association 1991 Proceedings of the Section on Survey Research Methods*.

Steffick, D. (1992). Analyzing longitudinal data linkages in a panel of individual tax returns. *American Statistical Association 1992 Proceedings of the Section on Social Statistics*.

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

UTILISATION DE DONNÉES ADMINISTRATIVES POUR ÉVALUER LA QUALITÉ DES DONNÉES SUR LE REVENU RECUEILLIES LORS DE L'ENQUÊTE «SURVEY OF INCOME AND PROGRAM PARTICIPATION»

J.F. Coder[1]

RÉSUMÉ

L'enquête «Survey of Income and Program Participation» (SIPP) sert à recueillir un ensemble étendu de renseignements sur la situation financière des ménages américains. Les interviews sont réalisées à des intervalles de 4 mois et les données sur le revenu et les antécédents professionnels sont recueillies pour chaque mois d'une période de référence de 4 mois. Dans la présente communication on étudie l'exactitude des données sur le revenu provenant des salaires et traitements recueillies dans le cadre de l'enquête SIPP de 1990, à l'aide d'un appariement «exact» entre les données de l'enquête et les données tirées des déclarations fédérales de revenus des particuliers pour 1990. Diverses mesures de l'importance et des caractéristiques de la réponse à l'enquête et des erreurs d'imputation sont présentées.

MOTS CLÉS: Erreur de réponse; revenu; qualité des données; couplage des données.

1. INTRODUCTION

Dans la présente communication j'examine l'importance et les caractéristiques de l'erreur de mesure[2] pour les données sur le revenu provenant des salaires et traitements recueillies dans le cadre de l'enquête SIPP. L'erreur de mesure a été définie ici comme l'écart entre les réponses à l'enquête pour les montants du revenu provenant des salaires et traitements et les montants «comparables» qui figurent sur les déclarations fédérales de revenus des particuliers pour l'année civile 1990. Des études basées sur des comparaisons des réponses à une enquête avec des données administratives, comme les renseignements figurant sur les déclarations de revenus utilisées dans la présente étude, sont relativement rares puisque l'accès aux données administratives est très limité et que le coût du couplage de données d'enquête avec des données administratives est élevé. Cette évaluation de l'erreur non due à l'échantillonnage pour le revenu provenant des salaires et traitements a été rendue possible à la suite d'un accord entre le Bureau of the Census et le Internal Revenue Service (IRS) qui permet au Bureau of the Census de coupler des répondants à une enquête avec leurs déclarations de revenus afin d'évaluer la qualité des données[3].

---

[1] J.F. Coder, Housing and Household Economic Statistics Division, Bureau of the Census, Washington, DC, U.S.A. 20233.

[2] Dans la présente communication, les termes «erreur de mesure», «erreur non due à l'échantillonnage», «erreur d'enquête» et «erreur de réponse» sont utilisés indifféremment pour désigner l'écart entre les données d'enquête et les données sur les déclarations de revenus. On admet que dans certains contextes ces termes peuvent avoir des définitions légèrement différentes.

[3] Dans le cadre d'un accord conclu avec le Internal Revenue Service (IRS), le Bureau of the Census reçoit chaque année un extrait de toutes les déclarations fédérales de revenus des particuliers. Ces données extraites sont utilisées pour élaborer des estimations de la population des États et des comtés pendant la période postcensitaire et afin d'évaluer la qualité des données recueillies dans le cadre d'enquêtes. Conformément au chapitre 13, la diffusion de renseignements qui permettraient au IRS ou à tout autre organisme ou personne d'identifier des répondants particuliers à une enquête est interdite. Le couplage des répondants à l'enquête avec les renseignements qui figurent sur les déclarations de revenus de ces personnes est effectué au Bureau of the Census et les fichiers couplés ainsi obtenus sont conservés dans des secteurs protégés dans les locaux occupés par le Bureau of the Census.

Les évaluations antérieures de la qualité des données sur le revenu recueillies dans des enquêtes-ménages montrent généralement des biais par défaut dans les estimations d'enquête des montants du revenu quand ces estimations sont comparées à des sources indépendantes comme les National Income and Product Accounts (comptes nationaux du revenu et des produits) (NIPA). Les comparaisons disponibles pour le supplément sur le revenu de la CPS de mars[4] montrent qu'au cours des dix dernières années, les estimations d'enquête du revenu global provenant des salaires et traitements faisaient l'objet d'un biais par défaut allant de 1 à 3 pour cent (U.S. Census 1991). Des comparaisons semblables, pour l'enquête SIPP, pour l'année civile 1984 ont montré l'existence d'un biais par défaut pouvant aller de 6 à 7 pour cent (Vaughan 1989).

Il existe des preuves qui montrent que ces biais par défaut au niveau global sont symptomatiques d'une structure sous-jacente et complexe de l'erreur. Des recherches récentes (par exemple Brownstone 1992; Coder 1990; Scholz 1990 et Lillard et coll. 1986) ont montré que l'erreur de mesure liée au revenu provenant des salaires et traitements n'est pas distribuée aléatoirement. À l'aide de la PSID, Brownstone a démontré que l'erreur de mesure est corrélée avec les gains «véritables». Coder trouve aussi cette corrélation et un biais par défaut important dans les mesures de l'inégalité pour la CPS de mars. Dans son étude portant sur le crédit d'impôt relatif au revenu gagné Scholz a observé, à l'aide du panel de 1984 de l'enquête SIPP, des sous-estimations des familles avec revenus provenant des salaires et traitements de $50 000 ou plus. Finalement, dans une étude des procédures employées pour faire des imputations dans les cas de réponses manquantes pour le montant des salaires et traitements lors de la CPS de mars, Lillard et coll. ont allégué qu'il y avait des biais par défaut importants dans l'attribution des montants salariaux aux non-répondants.

2. COLLECTE DES DONNÉES SUR LES SALAIRES ET TRAITEMENTS POUR L'ENQUÊTE SIPP

Cette évaluation de l'erreur de mesure dans le revenu provenant des salaires et traitements est basée sur la somme des montants des salaires et traitements relevés pour chaque mois de l'année civile 1990. Puisque le plan d'enquête de l'enquête SIPP utilisait une période de référence de quatre mois avec des interviews qui commençaient en février 1990 pour le premier groupe de renouvellement, afin d'obtenir des montants pour l'année civile il a fallu, selon le groupe de renouvellement, regrouper des données provenant de trois ou quatre interviews.

Les personnes faisant partie de l'univers des salariés comprenaient (1) les employés des sociétés ou des entreprises du secteur privé, (2) les employés des organismes à but non lucratif du secteur privé, (3) les employés des administrations fédérale, d'États ou locales, (4) les membres des Forces armées (militaires) et (5) les travailleurs autonomes propriétaires d'entreprises constituées en société.

Le questionnaire de l'enquête SIPP renferme deux sections identiques portant sur l'emploi des salariés (à l'exclusion des travailleurs autonomes propriétaires d'entreprises constituées en société) afin que des détails sur deux emplois salariés différents puissent être relevés pour la période de référence de quatre mois. Ces sections commencent par des questions sur les renseignements nécessaires afin d'attribuer les classifications types de profession et d'industrie. Viennent ensuite des questions pour déterminer la période pendant laquelle l'emploi a été occupé, les heures travaillées, les motifs du départ de l'emploi (si l'emploi n'a pas été occupé pendant toute la période de quatre mois), le mode de rémunération (p. ex., taux de rémunération horaire, salaire annuel) et l'affiliation syndicale.

Les sections sur l'emploi pour les salariés se terminent par des questions portant sur le montant du revenu provenant des salaires et traitements reçu chaque mois. Le concept de revenu utilisé est celui de la rémunération (salaires et traitements) «brute, avant retenues». On inclut dans le revenu provenant des salaires et traitements le salaire versé pour les heures normales et les gains sous forme de pourboires, de primes, de commissions et de primes d'heures supplémentaires. Cette définition exclut les avantages sociaux ou la rémunération en nature,

---

[4] Le supplément sur le revenu de la Current Population Survey (enquête sur la population actuelle) (CPS) de mars recueille des données sur le revenu et les antécédents professionnels auprès d'un échantillon d'environ 60 000 ménages. Dans cette enquête, qui est réalisée chaque année, on pose des questions détaillées sur les sources et les montants du revenu reçu au cours de l'année civile précédente. Cette enquête a été la principale source de données sur le revenu et la pauvreté pour les États-Unis depuis 1947.

telle que les repas, le logement, l'utilisation d'automobiles, etc. La définition du revenu avant retenues exige que le montant déclaré n'exclut pas les montants comme l'impôt sur le revenu et l'impôt sur la paye, les cotisations des employés à des régimes d'assurance-maladie et de pension, les régimes d'épargne, les cotisations syndicales, etc. D'après les principes de comptabilité, les montants sont inscrits au cours du mois où ils sont reçus et non au cours du mois où ils sont gagnés (le mois pendant lequel le travail est effectué).

Le revenu touché par les travailleurs autonomes propriétaires d'entreprises constituées en société est aussi compté comme revenu provenant des salaires et traitements bien que ce renseignement ne soit pas recueilli dans la partie du questionnaire décrite ci-dessus. Le concept utilisé pour recueillir les montants du revenu dans le cas des travailleurs autonomes propriétaires d'entreprises constituées en société est beaucoup moins bien défini que celui qui s'applique dans le cas des travailleurs salariés plus typiques. Pour le travailleur autonome, la définition de revenu mensuel comprend tout «salaire» régulier ou toute autre somme d'argent reçue de l'entreprise.

Les montants du revenu provenant des salaires et traitements sont imputés ou attribués quand le répondant n'a pas fourni de réponse. Le processus d'imputation est basé sur une technique standard du «hot deck». À l'aide de cette technique, on «apparie» un non-répondant à un répondant qui possède des caractéristiques «semblables». Le montant des salaires et traitements du répondant est alors attribué au non-répondant. Les caractéristiques utilisées au cours du processus d'appariement comprennent des variables telles que l'âge, la race, le sexe, le niveau de scolarité, la profession, les heures travaillées, les semaines travaillées, le lieu de résidence, etc.

3. COUPLAGE DES RENSEIGNEMENTS SUR LES SALAIRES FOURNIS POUR L'ENQUÊTE SIPP ET DANS LES DÉCLARATIONS DE REVENUS

Cette évaluation de la qualité des données sur les salaires et traitements recueillies dans le cadre de l'enquête SIPP a été rendue possible au moyen d'un couplage entre les données recueillies au cours de l'enquête et les données qui figurent sur les déclarations fédérales de revenus des particuliers. Les renseignements tirés des déclarations de revenus ont été fournis au Bureau of the Census par le IRS explicitement pour améliorer les estimations de la population dans les régions et pour effectuer des évaluations de la qualité des données.

Les données de l'equête SIPP et des déclarations de revenus ont été couplées à l'aide des numéros de sécurité sociale ((«social security numbers» (SSN)) recueillis au cours de l'enquête et mentionnés sur chaque déclaration de revenus. Les SSN recueillis au cours de l'enquête SIPP ont été confirmés par la Social Security Administration (SSA) afin d'en vérifier l'exactitude avant le couplage avec les déclarations de revenus[5].

Après la validation, on a tenté d'effectuer un couplage entre les 51 500 SSN de l'enquête et les approximativement 113 millions de déclarations de revenus produites pour 1990. Toute déclaration de revenus sur laquelle figurait un SSN correspondant à une personne dans l'échantillon de l'enquête SIPP a été extraite. Ce processus a donné un fichier de dossiers extraits renfermant environ 31 000 déclarations de revenus (les déclarations en double ont été extraites si les SSN tant de l'époux que de l'épouse correspondaient à une déclaration de revenus conjointe produite par le même couple marié). Les données provenant de ce fichier de déclarations de revenus appariées ont alors été fusionnées avec les données d'enquête pour créer un fichier couplé ou d'«appariement exact» qui a constitué la base des comparaisons présentées dans cette communication.

[5] La validation des SSN est une opération qui s'est faite en deux parties. Dans le cas des personnes qui ont déclaré un SSN pour l'enquête, on a confirmé ces SSN en les appariant avec les dossiers administratifs de la SSA. Dans cette opération, on a utilisé le nom, la date de naissance, la race et le sexe pour effectuer la validation. Dans le cas des personnes qui n'ont pas déclaré de SSN pour l'enquête, on a effectué une recherche dans les systèmes de dossiers de la SSA d'après le nom, la date de naissance, la race et le sexe. Aucune recherche n'a été effectuée pour les personnes qui avaient explicitement refusé de fournir un SSN à l'intervieweur de l'enquête SIPP.

4. L'UNIVERS ET LES LIMITATIONS DE L'ÉTUDE

4.1 L'univers de l'étude

L'univers utilisé pour cette étude n'est pas représentatif de la population dans son ensemble mais d'un sous-ensemble très particulier de l'échantillon de l'enquête SIPP. Ce sous-ensemble est composé de couples mariés[6] ayant des SSN validés tant pour l'époux que pour l'épouse, appariés à une déclaration de revenus conjointe de personnes mariées et dont le montant déclaré pour le revenu provenant des salaires et traitements n'est pas nul soit pour l'enquête SIPP, soit pour la déclaration de revenus. En tout, 5 703 couples mariés satisfaisaient à ces restrictions dans le fichier créé pour cette étude. Cela représente environ 62 pour cent du total de 9 267 unités époux-épouse présentes dans l'échantillon de 1990 de l'enquête SIPP en mars 1991. Du total de 9 267 unités, 864, ou environ 9 pour cent, ont été exclues parce qu'un des SSN ou les deux ne pouvaient être confirmés. Des 8 403 cas restants, 6 548 ont pu être appariés avec des déclarations de revenus. Aucun appariement n'a pu être effectué pour 902 cas, bien que des SSN validés étaient disponibles pour les deux conjoints. La majorité de ces cas correspondait à des couples qui n'ont pas produit de déclaration de revenus et pour lesquels un appariement était donc impossible. Les situations dans lesquelles le SSN d'un conjoint était apparié avec une déclaration de revenus qui n'était pas une déclaration conjointe ou qui ne renfermait pas de SSN «secondaire» (le SSN de l'autre conjoint) correspondaient à 857 cas additionnels éliminés de la présente analyse[7]. Finalement, 354 autres cas ont été exclus, surtout parce que seulement un des SSN des conjoints avait pu être apparié avec une déclaration conjointe renfermant à la fois un SSN principal et un SSN secondaire.

L'univers des personnes non mariées dans l'échantillon était exclu de l'étude. Ce groupe n'était pas inclus dans cette analyse dû aux problèmes liés aux erreurs de déclaration peuvent différer considérablement de ceux des couples mariés dont le revenu provenant des salaires et traitements représente la somme des montants pour l'épouse et pour l'époux. Une étude de l'univers des personnes non mariées sera entreprise séparément.

4.2 Concepts différents pour les salaires et traitements

Les concepts de revenu provenant des salaires et traitements pour les déclarations de revenus et pour l'enquête SIPP diffèrent sur plusieurs points. Les montants des salaires et traitements déclarés sur les déclarations de revenus peuvent ne pas refléter le montant réel gagné. Cela soulève plusieurs inquiétudes importantes ici. Tout d'abord, le concept pour l'enquête SIPP, du moins théoriquement, est défini comme le montant brut gagné. Cela inclurait les montants gagnés mais reportés de l'impôt courant. Toutefois, le montant indiqué sur les déclarations de revenus exclut ces montants de gains différés[8]. Deuxièmement, aux fins de l'impôt, le revenu provenant des salaires et traitements peut inclure certains genres de paiements en nature, dont aucun n'est inclus dans la définition utilisée pour l'enquête SIPP. Bien que l'inclusion des montants des paiements en nature dans les déclarations de revenus amène une surdéclaration de l'erreur «véritable», le nombre de ces cas est faible. Par contre, la proportion élevée de personnes qui participent maintenant à des régimes de revenu différé doit mener à une sous-déclaration du niveau d'erreur mesuré dans la présente étude à moins que les montants des gains différés ne soient aussi exclus, dans une large mesure, des montants déclarés pour l'enquête SIPP.

---

[6] Les couples mariés ont été définis plus précisément comme les couples qui étaient mariés en mars 1991. Puisque la validation des SSN était limitée aux personnes interviewées lors du premier cycle du panel, l'univers des couples mariés était aussi limité aux couples ou tant l'époux que l'épouse avaient été interviewés lors de l'interview initiale.

[7] Dans la majorité des cas, les déclarations de revenus conjointes de personnes mariées renferment les numéros de sécurité sociale (SSN) des deux conjoints et le SSN du déclarant principal est confirmé (les SSN sont classés comme le SSN principal et le SSN secondaire pour ces déclarations). Dans certains cas, toutefois, le SSN secondaire manque. Ces cas ont été éliminés même si un appariement avait pu être réalisé au moyen du SSN principal.

[8] Environ 20 pour cent des cas ont déclaré participer à un régime de gains différés. Le montant moyen différé était d'environ $2 800.

4.3 Base de sondage, sous-dénombrement de la population et pour l'enquête

La présente analyse est limitée aux erreurs attribuables aux problèmes de déclaration pour l'enquête. Les erreurs reliées (1) aux problèmes de la base de sondage, (2) au sous-dénombrement du recensement de la population et (3) aux sous-dénombrements différentiels des sous-groupes de la population dans le cadre de l'enquête n'ont pas été étudiées. Puisque nous savons qu'il existe des problèmes dans ces domaines et dans les méthodes de pondération utilisées pour corriger ces problèmes, la description de l'erreur totale n'est pas complète sans évaluation de la contribution de ces problèmes.

5. ÉVALUATION DES ERREURS NON DUES À L'ÉCHANTILLONNAGE

Pour examiner les erreurs non dues à l'échantillonnage dans le cas de l'enquête SIPP, j'ai choisi d'étudier une gamme étendue mais fondamentale de mesures afin de fournir une description du problème des erreurs. Parce qu'on peut imputer une partie importante du niveau total de l'erreur aux problèmes de données manquantes, la majorité des renseignements est présentée séparément selon le genre de déclaration. Deux groupes ont été définis en fonction du genre de déclaration. Le premier, désigné par l'expression «complet», comprend seulement les cas avec des données complètes pour les douze mois de l'année civile 1990 (c.-à-d. les cas pour lesquels il n'y a pas eu de mois de non-interview et pas de non-réponse aux questions sur les montants des salaires et traitements). Ces cas représentent environ 79 pour cent du total. Les cas qui restent (21 pour cent), font partie du deuxième groupe désigné par l'expression «incomplet». Ce dernier groupe est composé de cas avec un mois ou plus de non-interview pendant l'année ou avec non-réponse à une ou plusieurs questions portant sur les montants mensuels du revenu provenant des salaires et traitements.

Une restriction a été imposée sur le montant des salaires et traitements afin de réduire l'effet des «valeurs aberrantes». Au cours des premières étapes de l'étude, on a trouvé qu'un petit nombre de cas pour lesquels les montants des salaires et traitements indiqués sur les déclarations de revenus étaient très élevés (alors que des montants beaucoup plus faibles ont été déclarés à l'enquête) semblaient causer une distorsion excessive de la description des erreurs. C'est pourquoi j'ai limité tous les montants de revenu provenant des salaires et traitements afin qu'ils ne dépassent pas le montant le plus élevé déclaré lors de l'enquête. Ce montant était d'environ $700 000[9]. Toutes les mesures qui figurent dans cette communication reflètent donc cet ajustement. Cela n'a eu une incidence que sur un petit nombre des montants figurant sur les déclarations de revenus.

5.1 Différences dans les mesures sommaires fondamentales

5.1.1 Univers total

Les données dans le tableau A résument les biais par défaut dans la distribution des salaires et traitements pour l'enquête SIPP par rapport à celle tirée des déclarations de revenus. On remarque une sous-estimation de 7 pour cent de la moyenne et de 5 pour cent de la médiane. Il semble aussi y avoir un léger biais par défaut d'environ 1 pour cent dans le nombre des salariés. L'effet combiné des sous-estimations, pour l'enquête SIPP, des salariés et des montants des salaires et traitements touchés amène une sous-estimation d'environ 8 pour cent du montant global des salaires et traitements.

Le biais par défaut dans la variance de la distribution des salaires et traitements établie d'après l'enquête SIPP semble être plus grave. La variance de la distribution des salaires et traitements basée sur l'enquête SIPP n'est que d'environ 47 pour cent de la variance de la distribution comparable basée sur les déclarations de revenus[10]. Cet écart considérable dans les variances découle surtout de différences à la limite supérieure de la distribution, même après que l'on ait appliqué la limite de $700 000. Quand on a examiné la sensibilité des variances à des montants de salaires et traitements très élevés on a trouvé qu'il se produit une détérioration rapide au-dessus

---

[9] Plutôt que d'éliminer ces «valeurs aberrantes», j'ai choisi de les inclure mais d'en limiter le montant. Les résultats basés sur une stratégie qui élimine ces cas ont été très semblables à ceux qui figurent dans la communication.

[10] Si aucune correction n'est apportée pour les valeurs aberrantes, le rapport entre la variance basée sur la SIPP et la variance basée sur les déclarations de revenus n'est que d'environ .20.

du niveau des $400 000. Pour les cas dont les montants inscrits sur les déclarations de revenus sont inférieurs à $200 000, la variance de l'enquête SIPP est d'environ 90 pour cent de celle pour les déclarations de revenus. Si l'on porte à $400 000 la restriction appliquée à l'univers, cela n'entraîne qu'une légère diminution du rapport qui s'élève alors à environ 86 pour cent. À $700 000, toutefois, le rapport est réduit à 47 pour cent, tel que mentionné dans le tableau A.

Le coefficient de corrélation simple entre les montants de l'enquête SIPP et ceux des déclarations de revenus était de .754. Ce coefficient de corrélation s'est montré très sensible aux salaires et traitements élevés, diminuant rapidement quand les restrictions sur les valeurs aberrantes sont supprimées.

5.1.2 Cas complets

Des biais atténuants sont aussi évidents pour l'univers des cas complets. Pour la moyenne et la médiane, les biais sont un peu plus faibles que ceux relevés pour l'univers total (environ 4 pour cent tant pour la moyenne que pour la médiane). Cependant, le biais par défaut dans le nombre de salariés semble un peu plus élevé et celui pour la variance seulement légèrement plus faible. Pour l'ensemble des salaires, la valeur pour les cas complets était d'environ 7 pour cent inférieure à la valeur pour l'ensemble des salaires figurant sur les déclarations de revenus. Le coefficient de corrélation entre les montants de l'enquête SIPP et ceux des déclarations de revenus était d'environ 8 points (.834) plus élevé que pour l'univers total.

5.1.3 Cas incomplets

Sauf pour un léger biais surmontant dans le nombre de salariés, on trouve pour l'univers des cas incomplets des biais atténuant beaucoup plus élevés que ceux mentionnés pour les cas complets. Les moyennes et médianes pour l'enquête SIPP qui figurent au tableau A sont biaisés, abaissant de 15 et 11 pour cent respectivement. Si l'on passe à la comparaison des variances, on trouve que la situation n'est qu'un peu plus mauvaise que pour les cas complets, la variance des salaires et traitements, basée sur l'enquête SIPP étant environ 44 pour cent de la variance correspondante basée sur les déclarations de revenus.

Une étude des mesures sommaires relatives aux déclarations de revenus pour l'univers des cas complets et pour celui des cas incomplets fournit certaines preuves solides qu'il y a une différence entre les répondants et les non-répondants pour ce qui est des niveaux du revenu provenant des salaires et traitements. Le montant moyen et la variance sont beaucoup plus élevés (une moyenne plus élevée de 18 pour cent et une variance deux fois plus grosse) pour le groupe des cas incomplets composé de cas ayant fait l'objet d'une imputation et de cas d'interviews partielles.

On trouve, associé à ces écarts, le fait qu'un des buts du processus d'imputation semble ne pas avoir été atteint. Ce but est de réduire le biais imputable à la non-réponse à une question. La moyenne des montants des salaires et traitements imputés, dans le cadre de l'enquête, pour les cas incomplets est d'environ 5 pour cent plus élevée que la moyenne pour les cas complets, mais la moyenne relative aux déclarations de revenus pour les cas incomplets dépasse de 18 pour cent, comme on l'a mentionné ci-dessus, la moyenne relative aux déclarations de revenus pour les cas complets. Par conséquent, moins d'un tiers de l'écart de 18 pour cent, tel que mesuré par la moyenne, était comblé par les procédures d'imputation.

5.2 Effets de l'erreur non due à l'échantillonnage sur la distribution

La comparaison des mesures sommaires montre clairement que l'erreur de mesure n'est pas répartie uniformément. Les distances entre les moyennes et les médianes selon l'enquête SIPP et les déclarations de revenus diffèrent et des écarts considérables dans les variances de leurs distributions sont évidents. Afin d'évaluer l'effet net de ces erreurs sur la distribution des salaires, j'ai classé les cas séparément selon la taille des montants des salaires et traitements d'après l'enquête SIPP et les déclarations de revenus et étudié diverses caractéristiques des déciles des salaires obtenus à la suite de ce classement. Le tableau B donne des comparaisons des «limites» des déciles et des parts des déciles (deux mesures de l'inégalité utilisées couramment, l'indice de concentration de Gini et la variance du logarithme naturel (VarLn) des salaires sont aussi présentées).

Tableau A: Comparaisons sommaires des données sur les salaires et traitements provenant de l'enquête SIPP et des déclarations de revenus correspondantes pour 1990.

(Déclarations de revenus conjointes de personnes mariées appariées avec le revenu des salaires et traitements provenant d'une source précisée)

TOUS LES CAS

Mesure	Déclarations de revenus	SIPP	Écart en pourcentage
Nombre de cas	5 558	5 540	-0.9
Montant moyen	$43 630	$40 460	-7.3
Montant médian	$37 640	$35 760	-5.0
Variance	1.933e9	9.016e8	.466
Valeur maximum	+$2 000 000	+$700 000	.350
Salaires globaux (en milliers)	$243 800	$224 140	-8.1

CAS COMPLETS

Mesure	Déclarations de revenue	SIPP	Écart en pourcentage
Nombre de cas	4 409	4 326	-1.9
Montant moyen	$42 060	$40 020	-4.3
Montant médian	$37 190	$35 810	-3.7
Variance	1.446e9	7.874e8	.545
Valeur maximum	+$800 000	+$400 000	.500
Salaires globaux (en milliers)	$185 400	$173 100	-6.6

CAS INCOMPLETS

Mesure	Déclarations de revenus	SIPP	Écart en pourcentage
Nombre de cas	1 179	1 214	3.0
Montant moyen	$49 510	$42 010	-15.2
Montant médian	$39 720	$35 490	-10.6
Variance	3.342e9	1.472e9	.440
Valeur maximum	+$2 000 000	+$700 000	.350
Salaires globaux (en milliers)	$58 400	$51 000	-12.6

5.2.1 Limites des déciles. Globalement, l'enquête SIPP semble surestimer légèrement les niveaux du revenu provenant des salaires et traitements près de la limite inférieure de la distribution, mais à les surestimer aux autres points. Pour constater ce fait, il suffit d'examiner les limites présentées au tableau B. Ici, les limites selon l'enquête SIPP pour les deux premiers déciles sont supérieures aux limites selon les déclarations de revenus puis inférieures pour tous les déciles à compter du troisième. La taille des sous-estimations des limites après celles du deuxième décile est stable, chacune des sous-estimations étant de 4 à 5 pour cent.

Le cycle composé tout d'abord de surestimations puis de sous-estimations décrit ci-dessus pour l'univers total se répète aussi pour le groupe composé exclusivement de cas complets. Pour ce dernier univers, les sous-estimations dans le troisième décile et les déciles supérieurs étaient comprises entre 2 et 5 pour cent.

Les données sur les limites des déciles pour les cas incomplets, par contre, ne suivent pas le cycle de surestimations initiales suivies de sous-estimations. Pour ce groupe, on trouve une sous-estimation pour tous

les déciles. La sous-estimation est la plus faible pour les premiers déciles augmentant d'environ 4 pour cent pour le premier décile jusqu'à 11 pour cent à la limite du dernier décile.

5.2.2 Parts des déciles

Sans exception, l'enquête SIPP surestime la part des salaires et traitements reçus dans tous les déciles sauf le plus élevé. Pour l'enquête SIPP, on trouve une sous-estimation importante dans le dernier décile, ce qui reflète le déséquilibre cumulatif produit par les surestimations dans tous les groupes inférieurs. Les biais atténuants dans la part reçue par le décile supérieur sont importants, environ 11 pour cent globalement, 9 pour cent pour les cas complets et 15 pour cent pour l'univers des cas incomplets.

Tableau B: Seuils et parts des revenus pour les déciles du revenu provenant des salaires et traitements pour l'enquête SIPP et les déclarations de revenus appariées pour 1990.

(Déclarations de revenus conjointes de personnes mariées appariées avec le revenu des salaires ou traitements provenant d'une source précisée)

ÉTAT POUR LA DÉCLARATION

	Univers total		Cas complets		Cas incomplets	
Décile	Déclarations de revenus	SIPP	Déclarations de revenus	SIPP	Déclarations de revenus	SIPP
1	$9 499	$10 462	$8 859	$10 240	$11 861	$11 275
2	18 684	18 902	17 847	18 672	21 287	19 593
3	25 542	24 975	25 040	24 968	27 262	24 986
4	31 888	30 942	31 452	30 242	33 356	30 022
5	37 637	35 759	37 185	35 805	39 716	35 488
6	43 617	41 488	43 197	41 778	45 097	40 286
7	50 260	48 171	49 893	48 673	51 978	46 523
8	59 552	56 845	59 005	57 416	62 550	55 676
9	75 510	71 585	74 241	71 495	80 358	71 726
Parts des déciles						
Total	100.0	100.0	100.0	100.0	100.0	100.0
1	1.0	1.4	1.0	1.3	1.4	1.6
2	3.2	3.7	3.2	3.7	3.4	3.8
3	5.1	5.4	5.1	5.5	4.9	5.3
4	6.6	6.8	6.8	6.9	6.2	6.6
5	7.9	8.1	8.1	8.2	7.4	7.9
6	9.3	9.5	9.5	9.7	8.5	9.2
7	10.7	11.0	11.0	11.2	9.8	10.4
8	12.5	12.9	12.9	13.1	11.5	12.2
9	15.2	15.7	15.6	15.9	14.1	15.0
10	28.5	25.5	26.8	24.5	32.8	28.0
Gini	.390	.358	.381	.353	.417	.373
VarLn	1.180	.787	1.290	.822	.780	.651

5.2.3 Inégalité des salaires et traitements

Les estimations de l'enquête pour deux mesures de l'inégalité des salaires et traitements utilisées couramment, l'indice de concentration de Gini et la variance du logarithme naturel des salaires et traitements sont toutes les deux biaisées de façon atténuée. Globalement, les biais sont d'environ 8 pour cent pour l'indice de concentration de Gini et de 33 pour cent pour la variance du logarithme. Cette sous-estimation de 33 pour cent de la variance

du logarithme est considérablement inférieure à celle déjà mentionnée pour le calcul de la variance ordinaire et reflète les effets de «compression» de la transformation.

5.3 Mesures de l'erreur de classification

On trouve au tableau C des distributions de l'erreur exprimée sous forme de distance entre le décile d'après l'enquête SIPP et le décile d'après les déclarations de revenus. Le tableau montre aussi la proportion de cas dans chaque groupe de distance selon l'état pour la déclaration. Plus de la moitié (53 pour cent) de tous les cas ont été classés dans le décile approprié selon leur montant déclaré pour l'enquête SIPP. Ce taux d'exactitude est inférieur (40 pour cent) pour le groupe des cas incomplets où les classifications dépendent du montant attribué par les procédures d'imputation de l'enquête SIPP. Un autre groupe de cas de l'enquête SIPP (33 pour cent) ont été classés soit dans le décile au-dessous, soit dans le décile au-dessus de leur position selon les déclarations de revenus. La combinaison de ces deux groupes donne un total de 86 pour cent de tous les cas classés à un décile près du décile dont ils font partie d'après les déclarations de revenus.

L'exactitude de la classification est liée au statut pour la déclaration comme on le voit dans la partie droite du tableau C. La proportion du total des cas dans chaque catégorie de distance de déciles attribuable aux cas incomplets augmente à mesure que la distance entre les déciles de l'enquête SIPP et des déclarations de revenus augmente.

Tableau C: Sommaire des écarts dans la classification des déciles pour le revenu provenant des salaires et traitements selon l'enquête SIPP et les déclarations de revenus: 1990.

(Déclarations de revenus conjointes de personnes mariées appariées avec le revenu des salaires ou traitements provenant des deux sources)

	UNIVERS DES DÉCLARATIONS					
	Total	Complets	Incomplets	Total	Complets	Incomplets
Distance en déciles entre l'enquête SIPP et les déclarations de revenus						
Nombre	5 425	4 248	1 177	5 425	4 248	1 177
Pourcentage	100.0	100.0	100.0	100.0	78.3	21.7
Même décile	53.3	56.7	40.4	100.0	83.6	16.4
1	33.0	32.6	34.1	100.0	77.3	22.7
2	8.2	7.3	12.0	100.0	68.9	31.1
3	2.5	1.7	5.4	100.0	52.9	47.1
4	1.5	0.8	3.9	100.0	39.3	60.7
5	0.7	0.5	2.2	100.0	50.0	50.0
6	0.4	0.2	0.8	100.0	42.9	57.1
7	0.3	0.2	0.6	100.0	52.9	41.1
8	0.2	0.1	0.3	100.0	33.0	37.0
9	(z)	0.0	0.2	100.0	0.0	100.0

(z) Moins de .05 pour cent.

5.4 Décomposition de l'erreur dans les salaires globaux

On constate dans le tableau A que le montant global du revenu provenant des salaires et traitements basé sur l'enquête SIPP était d'environ 8 pour cent inférieur au montant global provenant des déclarations de revenus. Les cas incomplets contribuent une part disproportionnellement élevée du biais atténuant net. Bien qu'ils ne constituent qu'environ 21 pour cent des cas, on peut leur attribuer 38 pour cent de la sous-estimation globale pour l'enquête.

Même si l'effet net de l'erreur de déclaration est manifestement un biais atténuant, pour de nombreux cas de l'enquête SIPP, les montants des salaires et traitements dépassent les montants des déclarations de revenus correspondantes. Tant pour l'univers des cas complets que pour l'univers des cas incomplets, dans environ 38 pour cent des cas, le montant selon l'enquête SIPP est supérieur à celui selon les déclarations de revenus. La surdéclaration nette s'est élevée à environ $16.3 millions. Le niveau moyen de surdéclaration était d'environ $5 800 pour les cas complets mais de près de $14 000 pour l'univers des cas incomplets. Les cas avec des montants inférieurs pour l'enquête SIPP dépassent ceux avec des estimations supérieures tant pour le nombre que pour l'importance de l'erreur (écart entre le montant pour l'enquête SIPP et pour les déclarations de revenus). Pour les 62 pour cent des cas avec des montants inférieurs pour l'enquête SIPP, l'erreur moyenne était de $8 000 pour les cas complets et de $18 000 pour les cas incomplets.

Tableau D: Mesures choisies de l'erreur dans les données sur le revenu provenant des salaires et traitements fournies par le panel de 1990 de l'enquête SIPP: 1990.

(Déclarations de revenus conjointes de personnes mariées appariées avec le revenu des salaires ou traitements provenant des deux sources)

	ÉTAT POUR LA DÉCLARATION		
Mesure de l'erreur	Total	Complets	Incomplets
Écart moyen	$-3 920	$-2 980	$-7 310
Écart médian	$-1 230	$-1 020	$-2 440
Écart relatif moyen (%)	30.0	34.2	15.1
Écart relatif médian (%)	-3.8	-3.4	-7.3
Écart absolu moyen	$8 970	$6 980	$16 100
Écart absolu médian	$3 660	$3 240	$6 090
Somme des carrés des écarts (SCE)	4.664e12	2.095e12	2.569e12
Indice d'incohérence[1]	.442	.320	.653

[1] défini comme (((SIPP-déclaration de revenus)^2/n)/var déclaration de revenus)

Tableau E: Distribution cumulative des écarts absolus relatifs entre le revenu provenant des salaires et traitements selon l'enquête SIPP et les déclarations de revenus: 1990.

(Déclarations de revenus conjointes de personnes mariées appariées avec le revenu des salaires ou traitements provenant des deux sources)

	ÉTAT POUR LA DÉCLARATION		
Erreur de pourcentage	Total	Complets	Incomplets
Total	100.0	100.0	100.0
Moins de 1 pour cent	3.1	3.4	1.9
Moins de 2 pour cent	8.4	9.3	5.2
Moins de 3 pour cent	14.1	15.7	8.5
Moins de 4 pour cent	19.8	33.1	11.7
Moins de 5 pour cent	25.1	27.8	18.5
Moins de 10 pour cent	48.9	53.1	33.6
Moins de 15 pour cent	62.1	66.6	45.7
Moins de 20 pour cent	71.8	76.2	55.8
Plus de 20 pour cent	28.2	23.8	44.2

Les erreurs relatives au montant des salaires et traitements touchés servent à réduire plutôt qu'à accroître le biais atténuant dans l'agrégat de l'enquête. Un biais surmontant est créé parce que l'agrégat attribuable aux faux positifs pour l'enquête est plus de deux fois plus grand que le montant agrégé selon les déclarations de revenus pour les cas de l'enquête qui ne déclarent aucun montant pour le salaire.

5.5 Étude des erreurs elles-mêmes

Jusqu'ici, la communication a surtout porté sur les effets nets de l'erreur de mesure. L'importance et la distribution des erreurs elles-mêmes sont aussi importantes pour donner un aperçu complet du problème des erreurs. La présente section est donc consacrée à établir le profil des erreurs.

5.5.1 Mesures générales

Quand on tente d'établir le profil des caractéristiques des erreurs, il y a de nombreuses statistiques qui pourraient être étudiées. Un certain nombre d'entre elles sont incluses au tableau D et chacune semble fournir une perspective quelque peu différente. Par exemple, l'écart moyen entre le montant des salaires et traitements selon l'enquête SIPP et les déclarations de revenus était de -$3 920. Cette mesure reflète l'effet net tant de la surdéclaration que de la sous-déclaration. Si l'on passe à une mesure basée sur la valeur absolue de l'écart, ce qui élimine l'effet compensateur de la surdéclaration et de la sous-déclaration, l'erreur moyenne semblait être plus de deux fois plus grande, s'établissant à $8 970. Les mesures de l'erreur pour l'univers des cas incomplets montrent que le problème des erreurs est considérablement plus important pour ces cas.

Tableau F: Parts des écarts absolus relatifs par décile pour le revenu provenant des salaires et traitements selon l'enquête SIPP: 1990.

(Déclarations de revenus conjointes de personnes mariées appariées avec le revenu des salaires ou traitements provenant des deux sources)

	ÉTAT POUR LA DÉCLARATION		
Écart en déciles	Total	Complets	Incomplets
1	0.2	0.1	0.3
2	0.5	0.4	1.0
3	0.9	0.8	0.7
4	1.3	1.1	2.5
5	1.8	1.6	3.4
6	2.5	2.1	4.5
7	3.3	2.8	5.9
8	4.4	3.7	8.0
9	6.7	5.4	12.0
10	78.4	90.0	60.7

5.5.2 Distribution de la taille des erreurs

La distribution des erreurs selon leur taille (dans le présent cas, la taille absolue relative) constitue une mesure simple et très descriptive de l'erreur. On peut trouver cette mesure au tableau E. Une étude de ce tableau révèle que pour environ 25 pour cent des cas, le montant de l'enquête SIPP était à moins de ±5 pour cent du montant figurant sur les déclarations de revenus et pour environ la moitié des cas, les montants selon l'enquête et les déclarations de revenus différaient de moins de 10 pour cent. L'accord que l'on retrouve pour les cas complets est considérablement plus élevé que celui relevé pour les cas incomplets.

5.5.3 Parts des erreurs

Étant donné une distribution des erreurs, il semble justifié et révélateur d'avoir une mesure de la concentration de telles erreurs. Les concentrations des erreurs ont été calculées pour la différence absolue relative entre les montants de l'enquête SIPP et des déclarations de revenus. Les résultats sont présentés au tableau F.

Pour l'univers total, on trouve un niveau de concentration élevé pour cette mesure, près de 80 pour cent de l'erreur globale attribuable à ces cas se trouvant dans les 10 pour cent supérieurs à la distribution des erreurs. Les erreurs sont encore plus concentrées dans les 10 pour cent supérieurs des cas complets où l'on retrouve 90 pour cent de l'erreur globale. Par contre, l'erreur globale pour l'univers des cas incomplets est considérablement moins concentrée dans la partie supérieure de la distribution.

6. CONCLUSIONS ET RECOMMANDATIONS

Les observations présentées dans cette communication sont basées sur un couplage entre les données sur les salaires et traitements recueillies à l'aide du panel de 1990 de l'enquête SIPP et les données déclarées par les répondants à l'enquête SIPP sur les déclarations fédérales de revenus de ces personnes pour 1990. Cette analyse des données couplées a montré que les données sur les salaires et traitements pour l'enquête SIPP sont biaisées par défaut quand on les compare aux renseignements qui figurent sur les déclarations de revenus. En fonction du montant total du revenu provenant des salaires et traitements reçu par les couples mariés, ce biais correspond à une sous-estimation nette d'environ 8 pour cent pour l'enquête SIPP.

Cette statistique simple ne révèle toutefois pas la nature complexe du problème des erreurs de mesure. Cette évaluation a fait ressortir quatre dimensions principales du problème des erreurs. La première dimension est caractérisée par les erreurs «mineures», c'est-à-dire, celles qui sont attribuables à une forme d'erreur de réponse simple où le répondant déclare un montant qui est légèrement plus faible ou plus élevé que le montant «véritable». La deuxième dimension se rapporte au problème plus sérieux caractérisé par les erreurs «majeures». Ces dernières sont celles qui entraînent une erreur de classification importante dans la position du répondant pour ce qui est de la distribution des salaires et traitements. De telles erreurs pourraient entraîner un déplacement de deux déciles ou plus dans la position de tels cas relativement à la distribution. La troisième dimension de l'erreur est celle qui porte sur les données manquantes et sur le système utilisé par la suite afin d'imputer des valeurs pour les cas de non-réponse à l'enquête. La dimension finale de l'erreur est celle qui mène à une représentation inadéquate de la queue supérieure de la distribution des revenus et dans l'effet que ce problème a sur les mesures de la variance et de l'inégalité des salaires.

La majorité des réponses à l'enquête SIPP sont entachées des genres d'erreurs que l'on peut considérer être des erreurs «mineures». En fait, le biais atténuant dans le revenu médian provenant des salaires et traitements pour l'enquête SIPP n'est que de 4 pour cent pour les cas sans problème de non-réponse. La comparaison des valeurs sur les déclarations de revenus et déclarées pour l'enquête montre que dans plus de la moitié des cas, le montant déclaré pour l'enquête et dans les déclarations de revenus différait de moins de 10 pour cent (et, pour les deux tiers des cas, de moins de 15 pour cent). Une mesure de l'erreur de classification basée sur la position selon la distribution montre que le décile d'après l'enquête et le décile d'après les déclarations de revenus étaient identiques pour 57 pour cent des cas et que près de 90 pour cent de tous les cas se trouvaient à ±1 décile de celui qui est calculé à l'aide du montant figurant sur les déclarations de revenus.

La mesure de la concentration des erreurs montre qu'une bonne partie de l'erreur de mesure globale peut être attribuée à un petit nombre de cas. Généralement, il s'agit de cas qui montrent des problèmes dus à des erreurs «majeures» qui entraînent une erreur de classification importante par rapport à leur position dans la distribution des salaires. On a trouvé que les cas dans le décile des erreurs le plus élevé contribuaient environ 90 pour cent du montant global des erreurs. Bien que cette analyse ne tente pas d'étudier les causes de l'erreur de mesure, il faudrait affecter des ressources additionnelles dans ce but. Le Bureau of the Census effectue actuellement des recherches dans le cadre du programme de conception de mesures de remplacement (Moore et coll.). On espère que ces travaux fourniront certains aperçus des problèmes signalés ici. De plus, il serait utile d'étudier les questionnaires dans les cas pour lesquels on trouve les plus grandes erreurs. Une telle étude constituerait un moyen économique de trouver ce qui est à l'origine de ces erreurs.

Une des constatations les plus importantes qui figurent dans cette communication est celle qui porte sur le niveau d'erreur associé aux données manquantes et à l'imputation effectuée par la suite. L'imputation faite pour les données manquantes est un élément qui contribue de façon importante au problème global des erreurs dans l'enquête SIPP. Pour environ 21 pour cent de l'univers utilisé pour l'étude, on relevait soit 1) de la non-réponse à l'enquête pour les montants des salaires et traitements, soit 2) certains mois avec état de non-interview totale (situation où il n'y a pas eu de réponse pour un cycle). On peut attribuer à ce groupe près de 40 pour cent de la sous-estimation nette de 8 pour cent du revenu global provenant des salaires et traitements. À l'aide d'autres mesures de l'erreur, comme la somme des carrés des écarts, on trouve que près de 90 pour cent de l'erreur totale peut être attribuée au sous-univers des cas incomplets. D'après ces résultats, il semble évident qu'il faudrait entreprendre une étude du système d'imputation. Les couplages entre les enquêtes et les systèmes de dossiers administratifs fournissent un milieu unique pour élaborer et évaluer des systèmes d'imputation.

Traditionnellement, pour les enquêtes-ménages on a eu une très grande difficulté à fournir des estimations exactes de la queue supérieure des distributions des revenus et l'enquête SIPP ne fait pas exception. La comparaison des variances pour l'enquête SIPP et pour les déclarations de revenus dans le cas de la distribution des revenus provenant des salaires et traitements montre que la distribution de l'enquête SIPP minimise trop la variance des salaires et traitements et cette valeur inférieure est liée surtout aux valeurs supérieures de la distribution. La valeur inférieure de la variance peut être attribuée presque entièrement au fait que l'enquête SIPP ne peut recueillir les données correspondants à la partie la plus élevée de la distribution des salaires et traitements. Pour les niveaux de salaires et traitements inférieurs à $200 000, l'estimation de la variance basée sur l'enquête SIPP est biaisée de façon atténuée de seulement environ 10 pour cent. Pour les montants inférieurs à $700 000, le biais atténuant augmente pour dépasser 50 pour cent et, si l'on tient compte de tous les cas sans restriction, le biais est de 80 pour cent. Puisque le nombre de cas sur lesquels porte la présente étude est relativement faible, une autre étude du biais atténuant de la variance pour l'enquête SIPP est justifiée. À tout le moins, on devrait calculer la variance de la distribution du revenu provenant des salaires et traitements fondée sur le fichier de microdonnées à grande diffusion des Statistics of Income (statistiques du revenu) (SOI) pour 1990 afin d'obtenir une estimation fiable de la variance pour toutes les déclarations de revenus et non seulement pour celles appariées à l'univers de l'enquête SIPP utilisé dans la présente étude.

En plus de la recherche qui se poursuit mentionnée auparavant, il y a plusieurs autres domaines de recherche qui devraient être entrepris. Premièrement, cette évaluation de la qualité des données devrait être reprise, basée sur un couplage semblable entre les données de l'enquête SIPP et les données sur le total des prestations sociales que la Social Security Administration rendra bientôt disponibles. Le couplage entre les données de l'enquête SIPP et ces données permettrait une étude beaucoup plus approfondie des erreurs de déclaration que ce qui peut être effectué à l'aide des déclarations de revenus puisque le couplage serait basé sur une personne plutôt que sur une déclaration de revenus. Ainsi, on pourrait traiter séparément les époux et les épouses plutôt que de les considérer comme une unité. L'utilisation de ce fichier éliminerait aussi le problème d'incompatibilité dû au fait que les montants figurant sur les déclarations de revenus ne comprennent pas les gains différés puisque les montants du total des prestations sociales reflètent les niveaux de gains aux fins du calcul des impôts sur la paye pour la sécurité sociale.

Deuxièmement, il serait utile d'étendre la recherche présentée ici afin de déterminer le niveau d'erreurs observé pour divers sous-groupes de la population. Par exemple, pour quelles classifications de l'âge, des antécédents professionnels, du niveau de scolarité, des catégories de travailleurs, des professions, etc. trouve-t-on les plus grands problèmes d'erreur de mesure?

Troisièmement, il faudrait des recherches sur la pondération relative à la post-stratification basée sur les renseignements figurant sur les déclarations de revenus. Un travail effectué il y a plusieurs années au Bureau of the Census a montré qu'on pouvait réaliser une réduction importante des variances d'échantillonnage et une certaine réduction du biais suite à l'utilisation d'éléments de contrôle par pondération provenant de renseignements figurant sur les déclarations de revenus.

Finalement, dans ce qui pourrait être un projet à un peu plus long terme, on devrait étudier une inconnue importante dans le tableau de l'erreur de mesure, celle qui porte sur les non-interviews dans les ménages. Actuellement, des mécanismes de pondération grossiers sont appliqués sans aucune stratification pour tenir compte de la situation socio-économique.

BIBLIOGRAPHIE

Brownstone, D., et Valletta, R. (1992). Modeling measurement error bias in cross-section and longitudinal wage equations. Discussion présentée au «1992 Bureau of the Census Annual Research Conference».

Coder, J.F. (1990). Exploring nonsampling errors in the wage and salary data from the March current population survey. Discussion présentée à la réunion «1990 Allied Social Sciences Association/Society of Government Economists».

Groves, R.M. (1989). *Survey Errors and Survey Costs*, New York, John Wiley and Sons.

Herriot, R.A., et Spiers, E.F. (1980). Measuring the impact on income statistics of reporting differences between the current population survey and administrative sources. *Studies from Interagency Data Linkages*, Report No. 11, U.S. Department of Health, Education, and Welfare, Social Security Administration, Office of Research and Statistics, SSA Publication Numéro 13-11750, (mars).

Lillard, L., Smith, J.P., et Welch, F. (1986). What do we really know about wages? The importance of nonreporting and census imputation. *Journal of Political Economy*, 94, 3, Partie 1, 488-506 (juin).

Moore, J.C., Bogan, K.E., et Marquis, K.H. (1992). A cognative interviewing approach for the survey of income and program participation: development of procedures and initial test results. Discussion présentée au Symposium 92 de Statistique Canada (novembre).

Scheuren, F.H., Oh, L., Vogel, L., et Yuscavage, R. (1981). *Studies from Interagency Data Linkages*, Rapport n° 10, U.S. Department of Health, Education and Welfare, Social Security Administration, Office of Research and Statistics, SSA Publication Number 13-11750 (janvier).

Scholz, J.K. (1990). The participation rate of the earned income tax credit. La Follette Institute of Public Affairs (août).

United States Bureau of the Census (1991). P-60 n° 174, *Money Income of Households, Persons and Families in the United States: 1990* (août).

Vaughan, D.R. (9189). Reflections on the income estimates from the survey of income and program participation. *ORS Working Paper Number 39*, U.S. Department of Health and Human Services, Social Security Administration, Office of Research and Statistics (septembre).

CONFÉRENCIER SPÉCIAL INVITÉ

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

ESTIMATEURS POUR DES ENQUÊTES LONGITUDINALES AVEC APPLICATION À L'ENQUÊTE «CURRENT POPULATION SURVEY» DES É.-U.

W.A. Fuller[1], A. Adam et I.S. Yansaneh

RÉSUMÉ

On s'intéresse à l'estimation dans les enquêtes répétées avec chevauchement partiel des unités d'échantillonnage. On tient compte des effets liés au temps passé dans l'échantillon pour l'élaboration de certaines méthodes et on décrit la mise en application de ces méthodes dans des enquêtes qui traitent un grand nombre de caractéristiques. Différents estimateurs des caractéristiques de l'emploi fondés sur l'enquête «Current Population Survey» des É.-U. sont comparés.

MOTS-CLÉS: Plan de renouvellement; structure de covariance; meilleur estimateur linéaire sans biais.

1. INTRODUCTION

Nous allons nous intéresser à l'estimation dans le contexte d'enquêtes à passages répétés. Duncan et Kalton (1987) examinent divers types d'enquêtes répétées et les objectifs de ces enquêtes. Nous nous pencherons sur les cas où des déterminations sont faites de façon répétée au sujet de certains éléments de l'échantillon, mais où tous les éléments de l'échantillon ne figurent pas dans l'échantillon à chacune des exécutions de l'enquête.

Dans une étude réalisée il y a longtemps, Jessen (1942) utilisait les moindres carrés pour incorporer l'information recueillie à une occasion précédente dans l'estimation de l'occasion courante. Patterson (1950) a étudié l'estimation visant les échantillons avec renouvellement. Ces travaux ont été suivis de ceux de plusieurs auteurs, notamment Eckler (1955), Rao et Graham (1964), Gurney et Daly (1965), Raj (1965), Singh (1968), Wolter (1979), Huang et Ernst (1981) et Kumar et Lee (1983). Ces auteurs ont traité les quantités inconnues à chaque occasion comme des paramètres fixes.

Blight et Scott (1973), Scott et Smith (1974), Scott, Smith et Jones (1977), Smith (1978) et Jones (1979) ont examiné l'estimation en vertu de l'hypothèse selon laquelle les valeurs vraies sous-jacentes sont la réalisation d'une série chronologique.

Nous examinerons l'estimation dans le cas de l'enquête «Current Population Survey» des É.-U. Notre recherche a consisté à élaborer un modèle représentant la structure de covariance des observations de l'enquête «Current Population Survey», à estimer les paramètres de ce modèle pour deux importantes caractéristiques de la population active et à étudier différentes méthodes d'estimation. Nous supposerons que les valeurs vraies inconnues sont des paramètres fixes.

2. L'ENQUÊTE «CURRENT POPULATION SURVEY»

L'enquête «Current Population Survey» est une enquête menée à l'échelle nationale, visant à produire des estimations par État ainsi que pour le pays dans son ensemble. L'échantillon est un échantillon aréolaire stratifié comportant environ 717 strates dans 50 États. De l'ensemble des strates, 384 contiennent plus d'une unité

[1] W.A. Fuller, Iowa State University, 221, Snedecor Hall, Ames, IA 50011 É.-U.

primaire d'échantillonnage (UPÉ). Les UPÉ sont des secteurs géographiques, et dans les 384 strates, une seule UPÉ est sélectionnée en vue d'observation. Selon la terminologie du Census Bureau, les 333 strates restantes sont des unités primaires d'échantillonnage autoreprésentatives. En termes plus techniques, les unités primaires d'échantillonnage dans les 333 strates sont des sous-secteurs des subdivisions géographiques les plus grandes. Ces sous-secteurs sont des unités plus petites que les UPÉ dans l'autre groupe de 384 strates. L'enquête «Current Population Survey» est une enquête à grande échelle dans le cadre de laquelle environ 57 000 ménages, et environ 113 000 personnes, sont interviewés chaque mois.

Le plan d'enquête est établi de telle façon que les personnes demeurent des répondants pendant quelques mois. Un groupe de personnes particulier est introduit dans l'échantillon, est interviewé pendant quatre mois, reçoit un congé de 8 mois, puis est interviewé pendant une autre période de quatre mois. Après cette deuxième période de quatre mois, les personnes sont retranchées définitivement de l'échantillon. En tout temps, il y a 16 groupes en cause : 8 groupes sont interviewés, tandis que les 8 autres sont en congé. Un des groupes interviewés l'est pour la première fois, un autre pour la deuxième fois, etc. L'échantillon est donc équilibré vis-à-vis du nombre d'interviews subies par les répondants. Les méthodes de collecte des données et d'estimation comportent un certain nombre d'opérations complexes, par exemple un ajustement par la méthode du quotient en fonction des totaux de la population. Ces opérations ne seront pas discutées dans notre présentation.

Les données de base de notre étude étaient formées de deux parties. La première est un ensemble de 48 échantillons répétés pour les 12 mois de 1987. Ces échantillons répétés ont été construits selon un plan conçu par Fay (1989). Essentiellement, des poids sont attribués aux UPÉ d'une manière équilibrée, de telle sorte que chaque ensemble de poids donne une estimation non biaisée du total. Le carré de la différence entre l'estimation fondée sur n'importe quel des échantillons répétés et l'estimation globale est une estimation du quart de la variance de l'estimation globale. Les observations répétées portent sur chacun des 12 mois et sur chacun des 8 groupes de renouvellement interviewés. Il y a 12 mois pour les 8 groupes de renouvellement, et 48 échantillons répétés pour chacun. L'ensemble de données de 1987 contient donc 4 608 observations.

Le tableau 1 contient une représentation des données. Nous appelons les colonnes du tableau des séquences («streams»). La première entrée du premier mois vise les personnes qui sont interviewées pour la première fois. Autrement dit, A_{11} représente l'ensemble des personnes interviewées pour la première fois au premier mois. Ces personnes sont interviewées pour la deuxième fois au deuxième mois. La première entrée de la deuxième séquence vise les personnes qui sont interviewées pour la deuxième fois au premier mois. Si nous nous déplaçons vers le bas dans la deuxième colonne, nous voyons que les personnes qui étaient interviewées pour la deuxième fois au premier mois sont interviewées pour la quatrième fois au troisième mois. Ces personnes sont ensuite retranchées de l'échantillon et un nouveau groupe est amené. Dans ce cas, c'est le groupe E qui est interviewé pour la cinquième fois au quatrième mois. Le groupe E est interviewé pendant quatre mois, puis sort de l'échantillon et fait place à un nouveau groupe. Dans ce cas, c'est le groupe F, qui est interviewé pour la première fois au huitième mois. Ainsi, les groupes de renouvellement apparaissent dans une seule séquence. Ils entrent dans l'échantillon et en sortent selon le plan de renouvellement 4-8-4.

Tableau 1: Organisation des données de 1987.

Mois	Séquences							
1	$A_{1,1}$	$D_{1,2}$	$G_{1,3}$	$J_{1,4}$	$M_{1,5}$	$P_{1,6}$	$T_{1,7}$	$X_{1,8}$
2	$A_{2,2}$	$D_{2,3}$	$G_{2,4}$	$K_{2,5}$	$M_{2,6}$	$P_{2,7}$	$T_{2,8}$	$Y_{2,1}$
3	$A_{3,3}$	$D_{3,4}$	$H_{3,5}$	$K_{3,6}$	$M_{3,7}$	$P_{3,8}$	$U_{3,1}$	$Y_{3,2}$
4	$A_{4,4}$	$E_{4,5}$	$H_{4,6}$	$K_{4,7}$	$M_{4,8}$	$Q_{4,1}$	$U_{4,2}$	$Y_{4,3}$
5	$B_{5,5}$	$E_{5,6}$	$H_{5,7}$	$K_{5,8}$	$N_{5,1}$	$Q_{5,2}$	$U_{5,3}$	$Y_{5,4}$
6	$B_{6,6}$	$E_{6,7}$	$H_{6,8}$	$L_{6,1}$	$N_{6,2}$	$Q_{6,3}$	$U_{6,4}$	$Z_{6,5}$
7	$B_{7,7}$	$E_{7,8}$	$I_{7,1}$	$L_{7,2}$	$N_{7,3}$	$Q_{7,4}$	$V_{7,5}$	$Z_{7,6}$
8	$B_{8,8}$	$F_{8,1}$	$I_{8,2}$	$L_{8,3}$	$N_{8,4}$	$R_{8,5}$	$V_{8,6}$	$Z_{8,7}$
9	$C_{9,1}$	$F_{9,2}$	$I_{9,3}$	$L_{9,4}$	$O_{9,5}$	$R_{9,6}$	$V_{9,7}$	$Z_{9,8}$
10	$C_{10,2}$	$F_{10,3}$	$I_{10,4}$	$J_{10,5}$	$O_{10,6}$	$R_{10,7}$	$V_{10,8}$	$\Gamma_{10,1}$
11	$C_{11,3}$	$F_{11,4}$	$G_{11,5}$	$J_{11,6}$	$O_{11,7}$	$R_{11,8}$	$W_{11,1}$	$\Gamma_{11,2}$
12	$C_{12,4}$	$D_{12,5}$	$G_{12,6}$	$J_{12,7}$	$O_{12,8}$	$S_{12,1}$	$W_{12,2}$	$\Gamma_{12,3}$

3. STRUCTURE DE COVARIANCE DES ESTIMATEURS DE BASE

Nous avons commencé notre recherche en supposant un modèle d'analyse de variance pour ces données. Soit

$$y_{gjk} = \mu + u_j + \alpha_t + \tau_k + \gamma_g + \zeta_{tk} + \epsilon_{tkj}, \qquad (1)$$

où y_{gjk} est l'observation au temps t, c.-à-d. au mois t, pour le j-ième échantillon répété et le groupe de renouvellement rendu à la k-ième interview, μ est la moyenne globale, u_j est un effet lié à la répétition des échantillons, α_t est un effet lié au mois, τ_k est un effet lié au temps passé dans l'échantillon, γ_g est un effet lié au groupe de renouvellement et les ζ_{tk} sont des effets d'interaction. L'indice g, désignant un groupe de renouvellement, est entièrement déterminé par le mois t et le temps passé dans l'échantillon k. Voir le tableau 1.

Le tableau 2 contient les résultats de l'analyse de variance obtenus avec les données recueillies, selon le modèle utilisé. Les effets liés au mois prédominent. Il est clair que les niveaux du chômage et de la population active civile varient selon les mois. On note aussi des effets marqués liés au temps passé dans l'échantillon. La somme des carrés pour les effets liés au groupe de renouvellement est ajustée en fonction du mois et du temps passé dans l'échantillon. Après la suppression des effets liés au groupe de renouvellement, il reste 52 degrés de liberté sur les 95 degrés de liberté originaux. Les carrés moyens pour ces degrés de liberté sont donnés à la ligne appelée «Interactions».

Nous sommes disposés à faire l'hypothèse que les échantillons répétés, par leur construction, sont quasi-indépendants. La corrélation entre les observations faites à l'intérieur d'un même groupe de renouvellement signifie que les hypothèses nécessaires à l'exécution de tests F classiques pour ce tableau ne sont pas respectées. Toutefois, la plupart des observateurs seront disposés à conclure que des effets prononcés sont présents dans cet ensemble de données, notamment des effets liés au mois. L'effet lié au temps passé dans l'échantillon est aussi très important.

Tableau 2: Analyse de variance pour les personnes occupées, les chômeurs et la population active civile, 1987.

Source	Degrés de liberté	Carrés moyens		
		Personnes occupées	Chômeurs	PAC
Répétition des éch.	47	1.2134	0.1785	1.0762
Mois	11	3553.2435	268.4974	2377.7240
Temps passé dans l'éch.	7	458.1149	75.4759	891.2340
Groupes[1]	25	113.4492	20.7709	91.5742
Interactions	52	12.9719	6.7783	11.7550
Résidu	4465	0.2458	0.0554	0.2112

[1] Le carré moyen, pour les groupes, est ajusté en fonction du mois et du temps passé dans l'échantillon.

Afin d'estimer la structure de covariance de ces données, nous examinons un sous-ensemble des effets définis dans notre modèle original. Désignons par r_{gjk} la somme de l'effet lié à la répétition et de l'effet epsilon de notre modèle original. Nous utilisons l'indice g, pour le groupe, plutôt que pour le temps. Comme il a été mentionné plus haut, le fait de connaître g et k équivaut à connaître t et k. Nous écrivons

$$r_{gjk} = u_j + e_{gj} + a_{gjk}, \tag{2}$$

où u_j est l'effet lié à la répétition, e_{gj} est l'effet permanent lié au groupe de renouvellement et a_{gjk} est l'effet transitoire lié au groupe de renouvellement. L'effet lié à la répétition est un reflet de la différence entre les unités primaires d'échantillonnage. Pour les besoins de ce modèle, cet effet est supposé constant dans le temps. Par exemple, la moyenne dans le temps de certaines unités primaires d'échantillonnage est supérieure à celle d'autres unités primaires d'échantillonnage. L'effet e_{gj} est un effet semblable pour les groupes de renouvellement. La moyenne à long terme pour certains groupes de renouvellement est supposée supérieure à celle d'autres groupes. La composante finale de notre modèle, dénotée par a_{gjk}, vise à refléter la corrélation, dans les observations recueillies dans le même groupe de renouvellement, qui a tendance à diminuer à mesure que la durée entre les observations augmente. En vertu de notre modèle, peu importe le moment où nous observons le j-ième échantillon répété, nous obtenons un u_j. Peu importe le moment où nous observons ce groupe de personnes particulier, nous obtenons un e_{gj}. Mais il y a aussi un effet lié à un groupe particulier qui se dissipe avec le temps. Par exemple, une personne a une propension générale à avoir un emploi. Mais si nous observons cette personne à deux moments rapprochés, la personne est plus susceptible soit d'être en chômage, soit d'avoir un emploi, les deux fois. Les a_{gjk} servent à représenter l'effet dont la corrélation diminue à mesure qu'augmente l'intervalle entre les observations.

Nous supposons que l'effet transitoire lié au groupe de renouvellement satisfait une autorégression de troisième ordre,

$$a_{gjk} = \xi_1 a_{gj,k-1} + \xi_2 a_{gj,k-2} + \xi_3 a_{gj,k-3} + b_{gjk}, \tag{3}$$

où

$$b_{gjk} \sim \text{Ind}(0, \sigma_b^2).$$

Selon notre modèle, la corrélation entre les observations faites à deux occasions dans le même groupe de renouvellement est

$$\rho_r(h) = \frac{\sigma_u^2 + \sigma_e^2 + \rho_a(h)\sigma_a^2}{\sigma_u^2 + \sigma_e^2 + \sigma_a^2}, \tag{4}$$

où h est l'intervalle qui sépare les deux occasions et $\rho_a(h)$ est l'autocorrélation de l'effet a. La variance d'une observation choisie au hasard est $\sigma_u^2 + \sigma_e^2 + \sigma_a^2$. La covariance entre deux observations venant du même échantillon répété, mais de groupes de renouvellement différents, est σ_u^2 pour n'importe quel h.

Nous pouvons estimer l'autocovariance pour un groupe de renouvellement. L'autocovariance est une estimation de $\sigma_u^2 + \sigma_a^2 + \rho_a(h)\sigma_a^2$. Cet ensemble d'estimations de covariances ne permet pas, à lui seul, de séparer les effets. La ligne de l'erreur dans l'analyse de variance du tableau 2 donne une estimation de $\sigma_a^2 + \sigma_e^2$, et une analyse de variance distincte a été utilisée pour estimer σ_u^2. En ayant les estimations de ces trois quantités, nous pouvons obtenir une solution pour les autres paramètres.

Tableau 3: Autocorrélations moyennes dans un groupe de renouvellement pour l'enquête «Current Population Survey» de 1987.

Décalage	Nombre d'obs.	Personnes occupées	Chômeurs	Population active civile	Taux de chômage
1	66	0.8068	0.4979	0.7876	0.5187
		(0.0062)	(0.0136)	(0.0068)	(0.0132)
2	40	0.7332	0.3788	0.7197	0.4019
		(0.0106)	(0.0199)	(0.0111)	(0.0195)
3	18	0.6856	0.3230	0.6668	0.3484
		(0.0182)	(0.0312)	(0.0192)	(0.0306)
9	3	0.6732	0.1566	0.6377	0.2034
		(0.0461)	(0.0832)	(0.0504)	(0.0818)
10	4	0.7191	0.2691	0.6187	0.3159
		(0.0354)	(0.0686)	(0.0452)	(0.0665)
11	3	0.6038	0.1401	0.4910	0.2138
		(0.0536)	(0.0830)	(0.0614)	(0.0814)
Moy. 9-11	10	0.6708	0.1966	0.5861	0.2443
		(0.0252)	(0.0450)	(0.0298)	(0.0439)

Le tableau 3 contient les autocorrélations estimées pour le chômage et la population active civile. Les autocorrélations sont beaucoup plus fortes pour la population active civile que pour le chômage. Il y a un nombre limité de groupes de renouvellement observés aux décalages 9, 10 et 11, de sorte que la moyenne pour ces décalages a été incluse dans le tableau. La corrélation est voisine de 0.60 aux décalages 9 et 11 pour la population active civile.

Le tableau 4 contient les estimations des paramètres du processus autorégressif pour a_{gik}. Le tableau 5 contient les estimations des autres paramètres de notre modèle.

La figure 1 présente un graphique de l'autocorrélation estimée dans le cas des chômeurs. Il s'agit de la fonction d'autocorrélation pour un groupe de renouvellement particulier. Nous avons seulement observé les autocorrélations pour les périodes zéro à 3 et pour les périodes 9, 10 et 11. Toutes les autres autocorrélations indiquées par les points sont estimées à l'aide de notre modèle. Les carrés qui se trouvent dans la figure sont les autocorrélations estimées par Breau et Ernst (1983) à l'aide de données relatives aux années 1976 et 1977. Les autocorrélations estimées sont très voisines pour les deux périodes.

Figure 1: Estimations des autocorrélations pour les chômeurs.

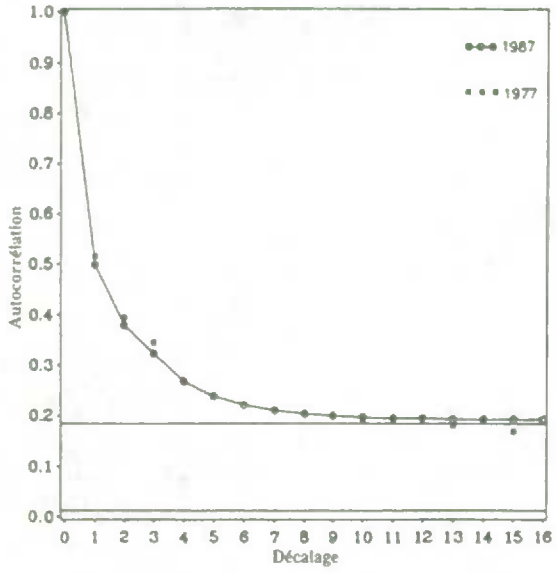
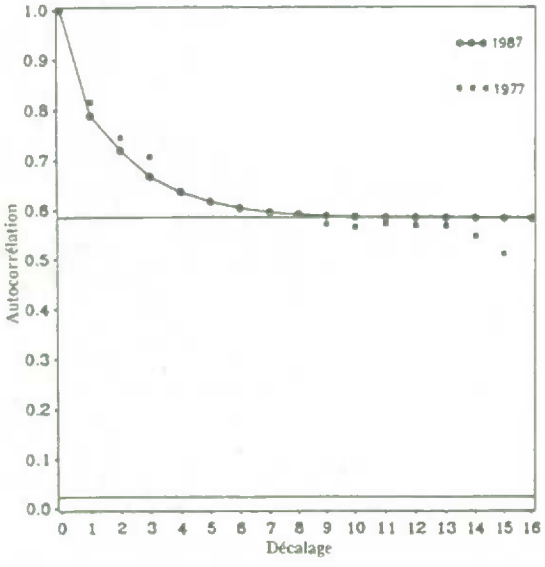


Figure 2: Estimations des autocorrélations pour la population active civile.



La figure 2 présente des estimations semblables des autocorrélations pour la population active civile. Encore une fois, les carrés représentent des estimations faites par Breau et Ernst d'après des données de 1976 et de 1977. Les estimations de Breau et Ernst sont un peu supérieures pour les autocorrélations sur de courtes périodes et un peu inférieures pour les autocorrélations sur des périodes plus longues.

Tableau 4: Estimations des paramètres des processus transitoires a_{gjk}.

Modèle $a_{gjk} = \xi_1 a_{gj,k-1} + \xi_2 a_{gj,k-2} + \xi_3 a_{gj,k-3} + b_{gjk}$					
Charactéristique	$\hat{\xi}_1$	$\hat{\xi}_2$	$\hat{\xi}_3$	$\hat{\sigma}_b^2$	$\hat{\sigma}_a^2$
Personnes occupées	0.40481	0.04270	-0.04945	0.06841	0.08263
Chômeurs	0.33422	0.08452	0.05267	0.03831	0.04508
Population active civile	0.43415	0.11345	0.00318	0.06756	0.08962

Tableau 5: Estimations de σ_u^2, σ_e^2, et σ_a^2.

Charactéristique	Composante de variance			
	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_a^2$	Total
Personnes occupées	0.00552	0.16319	0.08263	0.25134
Chômeurs	0.00065	0.01037	0.04508	0.05610
Population active civile	0.00499	0.12159	0.08962	0.21620

Les droites se trouvant dans les figures représentent les deux composantes permanentes. La droite du bas représente σ_u^2 en pourcentage de la variation totale. Celle du haut représente la somme $\sigma_u^2 + \sigma_e^2$ en pourcentage du total. Selon notre modèle, même après de longs intervalles de temps, il existe une certaine corrélation en raison de l'effet permanent lié aux groupes de renouvellement et de l'effet permanent lié aux unités primaires d'échantillonnage.

Pour la population active civile, σ_u^2 et les $\sigma_u^2 + \sigma_e^2$ constituent une fraction beaucoup plus élevée de la variation totale que dans le cas des chômeurs. Ainsi, l'autocorrélation visant la population active civile est importante même après des intervalles extrêmement longs. Le profil est sensiblement le même pour les personnes occupées que pour la population active civile. En fait, la corrélation à long terme est plus forte dans le cas des personnes occupées que dans celui de la population active civile.

L'autocorrélation pour le taux de chômage est très voisine de la fonction de corrélation visant les chômeurs. La variation du dénominateur, qui est la population active civile, contribue peu à la variation du taux.

4. COMPARAISON DE DIFFÉRENTS ESTIMATEURS

On obtient un estimateur simple pour la population active civile au temps t en prenant la moyenne des estimations relatives aux huit groupes de renouvellement. Nous appelons ce dernier l'estimateur direct ou l'estimateur de base et le désignons par $\bar{y}_{t0}$. Selon notre modèle, la variance de l'estimateur direct est

$$V\{\bar{y}_{t0.}\} = (64)^{-1}(64\sigma_u^2 + 8\sigma_e^2 + 8\sigma_a^2). \qquad (5)$$

Nous supposons que les groupes de renouvellement sont indépendants au sein des unités primaires d'échantillonnage et que la variation des unités primaires d'échantillonnage est reflétée dans σ_u^2, la variance des échantillons répétés. Les coefficients de variation pour les estimateurs directs sont d'environ 0.3 % pour les personnes occupées, d'environ 1.6 % pour les chômeurs et d'environ 0.2 % pour la population active civile.

Le tableau 6 donne la contribution des différentes composantes à la variance de l'estimateur direct. Les contributions sont très différentes pour les chômeurs et la population active civile. Dans le cas des chômeurs, environ 9 % de la variation provient, selon l'estimation, de l'effet lié à l'échantillon répété, tandis que pour la population active civile, la contribution liée à l'échantillon répété est d'environ 16 %. Environ 74 % de la variance, dans le cas du chômage, provient de l'effet transitoire lié au groupe de renouvellement, tandis que ce même effet ne représente que 36 % de la variance dans le cas de la population active civile. L'effet permanent lié au groupe de renouvellement est la composante la plus importante pour la population active civile, tandis qu'il ne représente que 17 % de la variance dans le cas des chômeurs.

Tableau 6: Variance de l'estimateur direct d'après huit séquences.

Propriété	Personnes occupées	Chômeurs	Population active civile	Taux de chômage
Variance	9.2793	1.9408	8.0361	8.3994×10^{-5}
% dû à $\hat{\sigma}_u^2$	15.22	8.60	15.89	8.56
% dû à $\hat{\sigma}_e^2$	56.28	17.10	48.41	17.76
% dû à $\hat{\sigma}_a^2$	28.50	74.30	35.70	73.68

L'estimateur actuellement utilisé dans l'enquête «Current Population Survey» est une moyenne pondérée de l'estimateur direct, d'une quantité formée de l'estimateur composite précédent plus une estimation du changement, et d'une troisième combinaison linéaire des estimations courantes de chacun des groupes de renouvellement. L'estimateur est le suivant:

$$\hat{\mu}_{tB.} = 0.6\bar{y}_{t0.} + 0.4(\hat{\mu}_{t-1,B.} + \hat{\delta}_{t,t-1}) + 0.05[2^{-1}(y_{t01} + y_{t05}) - 6^{-1} \sum_{k=2}^{4} (y_{t0k} + y_{t,0,k+4})], \tag{6}$$

où

$$\bar{y}_{t0.} = 8^{-1} \sum_{k=1}^{8} y_{t0k},$$

$$\hat{\delta}_{t,t-1} = 6^{-1} \left[\sum_{k=2}^{4} (y_{t0k} + y_{t,0,k+4}) - \sum_{k=2}^{4} (y_{t-1,0,k-1} + y_{t-1,0,k+3}) \right],$$

$\hat{\mu}_{tB.}$ est l'estimateur pour la période t, et $\hat{\delta}_{t,t-1}$ est un estimateur du changement construit d'après les groupes de renouvellement observés à la fois à la période t et à la période $t-1$. Jusqu'à 1985, l'estimateur ne contenait que les deux premiers termes. Le troisième terme de l'estimateur a procuré plusieurs avantages. Ce terme correspond à la différence entre la moyenne du premier et du cinquième groupe de renouvellement et la moyenne de tous les autres groupes de renouvellement. Il réduit les effets liés au temps passé dans l'échantillon qui apparaissaient dans l'estimateur original. Les groupes de renouvellement faisant partie de l'échantillon pour la première et la cinquième fois produisent des estimations plus élevées des chômeurs que ne le font les autres groupes de renouvellement. Par conséquent, la différence directe, $\delta_{t,t-1}$, est influencée par le fait que le groupe de renouvellement présent pour la première fois a une espérance plus élevée que le groupe dont c'est la deuxième présence. Les effets liés au temps passé dans l'échantillon ne s'annulent pas dans l'estimation de la différence. L'inclusion du troisième terme a pour résultat de rapprocher l'espérance de l'estimateur de l'espérance de l'estimateur direct. Ce terme réduit également la variance de l'estimateur par rapport à celle de l'estimateur à deux termes utilisé avant 1985.

Nous examinons maintenant le meilleur estimateur linéaire sans biais du niveau courant. Cet estimateur utilise toutes les données observées jusqu'au temps t pour donner la meilleure estimation au temps t. Selon la tradition de l'enquête «Current Population Survey», l'estimateur pour le temps t ne sera pas modifié à mesure que de nouvelles données deviendront disponibles.

Nous montrons que pour construire le meilleur estimateur linéaire sans biais, il n'est pas nécessaire d'emmagasiner toutes les observations précédentes. Cependant, toutes les observations précédentes relatives à n'importe quel groupe de renouvellement observé au temps t sont nécessaires. Dans le cas de l'enquête «Current Population Survey», l'estimateur tiendra compte de certaines observations remontant jusqu'à 15 mois

en arrière, car si le groupe de renouvellement est observé pour la dernière fois, c'est que ce groupe est associé à l'enquête depuis 16 mois.

Nous illustrons la construction du meilleur estimateur linéaire sans biais en utilisant un plan dans lequel 3 groupes sont observés à chaque point du temps. Nous supposons qu'un groupe est intégré à l'échantillon, est observé pendant 3 périodes et est ensuite retranché définitivement de l'échantillon. Selon ce plan simple, les meilleurs estimateurs aux temps $t-2$ et $t-1$ sont utilisés pour construire le meilleur estimateur linéaire sans biais au temps t. Il y a trois types d'observations au temps t, soit celles faites pour la première fois, celles faites pour la deuxième fois et celles faites pour la troisième fois. On suppose que les groupes sont indépendants. Ainsi, l'observation venant du groupe présent pour la première fois est indépendante de toutes les autres données. L'observation relative à un groupe présent pour la deuxième fois est en corrélation avec l'observation précédente pour ce groupe. Si nous faisons une régression de cette observation de deuxième période en fonction de l'observation précédente, nous produisons un écart, désigné par w_{t2}, qui n'est pas corrélé avec l'observation précédente. La combinaison linéaire créée à partir de l'observation de troisième période, w_{t3}, est indépendante de toutes les observations précédentes. Il y a cinq observations qui servent pour les besoins de l'estimateur. Il s'ensuit que le meilleur estimateur linéaire sans biais peut être construit selon le modèle linéaire. Autrement dit, toute l'information, jusqu'à la période courante, qui est utile à l'estimation de θ_t, est condensée dans les deux meilleurs estimateurs précédents et les trois observations courantes transformées. Nous avons donc un modèle linéaire en $(\theta_t, \theta_{t-1}, \theta_{t-2})$, où, selon notre notation précédente, $\theta_t = \mu + a_t$. Compte tenu de la matrice de covariance des cinq estimateurs, nous nous servons des moindres carrés généralisés pour construire notre meilleur estimateur du niveau courant.

Le modèle linéaire est

$$\begin{bmatrix} \hat{\theta}_{t-2} \\ \hat{\theta}_{t-1} \\ w_{t1} \\ w_{t2} \\ w_{t3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -b_{21} & 1 \\ -b_{32} & -b_{31} & 1 \end{bmatrix} \begin{bmatrix} \theta_{t-2} \\ \theta_{t-1} \\ \theta_t \end{bmatrix} + e_t,$$

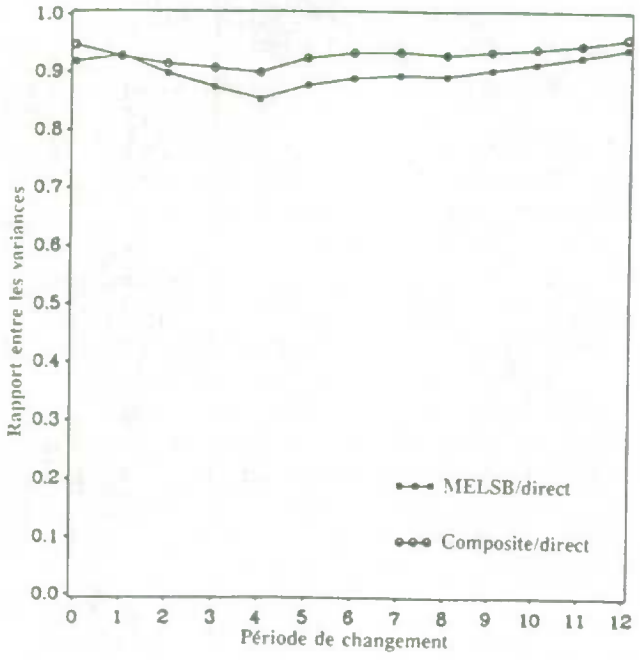
où, $w_{t1} = y_{t01}$, $w_{t2} = y_{t02} - b_{21} y_{t-1,0,1}$, $w_{t3} = y_{t03} - b_{32} y_{t-1,0,2} - b_{31} y_{t-2,0,1}$, e_t est le vecteur des différences entre les observations et leurs espérances, et les b_{ij} sont les coefficients de régression de la population. La matrice de covariance de $(\hat{\theta}_{t-2}, \hat{\theta}_{t-1})$ est connue, car elle découle de l'ajustement des moindres carrés effectué au temps *t-1*. Le vecteur (w_{t1}, w_{t2}, w_{t3}) n'est pas corrélé avec $(\hat{\theta}_{t-2}, \hat{\theta}_{t-1})$ et a une matrice de covariance diagonale déterminée par la structure de covariance des y_{t0j}.

Afin de construire le meilleur estimateur au temps t pour le plan de renouvellement 4-8-4, il est nécessaire d'emmagasiner 15 estimations pour les temps $t-1, t-2, \ldots, t-15$, ainsi que 60 observations précédentes relatives aux 15 groupes de renouvellement qui ont été observés précédemment. Bien que seulement 8 groupes soient observés à un moment quelconque, 16 groupes au total sont associés à l'enquête de ce moment. Les variances des différents estimateurs pour les chômeurs, la population active civile et les personnes occupées sont données aux tableaux 7, 8 et 9 respectivement.

Les variances des différents estimateurs pour les chômeurs sont comparées à la figure 3. La courbe formée de petits cercles représente la variance de l'estimateur composite courant divisée par la variance de l'estimateur direct pour chacune des périodes de changement, zéro dénotant l'estimateur du niveau courant.

La courbe formée de carrés représente la variance du meilleur estimateur linéaire sans biais (MELSB) divisée par la variance de l'estimateur direct. Les estimateurs utilisant des données antérieures ne sont que légèrement supérieurs à l'estimateur direct dans le cas des chômeurs. L'estimateur composite actuel du niveau courant a une variance se situant à environ 95 % de la variance de l'estimateur direct, tandis que la variance de l'estimateur des moindres carrés représente environ 92 % de celle de l'estimateur direct. Les deux estimateurs utilisant de l'information antérieure sont d'environ 7 % supérieurs à l'estimateur direct pour le changement survenu dans une période. Le gain maximal, d'environ 15 %, est obtenu avec le meilleur estimateur linéaire sans biais, pour un changement sur quatre mois.

Figure 3: Rapports entre les variances de différents estimateurs et les variances de l'estimateur direct pour les chômeurs.



Les variances de différents estimateurs dans le cas des personnes occupées sont comparées à la figure 4.

Figure 4: Rapports entre les variances de différents estimateurs et les variances de l'estimateur direct pour les personnes occupées.

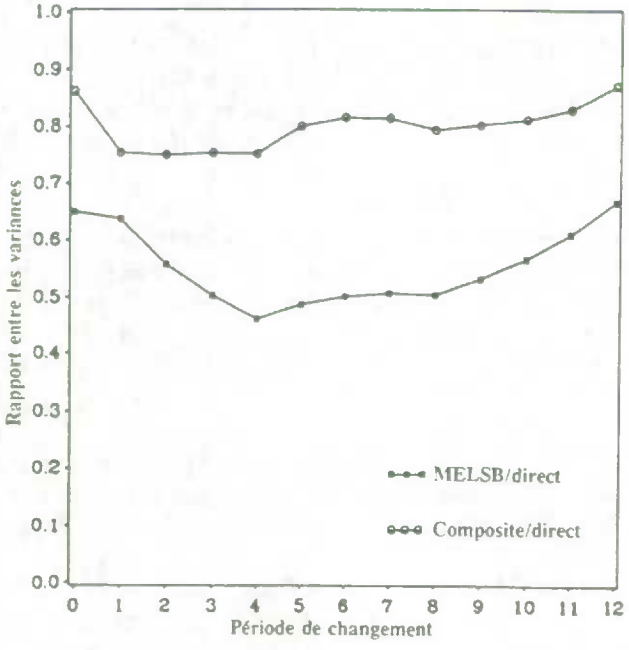


Tableau 7: Variances de différents estimateurs pour les chômeurs par rapport à la variance de l'estimateur direct du niveau courant.

Période de changement	Estimateur direct	Composite actuel	MELSB courant
0	1.000	0.947	0.918
1	1.155	1.070	1.073
2	1.490	1.361	1.338
3	1.686	1.528	1.473
4	1.830	1.645	1.562
5	1.830	1.691	1.606
6	1.830	1.708	1.628
7	1.830	1.710	1.636
8	1.830	1.701	1.634
9	1.787	1.671	1.614
10	1.745	1.641	1.595
11	1.703	1.614	1.578
12	1.660	1.593	1.564

La variance de l'estimateur composite actuel pour le changement survenu sur une période, dans le cas des personnes occupées, représente environ 86 % de la variance de l'estimateur direct pour le changement sur une période. La variance du changement sur une période, dans le cas du meilleur estimateur du niveau courant, est d'environ 65 % de la variance de l'estimateur direct pour le changement survenu sur une période. Le meilleur estimateur, dans le cas des personnes occupées, devient encore meilleur par rapport à l'estimateur direct et à l'estimateur composite à mesure que s'allonge la période visée par le changement. La variance du meilleur estimateur pour le changement sur 4 périodes est de moins de 50 % de celle de l'estimateur direct. L'estimateur composite actuel du changement survenu sur 4 périodes a une variance qui correspond à environ 75 % de celle de l'estimateur direct. Pour un écart de 11 périodes, l'estimateur composite courant a une variance se situant à environ 84 % de celle de l'estimateur direct, tandis que la variance du meilleur estimateur représente environ 68 % de celle de l'estimateur direct.

Tableau 8: Variances de différents estimateurs pour la population active civile par rapport à la variance de l'estimateur direct du niveau courant.

Période de changement	Estimateur direct	Composite actuel	MELSB courant
0	1.000	0.868	0.704
1	0.697	0.538	0.474
2	1.086	0.828	0.652
3	1.410	1.076	0.774
4	1.688	1.283	0.859
5	1.688	1.364	0.914
6	1.688	1.392	0.946
7	1.688	1.391	0.962
8	1.688	1.362	0.963
9	1.565	1.277	0.942
10	1.445	1.191	0.921
11	1.324	1.111	0.903
12	1.203	1.052	0.887

L'efficacité relative pour le taux de chômage est très voisine de celle obtenue pour les chômeurs. Celle obtenue dans le cas de la population active civile est à peu près semblable à celle visant les personnes occupées, mais

parce que les corrélations sont un peu plus fortes pour les personnes occupées, le MELSB donne des résultats légèrement meilleurs pour les personnes occupées que pour la population active civile.

Tableau 9: Variances de différents estimateurs pour les personnes occupées par rapport à la variance de l'estimateur direct du niveau courant.

Période de changement	Estimateur direct	Composite actuel	MELSB courant
0	1.000	0.862	0.650
1	0.677	0.511	0.432
2	1.083	0.813	0.604
3	1.413	1.065	0.711
4	1.701	1.279	0.784
5	1.701	1.363	0.829
6	1.701	1.390	0.855
7	1.701	1.388	0.865
8	1.701	1.353	0.860
9	1.560	1.255	0.832
10	1.419	1.154	0.806
11	1.278	1.061	0.782
12	1.137	0.992	0.761

Nous avons aussi calculé les variances d'estimateurs linéaires fondés sur 12 mois de données, 16 mois de données et 24 mois de données. La perte maximale d'efficacité résultant de l'usage de 12 mois de données plutôt que la totalité des données antérieures était de 8 % pour l'estimation du niveau courant, dans le cas des personnes occupées. La perte maximale pour l'estimateur fondé sur 16 mois de données était de 3 %. La perte maximale d'efficacité résultant de l'usage de 24 mois de données plutôt que la totalité des données antérieures était de 2 %. Certains estimateurs du changement sur une longue période, par exemple le changement d'une année à l'autre, avaient une variance plus faible lorsqu'ils étaient fondés sur seulement 24 mois de données.

Le plan de renouvellement actuel de l'enquête «Current Population Survey» prévoit que les personnes sont interrogées pendant 4 périodes, sont retirées de l'échantillon pour 8 périodes, puis y sont de nouveau intégrées pour 4 périodes. Le but de ce plan était de permettre un accroissement de l'efficacité des estimations des changements d'une année à l'autre.

Nous avons comparé les estimations obtenues avec ce plan de renouvellement à un plan dans lequel les personnes sont incluses dans l'échantillon pendant 6 périodes, puis en sortent définitivement, et un plan dans lequel les personnes font partie de l'échantillon pendant 8 périodes, puis en sont retranchées définitivement.

Le tableau 10 présente les variances des différents plans d'estimation par échantillonnage pour les chômeurs.

L'efficacité de l'estimateur du niveau courant selon un plan dans lequel les personnes demeurent dans l'échantillon pendant 6 périodes consécutives est légèrement inférieure à celle du plan 4-8-4, dans le cas des chômeurs. Les estimations des changements survenus sur une courte période sont légèrement supérieures dans le cas du plan à 6 périodes consécutives, mais pour les changements survenus sur une longue période, le plan 4-8-4 produit des estimations supérieures. Dans le cas du plan dans lequel les personnes demeurent dans l'échantillon pendant 8 périodes consécutives, qui a aussi été examiné, on note pour les chômeurs une légère perte pour le niveau courant par rapport au plan 4-8-4, mais un gain de l'ordre de 5 % pour les changements sur une courte période. Le plan 4-8-4 est supérieur à celui des 8 périodes consécutives pour les changements sur une plus longue période, notamment les changements d'une année à l'autre.

Tableau 10: Variances de différents estimateurs des chômeurs; la variance de l'estimateur direct du niveau courant est égale à un.

Quantité estimée	Estimateur composite actuel	Meilleur estimateur 4-8-4	Meilleur estimateur 8 pér. consécutives	Meilleur estimateur 6 pér. consécutives
Niveau courant	0.947	0.918	0.944	0.938
Changement 1 pér.	1.070	1.073	1.003	1.051
Changement 2 pér.	1.361	1.338	1.250	1.312
Changement 3 pér.	1.528	1.473	1.372	1.443
Changement 4 pér.	1.645	1.562	1.473	1.543
Changement 5 pér.	1.691	1.606	1.533	1.607
Changement 6 pér.	1.708	1.628	1.577	1.655
Changement 7 pér.	1.710	1.636	1.612	1.686
Changement 8 pér.	1.701	1.634	1.642	1.705
Changement 9 pér.	1.671	1.614	1.663	1.719
Changement 10 pér.	1.641	1.595	1.678	1.727
Changement 11 pér.	1.614	1.578	1.688	1.733
Changement 12 pér.	1.593	1.564	1.696	1.737
Moyenne de 12 pér.	0.255	0.249	0.301	0.266
Changement des moyennes de 12 pér.	0.273	0.262	0.372	0.359

Tableau 11: Variances de différents estimateurs de la population active civile; la variance de l'estimateur direct du niveau courant est égale à un.

Quantité estimée	Estimateur composite actuel	Meilleur estimateur 4-8-4	Meilleur estimateur 8 pér. consécutives	Meilleur estimateur 6 pér. consécutives
Niveau courant	0.868	0.706	0.796	0.783
Changement 1 pér.	0.538	0.474	0.430	0.470
Changement 2 pér.	0.828	0.652	0.589	0.651
Changement 3 pér.	1.076	0.774	0.709	0.786
Changement 4 pér.	1.283	0.859	0.793	0.883
Changement 5 pér.	1.364	0.913	0.858	0.962
Changement 6 pér.	1.392	0.946	0.913	1.032
Changement 7 pér.	1.391	0.961	0.963	1.088
Changement 8 pér.	1.362	0.962	1.010	1.133
Changement 9 pér.	1.277	0.941	1.049	1.170
Changement 10 pér.	1.191	0.920	1.083	1.199
Changement 11 pér.	1.111	0.902	1.111	1.223
Changement 12 pér.	1.052	0.886	1.135	1.242
Moyenne de 12 pér.	0.369	0.346	0.448	0.396
Changement des moyennes de 12 pér.	0.259	0.206	0.393	0.413

Un profil semblable peut être observé dans le cas de la population active civile. Il y a une perte au niveau courant pour le plan des 8 périodes consécutives par rapport au plan 4-8-4. En vertu du plan comportant des périodes consécutives, nous traitons 8 groupes à n'importe quelle période, tandis qu'en vertu du plan 4-8-4, nous

traitons 16 groupes. Les huit groupes additionnels apportent un supplément d'information. Par contre, les estimations du changement sur une durée allant jusqu'à environ 6 périodes sont supérieures dans le cas du plan de 8 périodes consécutives.

Tableau 12: Variances de différents estimateurs des personnes occupées; la variance de l'estimateur direct du niveau courant est égale à un.

Quantité estimée	Estimateur composite actuel	Meilleur estimateur 4-8-4	Meilleur estimateur 8 pér. conséc.	Meilleur estimateur 6 pér. conséc.
Niveau courant	0.862	0.653	0.761	0.759
Changement 1 pér.	0.511	0.432	0.395	0.434
Changement 2 pér.	0.813	0.604	0.559	0.619
Changement 3 pér.	1.065	0.710	0.669	0.747
Changement 4 pér.	1.279	0.783	0.731	0.829
Changement 5 pér.	1.363	0.828	0.782	0.901
Changement 6 pér.	1.390	0.854	0.828	0.970
Changement 7 pér.	1.388	0.863	0.874	1.026
Changement 8 pér.	1.353	0.858	0.960	1.071
Changement 9 pér.	1.255	0.830	0.960	1.108
Changement 10 pér.	1.154	0.803	0.993	1.139
Changement 11 pér.	1.061	0.779	1.021	1.165
Changement 12 pér.	0.992	0.758	1.046	1.186
Moyenne de 12 pér.	0.369	0.326	0.440	0.394
Changement des moyennes de 12 pér.	0.248	0.162	0.365	0.403

Les estimations des changements survenus sur plus de 8 périodes sont supérieures pour le plan 4-8-4, et l'estimation du changement d'une année à l'autre est d'environ 20 % supérieure pour le plan 4-8-4. Dans le cas des personnes occupées, le plan 4-8-4 l'emporte sur le plan des 6 périodes consécutives.

Selon le plan d'estimation actuel, les coefficients qui sont appliqués aux estimations directes pour construire les estimateurs relatifs aux personnes occupées sont les mêmes que ceux utilisés pour construire les estimations relatives aux chômeurs. Cela signifie qu'il y a convergence interne des estimations. L'espérance de chaque estimateur est exactement la même combinaison linéaire des effets liés au mois et des effets liés au temps passé dans l'échantillon.

Toutefois, parce que la structure d'autocorrélation est différente pour les personnes occupées et pour les chômeurs, et parce que les coefficients minimisent approximativement la variance pour les chômeurs, les estimations des personnes occupées sont inefficaces.

La méthode suivante pourrait être employée pour accroître l'efficacité des estimations des personnes occupées.

1. Construire des estimations de la population active civile qui soient les meilleures sous réserve que l'espérance de l'estimateur soit la même combinaison linéaire des effets liés au temps passé dans l'échantillon que celle de l'espérance de l'estimateur des chômeurs.

2. Au moyen des estimations de la population active civile et des chômeurs, construire des poids, pour les observations courantes, qui reproduisent les estimations des chômeurs et de la population active civile.

L'estimateur théoriquement optimal pour une caractéristique y utiliserait l'information passée relative à toutes les caractéristiques. En pratique, cela n'est pas possible. Notre recherche a démontré que les corrélations croisées entre les chômeurs et la population active civile sont faibles. Les six premières corrélations croisées sont

estimées à moins de 0.10 en valeur absolue. Il s'ensuit que les estimations fondées seulement sur les valeurs passées de la caractéristique estimée sont quasi optimales pour ces deux caractéristiques.

Nous avons étudié le comportement des effets liés au temps passé dans l'échantillon en construisant des combinaisons linéaires des estimations des 8 groupes de renouvellement de base de chaque période qui sont des fonctions linéaires uniquement des effets liés au temps passé dans l'échantillon pour la population. Les variations, sur une période de onze ans allant de 1980 à 1990, des contrastes entre effets liés au temps passé dans l'échantillon sont approximativement égales aux variances estimées des contrastes. Par conséquent, les données n'offrent aucune raison de rejeter l'hypothèse selon laquelle les effets liés au temps passé dans l'échantillon ont été constants durant cette période.

Malgré ce résultat, de nombreux praticiens n'accepteraient pas d'adopter un modèle dans lequel les effets liés au temps passé dans l'échantillon seraient constants sur plusieurs années. Nous présentons la variance d'estimations construites de manière à avoir la même espérance que celle de l'estimateur direct. Toutefois, il serait possible de construire des estimateurs ayant comme espérance n'importe quelle combinaison linéaire des effets liés au temps passé dans l'échantillon. Par exemple, on pourrait construire un estimateur ayant la même espérance que l'estimateur composite actuel.

Tableau 13: Variances d'estimateurs linéaires avec effets liés au temps passé dans l'échantillon; la variance de l'estimateur direct du niveau courant est égale à un.

Période de Changement	Chômeurs		Pop. active civile	
	MELSB 24	Récursif 36	MESLB 24	Récursif 36
0	0.928	0.923	0.763	0.733
1	1.089	1.075	0.490	0.480
2	1.348	1.342	0.682	0.663
3	1.487	1.479	0.816	0.789
4	1.578	1.569	0.911	0.877
5	1.623	1.613	0.974	0.934
6	1.646	1.635	1.015	0.970
7	1.656	1.644	1.039	0.987
8	1.655	1.642	1.047	0.990
9	1.635	1.622	1.035	0.972
10	1.617	1.603	1.023	0.954
11	1.602	1.587	1.014	0.939
12	1.589	1.573	1.009	0.926

Les résultats relatifs à deux estimateurs sont présentés au tableau 13. L'estimateur appelé «MELSB 24» est l'estimateur linéaire qui est fondé sur 24 mois de données et qui est le meilleur pour le niveau courant. L'estimateur désigné «Récursif 36» a été construit de la même manière que l'estimateur récursif examiné plus haut. Les effets liés au temps passé dans l'échantillon ont été inclus dans le modèle à titre d'effets fixes qui sont constants dans le temps. Toutefois, la variance des effets liés au temps passé dans l'échantillon, dans la «matrice de covariance» ayant servi à mettre à jour les estimations, a été maintenue au niveau propre à une période d'observation de 36 mois. Ainsi, l'estimateur est semblable à un estimateur lissé exponentiellement, c'est-à-dire que l'effet des observations du passé s'estompe à mesure que s'accroît l'intervalle.

La restriction selon laquelle l'espérance de l'estimateur linéaire doit être la même que celle de l'estimateur direct, fait en sorte que les variances du tableau 13 sont légèrement supérieures aux variances des estimateurs correspondants aux tableaux 10 et 11.

En exigeant que l'estimateur récursif ait la même espérance que l'estimateur direct, et en utilisant l'équivalent de 36 observations pour estimer les effets liés au temps passé dans l'échantillon, on accroît d'environ 0.5 % la

variance pour tous les estimateurs du changement visant les chômeurs, par rapport au meilleur estimateur exempt de restrictions quant au temps passé dans l'échantillon. Si seulement 24 mois de données sont utilisés dans l'estimation, la variance peut devenir jusqu'à 1.5 % supérieure à celle de l'estimateur exempt de restrictions pour les chômeurs.

Dans le cas de la population active civile, l'estimateur avec restriction fondé sur les effets liés au temps passé dans l'échantillon sur 36 périodes est de 1.3 % à 4.5 % moins efficace que l'estimateur exempt de restriction. L'estimateur du changement d'une année à l'autre utilisant seulement 24 mois de données est d'environ 14 % moins efficace que l'estimateur sans restriction du changement d'une année à l'autre.

5. SOMMAIRE

Nous avons déterminé trois sources de variation dans les estimations des caractéristiques de la population active. Les effets permanents liés aux groupes de renouvellement et aux unités primaires d'échantillonnage sont plus importants dans le cas de la population active civile et des personnes occupées que dans celui des chômeurs. L'effet transitoire des groupes de renouvellement est plus important dans le cas des chômeurs que dans celui de la population active civile. Il s'ensuit que l'utilisation de données passées procure des gains supérieurs dans l'estimation de la population active civile que dans l'estimation des chômeurs. Une méthode d'estimation des moindres carrés intégrale utilisant toute l'information passée est de beaucoup supérieure à l'estimateur composite actuel pour la population active civile. Seuls de très légers gains par rapport à la méthode actuelle sont possibles dans le cas des estimations des chômeurs. Une procédure de construction de poids fondée sur les moindres carrés est décrite; cette procédure peut servir à produire des estimateurs à convergence interne, lorsque différentes combinaisons linéaires des observations passées sont utilisées dans l'estimation des chômeurs et des personnes occupées. Pour les caractéristiques de la population active, l'inclusion des effets liés au temps passé dans l'échantillon dans l'estimation aurait des répercussions modestes sur l'efficacité des estimateurs.

REMERCIEMENTS

Cette recherche a été partiellement subventionnée par le U.S. Bureau of Labor Statistics et le U.S. Bureau of the Census, en vertu du contrat J-9-J-8-0082 et des «Joint Statistical Agreements» JSA 91-1 et JSA 91-21.

BIBLIOGRAPHIE

Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-29.

Bailar, B.A. (1978). Rotation sampling biases and their effects on estimates from panel surveys. Dans N. Krishnan Namboodiri, Ed. *Survey Sampling and Measurement*, 385-407. Academic Press, New York.

Blight, B.J.N., et Scott, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Series B35, 61-66.

Breau, P., et Ernst, L.R. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.

Cochran, W.G. (1977). *Sampling Techniques, Third Edition*, John Wiley, New York.

Duncan, G.J., et Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.

Eckler, A.R. (1987). Rotation sampling. *Annals of Mathematical Statistics*, 26, 664-685.

Fay, R.E. (1984). Some properties of estimates of variance based on replication methods. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 495-500.

Fay, R.E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 212-217.

Fuller, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

Gurney, M., et Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, 242-257.

Huang, L.R., et Ernst, L.R. (1981). Comparison of an alternative estimator to the current composite estimator in the current population survey. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 303-308.

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.

Jones, R.G. (1979). The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*, 21, 45-56.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 42, 221-226.

Jones, R.H. (1970). Recursive estimation of a subset of regression coefficients. *Annals of Mathematical Statistics*, 41, 688-691.

Kumar, S., et Lee, H. (1983). Évaluation de l'application d'estimateurs composites à l'enquête sur la population active du Canada. *Techniques d'enquête*, 9, 2, 196-221.

Odell, R.L, et Lewis, T.O. (1971). Best linear recursive estimation. *Journal of the American Statistical Association*, 66, 893-896.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *The Journal of the Royal Statistical Society, Series B*, 12, 241-255.

Plackett, R.L. (1950). Some theorems on least squares. *Biometrika*, 37, 149-157.

Raj, D. (1965). On sampling over two occasions with probability proportionate to size. *Annals of Mathematical Statistics*, 36, 327-330.

Rao, J.N.K., et Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

Sallas, W.M., et Harville, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 869.

Scott, A.J., et Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

Scott, A.J., Smith, T.M.F., et Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.

Smith, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. Dans N. Krishan Namboodiri, Ed. *Survey Sampling and Measurement*, Academic Press, New York.

Tam, S.M. (1986). Optimal prediction in stochastic regression models with application to the analysis of repeated surveys. *Australian Journal of Statistics*, 28, 345-353.

Wolter, K. (1979). Composite estimation in finite populations. *Journal of the Americal Statistical Association*, 74, 604-613.

ALLOCUTION DE CLÔTURE

Recueil du Symposium 92 de Statistique Canada
Conception et analyse des enquêtes longitudinales
Novembre 1992

ALLOCUTION DE CLÔTURE

G.J. Brackstone[1]

Nous voici arrivés au terme du Symposium 92. Nous avons entendu un grand nombre de communications de grande qualité, depuis le survol général des grandes questions fait par Graham Kalton lundi matin, jusqu'à la présentation instructive de Wayne Fuller de cet après-midi, portant sur l'estimation dans le domaine des enquêtes longitudinales.

Nous avons traité d'une vaste gamme de questions méthodologiques et d'un large éventail d'applications. Sans vouloir faire une synthèse, j'aimerais signaler quelques points que j'ai relevés au cours du symposium et qui me semblent importants. Dans la toute première session, on nous a dit que les enquêtes longitudinales ont tendance à être complexes, tout en nous invitant instamment à les garder le plus simple possible. C'est là un conseil judicieux. Le défi posé par les enquêtes longitudinales est suffisamment compliqué pour que nous évitions de lui ajouter une complexité inutile. En même temps, nous étions exhortés à ne jamais perdre de vue les objectifs de base -- un autre conseil précieux. Il semble toujours y avoir une nouvelle exigence méritant d'être incluse dans une enquête longitudinale, parce que le coût marginal d'un tel ajout apparaît peu élevé. Mais tôt ou tard ces nouvelles exigences faussent le plan et ajoutent une complexité pouvant faire obstacle à l'atteinte des objectifs de base.

On nous a dit qu'en raison de la richesse des bases de données longitudinales, leur utilisation pose des défis particuliers. Premièrement, la pleine valeur de ces bases de données ne peut être réalisée au moyen de croisement de données de plus en plus complexes; d'autres méthodes d'analyse tenant compte de la dimension temporelle des données sont nécessaires. Deuxièmement, c'est un défi que de rendre ces bases de données accessibles aux analystes tout en protégeant leur caractère confidentiel et en évitant de détruire leur richesse.

Enfin, il a été question abondamment des dossiers administratifs et de leur importance dans l'élaboration de données longitudinales, que ce soit comme source directe de données, comme complément à des données d'enquête ou comme source d'évaluation des données.

C'est à vous de juger si le symposium a été un succès. J'espère que chacun a entendu ou appris, au cours du symposium, quelque chose d'important qu'il pourra mettre en application une fois qu'il reprendra ses activités courantes. Du point de vue de la participation, le symposium a certainement été un succès. Nous avions 420 participants inscrits. Heureusement que tous n'ont pas assisté à toutes les sessions, car nous aurions eu de sérieux problèmes d'espace. Les participants venaient de neuf pays différents.

Encore une fois, nous produirons une recueil de ce symposium, qui sera envoyé à tous les participants inscrits de l'extérieur de Statistique Canada. À Statistique Canada, nous veillerons à ce que des exemplaires soient disponibles pour chacune des divisions. Il est possible de commander les recueils des symposiums précédents, ainsi que des exemplaires additionnels des recueil du présent symposium.

Je voudrais exprimer ma reconnaissance et mes remerciements à plusieurs personnes qui ont contribué à la bonne marche de ce symposium. En premier lieu, les membres du comité organisateur, soit John Armstrong, Nancy Darcovich et Pierre Lavallée, méritent nos remerciements. Ils ont planifié un programme intéressant et équilibré, et ont réussi à attirer bon nombre des praticiens les plus expérimentés et des théoriciens les plus compétents du domaine. Seuls ceux d'entre vous qui ont déjà accompli cette tâche peuvent apprécier la quantité

---

[1] G.J. Brackstone, Statisticien en chef adjoint, Secteur de l'informatique et de la méthodologie, 26-J, Immeuble R.H. Coats, Parc Tunney, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

de travail qu'exige la planification d'un tel événement. Au surplus, ces organisateurs ne peuvent profiter pleinement des fruits de leurs efforts, étant occupés à résoudre les difficultés qui se posent tout au long du symposium lui-même.

J'aimerais encore une fois remercier les organismes assumant le co-parrainage de cet événement, soit le Laboratoire de recherche en statistique et probabilité de l'Université Carleton et de l'Université d'Ottawa, ainsi que la Direction d'hygiène du milieu de Santé et Bien-être social Canada. Je tiens également à remercier Dan Krewski et Avi Singh, qui ont chacun organisé l'une des sessions de ce symposium.

Plusieurs personnes ont apporté leur contribution sur le plan de l'administration et de l'accueil : Hélène St-Jean a participé à de nombreux aspects de l'organisation; Suzanne Bonnell, Carole Jean-Marie, Carmen Lacroix, Christine Larabie et Lynn Savage ont apporté leur contribution pendant le symposium lui-même. Nous tenons à remercier spécialement Carolyn Zirbser pour les nombreuses heures passées à préparer la documentation et à veiller aux préparatifs du Symposium 92.

J'aimerais remercier également nos interprètes pour leur excellent service, ainsi que tous les conférenciers, les présidents de session et les participants qui ont permis le bon déroulement de ce symposium.

L'an prochain, nous allons déroger à la tradition. Pour notre dixième symposium, nous allons nous déplacer à Buffalo (New York). Nous assumons le co-parrainage de la Conférence internationale sur les enquêtes auprès des établissements, qui aura lieu du 27 au 30 juin 1993 à Buffalo, et qui remplacera le symposium que nous tenons habituellement ici à Ottawa. Je vous invite à participer à cette conférence, qui traitera des méthodes d'enquête sur les entreprises, les fermes et les institutions. Vous trouverez plus de détails dans les brochures qui se trouvent sur la table des documents d'information.

Le symposium est maintenant terminé. Merci de votre présence, et bon voyage.