Statistique Canada Statistics Canada

TECHNIQUES D'ENQUÊTE

UNE REVUE
DE
STATISTIQUE CANADA

VOLUME 11, NUMÉRO 2
DÉCEMBRE 1985

Canada

Statistique Canada

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

Décembre 1985

Publication autorisée par
le ministre des Approvisionnements
et Services Canada

© Ministre des Approvisionnements
et Services Canada 1986

Mai 1986
8-3200-501

Prix: Canada, $10.00, $20.00 par année
Autres pays, $11.50, $23.00 par année

Paiement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 11, n° 2

ISSN 0714-0045

Ottawa

10102580S3
STATISTICS CANADA LIBRARY

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

COMITÉ DE RÉDACTION

Président R. Platek, *Statistique Canada*

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

Rédacteurs associés
K.G. Basavarajappa, *Statistique Canada*
D.R. Bellhouse, *Université Western Ontario*
E.B. Dagum, *Statistique Canada*
J.F. Gentleman, *Statistique Canada*
G.J.C. Hole, *Statistique Canada*
T.M. Jeays, *Statistique Canada*
G. Kalton, *Université du Michigan*
C. Patrick, *Statistique Canada*
J.N.K. Rao, *Université Carleton*
C.E. Särndal, *Université de Montreal*
V. Tremblay, *Université de Montreal*

Rédacteur adjoint H. Lee, *Statistique Canada*

COMITÉ DE DIRECTION

R. Platek (Président), E.B. Dagum, G.J.C. Hole, H. Lee, C. Patrick, M.P. Singh

---

POLITIQUE DE RÉDACTION

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

Présentation de textes pour la revue

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

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 10.00$ par copie, 20.00$ par année au Canada, et de 11.50$ par copie, 23.00$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Vente et distribution des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 11, numéro 2, décembre 1985

TABLE DES MATIÈRES

M.B. WILK
Statisticiens et statisticiens 101

J.D. DREW, Y. BÉLANGER, et P. FOY
La stratification dans l'enquête sur la population active du Canada 109

D.R. BELLHOUSE
Échantillonnage des questionnaires manuscrits de recensement reproduits sur microfilm 125

B.C. SAXENA, P. NARAIN, et A.K. SRIVASTANA
Estimation du total pour deux caractères dans les enquêtes à bases de sondage multiples 135

E.B. DAGUM et M. MORRY
Désaisonnalisation des séries pour la population active en périodes de récession et de non-récession 149

E.B. DAGUM, G. HUOT, N. GAIT, et N. LANIEL
Formes de relation entre le niveau total de chômage et le nombre de bénéficiaires de l'assurance-chômage au Canada 163

L. SWAIN
Principes fondamentaux pour le développement des questionnaires 181

R.B.P. VERMA et P. PARENT
Vue d'ensemble des avantages et inconvénients des fichiers de données administratives choisis 193

R.D. SHARMA et C. WONG
Utilisation de fichiers de données administratives pour la production d'estimations de la migration: une étude de cas du fichier des permis de conduire en Ontario 203

F. AHMAD, R. CHOW, O. DEVRIES, A. HASHMI, et Y. MARCOGLIESE
Utilisation des dossiers de l'assurance-maladie de l'Alberta pour la production d'estimations régionales de la population 211

D.G. McRAE
Utilisation des comptes de l'Hydro dans le modèle d'estimation par régression de la population en Colombie-Britannique 223

D.S. O'NEIL et C.D. McINTOSH
Estimation de la répartition par âge et par sexe de la population totale des petites régions 229

R.B.P. VERMA, K.G. BASAVARAJAPPA, et R.K. BENDER
Estimation de la population par âge et par sexe 237

R.K. BENDER
Estimation de la population des petites régions: expérience de Statistique Canada... 247

Collaboration à la rédaction 252

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 101–107
Statistique Canada

Statisticiens et statisticiens[1]

MARTIN B. WILK[2]

Je suis honoré d'avoir été invité à prononcer une allocution à cette conférence annuelle de la Société statistique du Canada.

Mais à cet honneur est attachée, je le crains, la responsabilité de dire des choses intéressantes. C'est une tâche qui n'est pas facile. C'est pourquoi j'ai pensé procéder par étapes, en choisissant d'abord un titre pour mon exposé. Puis en tâchant d'expliquer le sens de ce titre. Cette explication allait constituer le fond de mon allocution. Malheureusement, je ne suis pas encore certain de ce que veut dire le titre que j'ai choisi. Mais je ne me laisserai pas arrêter par cela. Évidemment, comme disait Yogi Berra, si vous ne savez pas où vous allez, il se peut que vous n'y arriviez jamais.

Il y a beaucoup de gens à qui l'on donne le titre de statisticien et qui font des travaux très divers dont on peut dire qu'ils sont de la statistique. En fait, à diverses époques de ma carrière variée, j'ai moi-même été plusieurs sortes de statisticien. Cette observation me conduit à me demander quelles relations peuvent exister entre les différents genres de statisticien et de statistique.

On pourrait définir deux types de statistique: d'abord la statistique probabiliste, puis la production de bases de données dont s'occupent les bureaux de la statistique. Que faut-il entendre par statistique probabiliste? Sans chercher à être précis, je dirai que cette expression englobe les matières normalement traitées dans les manuels et les cours, par exemple l'analyse de variance, les tests de validité de l'ajustement, les plans d'expériences, les composantes de la variance, l'estimation par la méthode de Bayes et ainsi de suite.

Quant aux résultats des travaux des bureaux de la statistique comme Statistique Canada, le Bureau de la statistique du Manitoba et le U.S. Bureau of the Census, vous les lisez tous les jours dans les journaux.

Ces deux genres de travaux sont *considérés* comme liés entre eux et, je crois, *le sont*. On pourrait dire que la relation entre les deux est à la fois *réelle* et *imaginaire* — et je ne suis pas du tout certain de ce qu'il convient de ranger dans l'une et l'autre catégorie.

Considérons donc certaines réalités à quoi correspondent ces deux catégories, que nous appellerons la statistique théorique et la statistique pratique tout simplement pour ne pas employer des expressions comme "statisticien probabiliste".

La Société statistique du Canada semble composée surtout de théoriciens de la statistique. En effet, un sondage récent a montré qu'elle compte 66% d'universitaires et 21% de membres du personnel d'organismes gouvernementaux. Des 2,000 spécialistes que comprend la société, 32 seulement appartiennent à Statistique Canada.

Les personnes inscrites à cette conférence sont vraisemblablement pour la plupart des théoriciens qui donc s'intéressent avant tout à la statistique probabiliste. Non seulement n'y a-t-il que quelques personnes de Statistique Canada, mais j'ai dû aussi constater que les chefs de service de Statistique Canada ne se sont pas montrés très intéressés à envoyer de leur personnel à la conférence.

---

[1] Allocution prononcée à la conférence annuelle de la Société statistique du Canada.

[2] M. Martin B. Wilk, ancien statisticien en chef du Canada, est actuellement conseiller principal auprès du Conseil privé et président de la Société statistique du Canada.

Voyons maintenant des exemples de ce qui se fait comme travaux dans nos deux catégories de l'activité statistique. Le journal officiel de la Société statistique du Canada est la *Revue Canadienne de Statistique*, qui est une publication trimestrielle. L'organe officiel de Statistique Canada est *Le Quotidien*, qui a paru 256 fois l'an dernier.

Une comparaison des titres des articles publiés offre un grand intérêt. J'ai choisi au hasard dans *La Revue Canadienne de Statistique* 15 expressions clés parmi 122 qui pouvaient être représentatives des articles publiés en 1983.

Cette liste d'expressions donne une idée des sujets sur lesquels les théoriciens de la statistique lisent et écrivent.

Distributions d'abondance
Propriétés asymptotiques
Distribution centrée de Wishart
Distribution de khi carré
Valeurs limites d'acceptation
Théorie de la décision
Analyse de courbe de croissance
Filtre linéaire
Processus logistique
Études longitudinales
Modèle linéaire à plusieurs variables
Estimateur de mouvement
Séries chronologiques spatiales
Propriétés de structure
Estimation (pondérée) par les moindres carrés

Ces expressions sont dans toutes les bouches à cette conférence. Mais elles désignent des thèmes qui *ne sont pas* du tout ceux dont s'occupent les praticiens de la statistique. D'ailleurs beaucoup de praticiens, la plupart peut-être, ne connaissent pas ces expressions ou ne s'y intéressent.

Les communiqués parus dans *Le Quotidien* du 29 avril 1925 donnent une idée de la production de Statistique Canada.

- nombre total de porcs au Canada (plus de dix millions)
- nombre de tonnes d'orge exportées (plus de 150,000 en mars 1985)
- nombre de mètres carrés de laine minérale expédiés (plus de six millions)

On peut encore se faire une idée de la production de Statistique Canada au moyen du tableau des principaux indicateurs statistiques mis à jour toutes les semaines dans la publication faits saillants, distribuée aux ministres et sous-ministres. Voici quelques-uns de ces indicateurs :

Produit national brut
Mises en chantier de logements
Taux d'escompte
Taux de chômage
Hausse de l'indice des prix à la consommation
Gains hebdomadaires

Il y a aussi les mesures qui ont trait à d'autres secteurs de la société canadienne comme l'économie, les affaires, le commerce, les finances, les services sociaux et le monde du travail.

Statistique Canada produit des études sur des sujets comme le divorce au Canada, la santé des canadiens, la condition féminine, les indicateurs de conjoncture, les indicateurs de la science et de la technologie, les caractéristiques linguistiques des canadiens et ainsi de suite.

Je tiens à préciser que je ne cherche pas à porter un jugement sur la valeur relative des deux types de production statistique dont je parle ici. Ces deux genres d'activité sont utiles à la société, comme en témoigne le fait que l'une et l'autre a ses utilisateurs. Du point de vue social, ces deux types d'activité sont l'un et l'autre justifiés.

Mais ce que *je cherche à faire*, c'est analyser la relation qu'il peut y avoir entre ces deux formes d'activité qu'on appelle *statistique* dans un cas comme dans l'autre et qui sont l'occupation de spécialistes qu'on appelle tous des *statisticiens*.

Évidemment, nous pourrions nous contenter de dire qu'il s'agit d'un simple cas de polysémie, qui est le fait pour un mot d'avoir plus d'un sens. Ou nous pourrions tout simplement ignorer le problème. Mais cela ne serait ni sage, ni très positif.

Vous connaissez tous l'ouvrage classique de Kendall et Stewart sur la théorie de la statistique. Le premier volume a 396 pages de texte sans compter les tableaux et l'index. Ces 396 pages parlent de concepts théoriques de la statistique probabiliste et de la construction mathématique de formules diverses.

L'ouvrage commence par le passage suivant de O. Henry que citent les auteurs.

> "Asseyons-nous sur ce tronc d'arbre, dis-je,et oublions l'inhumanité et les ribauderies des poètes. C'est dans les glorieuses colonnes du fait positif et des mesures légales qu'il faut chercher la vraie beauté. Tenez, Mrs Sampson, dis-je, il y a des chiffres, qui surpassent en merveilles tous les poèmes, jusque dans ce tronc d'arbre sur lequel nous sommes assis. Les cercles de l'aubier montrent qu'il est mort à 60 ans. Enterré à 600 mètres de profondeur, il se transformerait en charbon d'ici 3 000 ans. La mine de charbon la plus profonde de la Terre se situe à Killingworth, près de Newcastle. Une caisse longue de 4 pieds, large de 3 pieds et haute de 2 pieds 8 pouces peut contenir une tonne de charbon. Lorsqu'une artère est rompue, comprimez-la au-dessus de la blessure. Une jambe humaine possède trente os. La tour de Londres fut brûlée en 1841.
>
> Continuez, Mr. Pratt, dit Mrs Sampson. De telles pensées sont si originales et revigorantes! J'estime que les statistiques sont aussi admirables qu'il est possible de l'être." (*Le Manuel du Mariage)*

Je trouve que ce passage est très beau. Et, naturellement, l'ouvrage est un bon exemple de clarté dans un texte savant. Mais je me demande quel rapport il peut y avoir entre le passage cité et le contenu du livre. Les auteurs y voient-ils un lien étroit? le passage, qui semble faire allusion au genre de préoccupation que peut avoir un praticien de la statistique, est-il cité pour justifier ou expliquer la superstructure de la statistique probabiliste qui se bâtit sur le travail des praticiens?

Les auteurs pensent-ils que les concepts et les formules développés dans leur traitement de la statistique probabiliste doivent guider ou fonder les travaux des praticiens et des bureaux de la statistique? Ou bien pensent-ils que la justification de la statistique probabiliste réside dans le fait que ses techniques sont utilisées dans les productions des bureaux de la statistique?

Qu'y a-t-il de *réel* et d'*imaginaire* dans ce rapport?

Il y a un peu d'énigme dans le rapport qui peut exister entre les travaux des théoriciens et des praticiens. Les choses semblent se passer comme ceci :

- les statistiques produites ont une valeur parce que leur production repose sur des méthodes reconnues;
- la méthode elle-même découle d'une théorie;
- mais la théorie statistique est formulée au moyen de concepts et de la logique mathématique, eux-mêmes fondés sur des postulats invérifiables.

Qu'est-ce qui justifie les postulats, les concepts et la théorie?

En science, de façon générale, une théorie est jugée valable par l'utilité des résultats que permet d'obtenir la technique qui en découle.

En réalité, une technique est souvent inventée sans *théorie* tandis que son utilité en fait répandre l'emploi. Le bronze et l'acier damassé ont été mis au point parce qu'ils étaient utiles, non parce qu'on avait formulé une théorie mathématique cohérente à propos de la métallurgie.

Pour juger si la statistique probabiliste est une bonne chose, nous devrions nous demander s'il en découle une technique permettant de produire des données utiles et valables.

Au lieu de cela, les statisticiens ont tendance à faire le contraire, c'est-à-dire à se demander si les travaux des praticiens sont valables à la lumière des préceptes de la statistique probabiliste.

La statistique probabiliste a produit des concepts, des modèles et des méthodes en grand nombre. Voici quelques exemples.

Prise de décision dans un contexte d'incertitude
Probabilité subjective
Science de l'inférence
Inférence de la vraisemblance
Estimation par la méthode de Bayes
Analyse de séries chronologiques
Tests d'hypothèses
Tests de signification
Estimation du niveau de confiance
Estimation d'erreurs de sondage
Méthodes de classification
Analyse de régression
Composantes de la variance
Plan d'expérience
Plan d'enquête par sondage
Estimateurs sans biais

Beaucoup d'auteurs ont dit que le concept le plus fondamental de la statistique probabiliste appliquée était l'*évaluation objective de l'incertitude.*

Mais je dois vous avouer que cette notion, si séduisante qu'elle puisse être et quelle que soit sa profondeur théorique, ne correspond pas à la réalité des travaux et du mandat des bureaux de la statistique.

Qu'il me soit permis de montrer par l'exemple l'importance du travail des praticiens. Vous pouvez faire vous-mêmes une expérience. Dressez la liste des questions qui vous semblent avoir un intérêt pour la société canadienne. Votre liste touchera à des points comme l'emploi et le chômage, le revenu des personnes âgées, la condition féminine, la croissance économique, les échanges commerciaux et la balance des paiements, la formation des familles, la distribution de la population, le déficit de l'état, et ainsi de suite.

Si vous y réfléchissez, vous constaterez que, pour la majorité de ces questions, vos impressions et vos connaissances dépendent assez directement des données statistiques produites par les praticiens, *principalement* par ceux de Statistique Canada. On pourrait faire la même constatation pour n'importe quel pays.

Pour préciser encore ma pensée par un exemple précis, j'aimerais donner ici un bref aperçu des applications possibles de l'indice des prix à la consommation.

L'indice des prix à la consommation est mis à jour chaque mois par Statistique Canada d'après des relevés mensuels des prix relatifs à un panier déterminé de biens et de services. Cet indice est le plus utilisé des indicateurs du taux d'inflation. On l'appelle souvent l'indice du coût de la vie. L'indice des prix à la consommation a un effet direct ou indirect sur presque tous les canadiens. Cet indice ou l'une ou l'autre des composantes dont il est la moyenne pondérée est utilisée pour faire des calculs ou établir des définitions dans les domaines suivants: impôt, conventions collectives, allocations familiales, pensions de sécurité de la vieillesse, contrats de location, assurances, pensions alimentaires, paiements pour enfants à charge, paiements pour enfants d'anciens combattants, remboursement de prêts étudiants. Il entre aussi en ligne de compte dans beaucoup d'autres formes de contrats et de textes de réglementation.

Mais revenons à l'évaluation objective de l'erreur, notion censément la plus importante de la statistique probabiliste. Statistique Canada ne produit pas de mesure de l'erreur d'estimation de l'indice des prix à la consommation. Nous *ne* publions *pas* d'estimations par intervalle de cet indice. Nous ne testons pas l'hypothèse de la stabilité de cet indice d'un mois à l'autre. Et nous ne produisons pas d'estimations composites dont la fonction serait de réduire la variance de l'erreur aléatoire.

On nous adresse de temps à autre des questions et des critiques sur cet état de choses, même par des gens qui ne sont ni des statisticiens, ni des scientifiques. Il semble que l'habitude de se voir communiquer des résultats de sondage d'opinion commence à donner au public le sentiment qu'une estimation d'erreur doit accompagner toute publication d'estimations. L'expression "19 fois sur 20" fait maintenant partie du vocabulaire de tous les lecteurs de journaux. Bien entendu, les sondages d'opinion existent depuis longtemps. George Gallup fait remonter l'usage de ce genre de sondage au moins jusqu'à 1824, année où un journal de Pennsylvanie publiait les résultats de ce qu'on avait alors appelé un "vote blanc pris sans considération de l'allégeance partisane". Les moyens de communication modernes et l'ordinateur ont fait proliférer les sondages. L'omniprésence des sondages a rendu le public plus conscient du fait qu'un statisticien (ou quiconque) peut effectuer une enquête par échantillonnage, faire des inférences et attribuer une mesure d'incertitude aux estimations.

Dans son rapport de 1983, le vérificateur Général du Canada posait la question de la mesure de la qualité des données produites par Statistique Canada. Le vérificateur recommandait que Statistique Canada définisse et diffuse un plus grand nombre de mesures de la qualité de ses données. L'agence fédérale a répondu officiellement que cette recommandation ne pouvait être appliquée intégralement vu l'impossibilité de produire des mesures de la qualité pour plusieurs genres de données, en particulier pour les données synthétiques. La réponse de Statistique Canada disait encore qu'il serait plus réaliste de donner une *description* complète des facteurs connus de baisse possible de la qualité et, quand il était possible d'en établir, des mesures de la qualité.

Certes, Statistique Canada publierait plus d'estimations d'erreurs si la chose était jugée possible. Il ne faut pas voir là une volonté de rejeter la possibilité de l'erreur. Comme disait à ses étudiants le professeur R.C. Bose, l'erreur est humaine, et les statisticiens sont humains.

Toutefois, les estimations d'erreurs reposent habituellement sur des hypothèses qui simplifient à outrance la situation. Par exemple, l'enquête sur la population active utilise des échantillons de ménages, non des échantillons de particuliers, qui ont les mêmes chances d'être choisis. En outre, étant donné la nature du plan de sondage, les ménages eux-mêmes n'ont pas des chances égales de faire partie de l'échantillon. La fraction de sondage est d'environ 1 sur 125, à l'échelle nationale, et peut atteindre 1 sur 24 dans les provinces à faible population. Pouvons-nous, alors, supposer que tous les particuliers sont indépendants et ont des chances égales d'être en chômage? Les données sont recueillies au moyen d'interviews, et l'interviewer comme le répondant peut commettre une erreur par inadvertance ou délibérément. Pouvons-nous ignorer toutes les sources possibles d'erreur hors l'erreur d'échantillonnage? Les personnes composant un ménage donné font l'objet d'un sondage pendant six mois consécutifs, un ménage sur six étant chaque mois retiré de l'échantillon tandis qu'un autre y est introduit pour le remplacer. Ainsi, dans n'importe quel mois donné, les répondants n'ont pas tous rempli le questionnaire le même nombre de fois. Pouvons-nous supposer que les six réponses sont indépendantes dans le temps? Pendant les six mois de sondage, un logement peut changer d'occupants. Et, naturellement, il y a les problèmes habituels de non-réponse, d'observations aberrantes, d'erreurs d'entrée des données, de calcul, d'impression, etc. Le problème des écarts par rapport à l'hypothèse "normale" de la théorie statistique reste un sujet de constante préoccupation pour certains théoriciens et praticiens de Statistique Canada.

D'une part, donc, la statistique probabiliste a produit la notion utile et importante d'estimation objective de l'erreur; le public a été initié à cette notion et s'attend à la voir appliquée.

D'autre part, il y a beaucoup d'informations statistiques très utilisées qui sont produites par des gens qu'on appelle des statisticiens, mais pour lesquelles des mesures de l'erreur ne sont ni ne peuvent actuellement être données.

Théoriquement, il semble y avoir plusieurs possibilités.

a) on pourrait ne plus englober dans la science statistique, et l'activité des bureaux de la statistique et celle des statisticiens probabilistes, et abandonner le principe d'une relation entre les deux.
b) la statistique probabiliste pourrait orienter ses travaux dans le sens de la recherche d'une technique adaptée à la réalité de l'élaboration de données complexes.
c) les statisticiens pourraient s'employer à rééduquer le public de façon qu'il cesse de croire à la possibilité de mesures nettes et objectives de l'incertitude.

Dans la pratique cependant, seulement la possibilité b) peut être envisagée. De fait, elle serait plus fructueuse pour tous ceux dont la statistique est le métier.

Dans un article paru dans la revue *Science* l'an dernier, un spécialiste de la philosophie des sciences, Ian Hacking, affirmait que les statisticiens avaient discrètement transformé le monde dans lequel nous vivons, non par la découverte de faits nouveaux ou par des progrès techniques, mais en changeant notre manière de raisonner, de faire des expériences et de nous former ensuite une opinion.

Cela fait plaisir de lire un jugement comme celui-là sur l'importance de la statistique probabiliste telle qu'en ont jeté les bases des gens comme Fisher, Neyman, Pearson, Wald et d'autres.

Mais j'aimerais souligner, pour en revenir au thème principal de mon exposé, qu'il y a une autre catégorie de statisticiens — les praticiens des bureux de la statistique — qui ont joué un rôle dans cette transformation discrète du monde, encore que d'une manière qui est l'opposé du point de vue de M. Hacking.

Les praticiens de la statistique découvrent de nouveaux faits, définissent de nouvelles notions et élaborent des définitions opératoires dont les applications peuvent servir au public.

Ils font effectivement des découvertes techniques dans le domaine du calcul, de la diffusion électronique de l'information, de l'infographie, des systèmes de classification, des systèmes de comptabilité nationale, etc.

Encore une fois, mon intention n'est pas ici de faire, implicitement ou explicitement, un jugement sur la valeur relative de deux types d'activité. Ce que je me demande, c'est ce qu'il y a de *réel* et d'*imaginaire* dans le rapport entre la statistique pratique et la statistique théorique.

La plupart des choses que font les praticiens ne découlent pas, en réalité, des constructions mathématiques, des théories et des présupposés de la statistique probabiliste, ni n'ont avec eux un lien immédiat. Et pourtant les praticiens se sentent d'une certaine façon obligés de faire semblant d'accorder de l'importance à un lien fondamental qu'auraient leurs travaux avec les concepts théoriques.

Par ailleurs, les théoriciens continuent de penser plus ou moins confusément que si de plus nombreux praticiens avaient des préoccupations théoriques, alors la statistique probabiliste pourrait avoir un effet *réel* sur les travaux des bureaux de la statistique.

Le potentiel que peut receler une collaboration plus efficace des praticiens et des théoriciens ne sera pas libéré sans un effort des uns et des autres.

Et ce n'est pas moi qui peux ce soir proposer des solutions qui auraient le caractère de révélations.

Cependant, il faut évidemment établir de meilleurs canaux de communication. Statistique Canada a créé en ce sens un programme de bourses de perfectionnement et de stage.

Également dans cet esprit, Statistique Canada a constitué un réseau de comités consultatifs, dont un sur la méthodologie statistique.

Statistique Canada a retenu les services de plusieurs statisticiens probabilistes comme consultants.

Je crois qu'il y a beaucoup de possibilités que pourraient exploiter Statistique Canada et les universités en organisant des ateliers mixtes et en collaborant à divers projets.

Il faudrait une meilleure compréhension mutuelle chez les praticiens et les théoriciens. Peut-être conviendrait-il de changer les critères et les normes ayant trait à la publication des données?

Peut-être faut-il modifier les bases sur lesquelles sont évaluées les demandes de subventions adressées au conseil de recherches en sciences naturelles et en génie du Canada?

Peut-être qu'il serait utile de modifier les programmes de formation. Peut-être que Statistique Canada devrait offrir un prix pour la mise au point de solutions productives dans des secteurs d'exploitation des bureaux de la statistique où le besoin s'en fait particulièrement sentir. Peut-être faudrait-il avoir une liste permanente des dix solutions les plus nécessaires; cette liste serait un stimulant pour les statisticiens probabilistes et un moyen de rester en contact avec eux.

Peut-être la société Statistique du Canada devrait-elle faire en sorte de créer une tradition selon laquelle l'allocution qui suit le dîner de la conférence annuelle aurait toujours pour thème le rapport entre statisticiens et statisticiens?

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 109–124
Statistique Canada

La stratification dans l'enquête sur la population active du Canada

J.D. DREW, Y. BÉLANGER, et P. FOY[1]

RÉSUMÉ

On décrit l'utilisation d'un algorithme de classification multi-variée pour effectuer la stratification pour l'enquête sur la population active. L'algorithme élaboré par Friedman et Rubin (1967), est modifié de manière à traiter la formation de strates géographiquement contiguës, et à faire la délimitation d'unités primaires d'échantillonnage (U.P.É.) hétérogènes mais compactes à l'intérieur des strates. Des études portant sur les variables de stratification, la robustesse de la stratification au fil du temps, et le type de stratification sont décrites.

MOTS CLÉS: Algorithme de classification multi-variée; stratification géographique; enquête permanente.

1. INTRODUCTION

L'enquête sur la population active du Canada (E.P.A.) est remaniée après chaque recensement décennal de la population et des logements. Dans le cadre du remaniement suivant le recensement de 1981, on a mené un programme intensif de recherche sur divers aspects du plan de sondage (Singh, Drew, et Choudhry 1984). Ce rapport décrit la partie du programme de recherche portant sur les méthodes de stratification.

Étant donné que l'E.P.A. est utilisée, non seulement pour fournir de l'information sur les caractéristiques de la population active, mais aussi comme plan de sondage général pour diverses autres enquêtes-ménages, l'un des principaux objectifs du remaniement était d'accroître la flexibilité de l'E.P.A. pour des applications générales. La stratification a été considérée comme un moyen d'améliorer l'efficacité du plan d'échantillonnage en ce qui a trait aux applications générales, de même que pour des variables d'intérêt particulier pour l'E.P.A., en adoptant des méthodes plus rigoureuses que dans l'ancien plan.

On a donc décidé de considérer l'utilisation d'algorithmes de classification multi-variée pour stratifier et mettre en grappes et de les comparer avec les méthodes utilisées dans l'ancien plan de sondage. On a choisi un algorithme non-hiérarchique élaboré par Friedman et Rubin (1967), en nous fiant aux résultats d'évaluations des différents algorithmes faites par Judkins et Singh (1981) dans le cadre de remaniement du Current Population Survey du U.S. Bureau of Census. La description de l'algorithme de base, et des extensions que nous avons développées se retrouve dans la section 2.

Les sections 3 et 4 décrivent les études d'évaluation et la stratification finalement adoptée dans les deux grands types de secteurs du plan de sondage de l'E.P.A., soient les Unités Non Auto Représentatives (U.N.A.R.) et les Unités Auto Représentatives (U.A.R.). La section 4 décrit également comment on a adapté l'algorithme pour délimiter les unités primaires d'échantillonnage à l'intérieur des strates de la partie N.A.R.

On conclut dans la section 5 avec quelques observations sur la possibilité d'adapter le système développé à d'autres applications.

[1] J.D. Drew et Y. Bélanger, Division des méthodes de recensement et d'enquêtes-ménages, et P. Foy, Division des méthodes d'enquête pour les entreprises, Statistique Canada, Ottawa, (Ontario), K1A 0T6.

2. ALGORITHME DE STRATIFICATION

L'algorithme de base utilisé pour la stratification est un algorithme multi-varié, non-hiérarchique élaboré par Friedman et Rubin (1967). Ce choix repose sur les résultats d'études faites par Judkins et Singh (1981) et Kostanich, Judkins, Singh et Schantz (1981), qui ont évalué plusieurs algorithmes de stratification pour le Current Population Survey du U.S. Bureau of the Census.

Ces derniers ont modifié la fonction objective de l'algorithme dans le cas de l'échantillonnage avec probabilité proportionnelle à la taille (PPT), et nous avons ajouté la capacité de formuler des strates compactes et contiguës. Une description plus complète de ce qui suit se retrouve dans Foy (1984).

2.1 La fonction objective de l'algorithme

L'algorithme vise à partitionner les unités de stratification (secteurs de dénombrement du recensement) dans des strates les plus homogènes possible à l'égard de plusieurs variables d'intérêt, c'est-à-dire en minimisant les sommes des carrés à l'intérieur de chaque strate.

Les expressions pour les sommes des carrés dans le cas de l'échantillonnage avec PPT suivent après l'introduction de la notation utilisée:

L = nombre de strates à former,

N = nombre total d'unités (secteurs de dénombrement),

N_k = nombre d'unités dans le groupe (strate) k; $(N_1 + N_2 + \ldots + N_L = N)$,

T_{jk} = mesure de la taille de l'unité j du groupe k,

$T_{.k}$ = mesure de la taille du groupe k,

$T..$ = taille totale,

${}_iX_{jk}$ = valeur observée de la variable i pour l'unité j du groupe k,

${}_iX_{.k}$ = total des valeurs observées de la variable i dans le groupe k,

${}_iX..$ = total des valeurs observées de la variable i,

W_i = facteur de pondération de la variable i (voir section 2.4 pour plus de détails),

p = nombre de variables d'intérêt.

Ainsi, l'expression de la somme des carrés totale, avec PPT, de la variable i est donnée par

$$SCT_i = \sum_{k=1}^{L} \sum_{j=1}^{N_k} \frac{T_{jk}}{T..} \left(\frac{T..}{T_{jk}} {}_iX_{jk} - {}_iX.. \right)^2.$$

Ceci est aussi l'expression de la variance de l'estimateur de ${}_iX..$ lorsqu'une unité est choisie avec PPT. La somme des carrés totale pondérée sur toutes les variables est donc

$$SCT = \sum_{i=1}^{p} W_i SCT_i.$$

Quant aux sommes des carrés intra-groupes et inter-groupes, elles sont obtenues respectivement par les expressions suivantes:

$$SCW_i = \sum_{k=1}^{L} \frac{T..}{T_{.k}} \sum_{j=1}^{N_k} \frac{T_{jk}}{T_{.k}} \left(\frac{T_{.k}}{T_{jk}} {}_iX_{jk} - {}_iX_{.k} \right)^2$$

et

$$SCB_i = \sum_{k=1}^{L} \frac{T_{.k}}{T_{..}} \left(\frac{T_{..}}{T_{.k}} {}_iX_{.k} - {}_iX_{..} \right)^2.$$

Leurs sommes des carrés pondérées sur toutes les variables sont données respectivement par

$$SCW = \sum_{i=1}^{p} W_i SCW_i$$

et

$$SCB = \sum_{i=1}^{p} W_i SCB_i$$

La somme des carrés intra-groupes de la variable i, SCW_i, est aussi l'expression de la variance de l'estimateur de ${}_iX_{..}$ lorsqu'une strate est choisie avec PPT et subséquemment une unité de cette strate est choisie avec PPT.

Comme d'habitude, nous avons le résultat suivant:

$$SCT_i = SCW_i + SCB_i, \ (i = 1, \ldots, p)$$

et

$$SCT = SCW + SCB.$$

La fonction objective du programme de stratification se trouve à être SCW, la somme des carrés intra-groupes pondérée sur toutes les variables. Elle doit être minimisée. On définit l'indice de stratification associé à la variable i, I_i, comme étant:

$$I_i = 100 \times \frac{SCB_i}{SCT_i} \qquad i = 1, \ldots, p$$

Une valeur élevée de l'indice indique une bonne stratification.

2.2 Recherche de la meilleure classification

Afin d'identifier la meilleure clasification, une méthode consisterait à générer toutes les partitions posibles de N unités en L groupes. Il suffirait ensuite de retenir celle qui minimise la fonction objective. Cette tactique est rarement faisable puisque le nombre de partitions possible peut être exagérément grand.

Friedman et Rubin (1967) suggèrent l'algorithme suivant. Commençons d'abord avec une partition quelconque des N unités en L groupes. Considérons maintenant le transfert d'une unité à un groupe autre que celui auquel elle appartient. Cette unité sera transférée au groupe qui apportera la plus forte réduction de la fonction objective. Toutefois, l'unité demeurera dans son groupe original si aucun transfert n'apporte de réduction. En se servant maintenant de la partition ainsi engendrée, traitons la deuxième unité de la même manière, ensuite la troisième, jusqu'à la $N^{ième}$. L'application de cette procédure à chaque unité devient une itération que les auteurs appellent "hill-climbing pass". Après plusieurs itérations, l'algorithme atteint un point où aucun transfert de quelqu'unité que ce soit ne permet de réduire la fonction objective. Ce point est dit un minimum local de la fonction objective parce qu'il dépend de la partition initiale utilisée. Une autre partition initiale aurait pu atteindre un autre point

où la valeur de la fonction objective est encore plus faible. Pour aller au-delà du minimum local, Friedman et Rubin décrivent deux procédures, soient le "forcing pass" et le "reassignment pass". En appliquant leur algorithme aux données décrites dans leur article, ils obtiennent la plus haute valeur connue de la fonction objective 10 fois sur 14 tentatives en utilisant différentes partitions initiales (ils utilisent une autre fonction objective qui est maximisée). Avec des données moins bien structurées, la plus haute valeur est atteinte 3 fois sur 11, sans toutefois l'assurance d'avoir obtenu la solution optimale. À leur avis, les méthodes du "forcing pass" et du "reassignment pass" ne sont utiles qu'à l'occasion. Ils ont plus confiance aux résultats obtenus en utilisant plusieurs partitions initiales. Ce point de vue est soutenu par Judkins et Singh (1981). On a donc décidé d'utiliser la technique de plusieurs partitions initiales.

Parce que l'algorithme ne déplace qu'une unité à la fois, le calcul de la fonction objective en est simplifié. Après le calcul initial de la fonction objective, il suffit de recalculer la contribution à la fonction objective des deux groupes impliqués dans le transfert de l'unité en jeu.

2.3 La contiguïté

Dans les plans de sondage antérieurs de l'E.P.A., on avait adopté des strates formées d'unités géographiques contiguës, c'est-à-dire que chaque unité d'une strate donnée devait toucher à au moins une autre unité de la même strate. Une des raisons principales était une présomption que de telles strates préserveraient l'efficacité du plan d'échantillonnage sur une plus longue période de temps que dans le cas des strates formées d'unités non-contiguës.

En vue d'évaluer cette présomption et d'adopter la meilleure stratification possible, nous avons considéré deux moyens de tenir compte de la géographie dans la stratification. La première méthode est élaborée par Dahmström et Hagnell (1978), et consiste en l'utilisation de centroïdes comme variables d'intérêt. Cette méthode fait appel à deux variables géographiques (centroïdes) qui sont des transformations de longitude et de latitude. Cette méthode donne des strates compactes, c'est-à-dire des strates où la distance entre les unités est rendue minimale par la minimisation de la somme des carrés intra-groupe usuelle des centroïdes. Cette minimisation est tempérée par la minimisation des autres variables d'intérêt. Aussi il n'y a pas de garantie que les strates soient ainsi formées d'unités contiguës.

L'autre méthode, que nous appelons l'approche de vecteurs de contiguïté, est nouvelle. Elle garantie des strates contiguës, mais pas nécessairement compactes. Des études décrites dans la section 3, ont porté sur l'utilisation de l'une ou l'autre des méthodes seulement ou d'une combinaison des deux méthodes.

2.3.1 Vecteurs de contiguïté

Afin d'assurer la formation de strates contiguës, on a procédé de la façon suivante. On fait l'optimisation comme décrite dans la section précédente mais en commençant dans ce cas avec une partition initiale qui est contiguë, et en permettant le déplacement de l'unité "j" de la strate A à la strate B, seulement si en plus de réduire les sommes des carrés, les conditions suivantes sont respectées:

(i) l'unité j est contiguë à une unité de la strate B

(ii) le déplacement de l'unité j à la strate B ne dérangera pas la contiguïté de la strate A.

Pour vérifier ces deux conditions, il est essentiel de connaître les liens de contiguïté entre les unités. Par conséquent, à chaque unité doit être assigné un vecteur de contiguïté qui contient la liste des unités qui y sont contiguës.

La première condition est facile à vérifier. Pour s'assurer que l'unité "j" est contiguë à une unité de la strate B, il suffit de trouver dans son vecteur de contiguïté une unité qui appartient à la strate B.

La deuxième condition est plus difficile à vérifier. Le principe est qu'une strate est dite contiguë si chaque paire d'unités de cette strate peut être reliée par une chaîne continue d'unités provenant de cette même strate. Soit l'unité j que l'on veut déplacer de la strate A à la strate B. Il est donc nécessaire de chercher, pour chaque paire d'unités du vecteur de contiguïté de l'unité j appartenant à la strate A, un autre lien parmi les unités de la strate A. À ce stage, le problème se compare à celui de trouver son chemin dans un labyrinthe.

Un algorithme a aussi été conçu pour créer de façon aléatoire des partitions initiales contiguës.

2.4 Pondération des variables

Les facteurs de pondération revêtent une importance toute particulière. Ce sont eux qui établissent l'apport de chaque variable à la classification.

Il est habituellement préférable de standardiser les variables, en rendant les facteurs de pondération inversement proportionnels à la somme des carrés totale de chaque variable. Cette standardisation permet d'obtenir un apport comparable de chaque variable à la classification.

Si après la standardisation on veut qu'une ou plusieurs variables se voient accorder plus d'importance relativement aux autres variables dans l'optimisation, on peut le faire en spécifiant un poids supérieur à 1 (la normale). Par exemple, une variable avec un poids de 2 aurait une importance double. Comme décrit dans la section 3.2 on a dû essayer plusieurs combinaisons de poids pour les variables à caractère géographique et non-géographique dans le but d'obtenir des strates compactes sans trop affecter la minimisation des autres variables.

3. LA STRATIFICATION DANS LES UNITÉS NON AUTOREPRÉSENTATIVES

3.1 Ancienne conception (Platek et Singh 1976)

Pour l'E.P.A., chacune des dix provinces canadiennes est divisée en un certain nombre de régions économiques (R.É.), composées de secteurs ayant des structures économiques semblables. Les frontières des R.É. sont déterminées en consultation avec les provinces. Ces R.É. sont considérées comme des strates primaires. L'étape de stratification suivante est la partition de chaque R.É. en unités autoreprésentatives (U.A.R.) et unités non autoreprésentatives (U.N.A.R.). Les unités autoreprésentatives sont les villes où l'échantillon prévu est assez important pour constituer au moins une tâche d'interviewer; la partie N.A.R. comprend le reste de la R.É. Des plans d'échantillonnage différents sont suivis dans les U.A.R. et U.N.A.R., parce que la population dans les U.N.A.R. est beaucoup plus dispersée, rendant nécessaire un plus grand nombre d'étapes d'échantillonnage. Pour ces mêmes raisons, on retient le concept des U.A.R. et U.N.A.R. dans le remaniement.

Dans l'ancien plan, on a stratifié la partie N.A.R. de chaque R.É. en un maximum de 5 strates contiguës ayant une population entre 36,000 et 75,000 personnes, par rapport aux caractéristiques principales de la population du recensement de 1971, comme décrit ci-dessous, et comme élaboré en plus de détail par Platek et Singh (1976).

La population active a été répartie en 7 catégories selon l'industrie. Dans chaque R.É., on a choisi les 3 catégories les plus importantes selon certains critères spécifiques. Les municipalités groupées, qui représentent les régions géographiques comprises dans une municipalité rurale, et qui, par conséquent, renferment souvent des municipalités urbaines géographiquement plus petites, ont été les unités de stratification. En comparant, pour chacune de ces unités, les proportions de la population active appartenant à chacune des trois catégories, avec les proportions correspondantes au niveau de la R.É., on a pu identifier les unités montrant une certaine ressemblance entre elles pour les regrouper dans des strates. Cette comparaison a été effectuée visuellement à l'aide de graphiques. En général, il a fallu procéder à certains rajustements pour satisfaire les exigences relatives à la taille et à la contiguïté des strates.

À l'intérieur de chaque strate, de 12 à 15 U.P.É. ont été formées, de façon à être semblables à la strate relativement aux variables de stratification, et relativement au rapport entre la population rurale et urbaine. Les parties rurales des U.P.É. ont été formées de S.D. contigus, et les parties urbaines ont été choisies de façon à être aussi près géographiquement que possible de la partie rurale. Les tailles des strates et des U.P.É. ont été déterminées de sorte qu'avec deux U.P.É. sélectionnées par strate, la taille de l'échantillon prévue soit équivalente à une tâche d'interviewer. Suivant ces critères, selon la province, la population des U.P.É. variait entre 3,000 et 5,000 personnes. À l'intérieur des U.P.É., l'échantillonnage a été fait en 2 ou 3 étapes.

3.2 Les études sur la stratification au moment du remaniement

Le but de nos études était de tirer des conclusions permettant de prendre des décisions relatives aux aspects suivants de la stratification: variables à utiliser, type de strates (entièrement rurales, entièrement urbaines ou mixtes), et importance à accorder à la contiguïté. Étant donné le temps très limité pour les études avant le moment de la formation des nouvelles strates et U.P.É., et l'attente générale que des strates formées d'unités contiguës seraient préférables au fil du temps à des strates formées d'unités non-contiguës, les deux premiers aspects ont été jugés prioritaires.

En ce qui concerne la contiguïté, on a dû expérimenter pour trouver le meilleur moyen de l'atteindre, soit par les vecteurs de contiguïté, les centroïdes ou une combinaison des deux. Cependant, après le remaniement, on a entrepris une étude plus approfondie examinant le choix entre des strates contiguës et des strates non-contiguës.

3.2.1 Étude sur les variables et le type de stratification

Une contrainte de la méthode de stratification utilisée dans l'ancien plan de sondage était le nombre limité de variables de stratification que l'on pouvait considérer (3 par R.É.).

Avec le nouvel algorithme, on s'est libéré de cette contrainte. En plus des sept variables d'industrie, on a voulu voir l'effet causé par l'utilisation de variables reliées au sujet de l'enquête, telles que: taux de personnes occupées, taux de chômage et niveau de revenu, ainsi que par des caractéristiques telles que: niveau d'éducation, type de logement et nombre total d'individus. Ces dernières caractéristiques se sont révélées très efficaces dans des études semblables menées par le U.S. Bureau of the Census pour le Current Population Survey.

Le tableau 1 décrit les diverses options étudiées en ce qui a trait au choix des variables.

En ce qui concerne le type de stratification, on a décider d'étudier l'effet de former des strates séparément pour les parties rurales et urbaines des R.É., comme une alternative à la méthode mixte de l'ancien plan.

Les contraintes du plan de sondage selon lesquelles les U.P.É. doivent avoir des populations presque égales, tandis que le rapport entre la population rurale et la population urbaine doit demeurer presque le même pour chaque U.P.É., ont résulté par le passé en un manque de contiguïté entre les parties rurales et urbaines des U.P.É. Ceci a mené à une érosion de la présumée correspondance entre l'U.P.É. et la tâche d'intervieweur. On a pensé que les strates séparées en parties rurales et urbaines, qui pourraient être sous-stratifiées d'une façon optimale, constituaient une solution possible à ce problème.

L'étude a porté sur 11 régions économiques réparties à travers tout le Canada. On a créé les strates en utilisant des données du recensement de 1971, et on les a évaluées en utilisant les données du recensement de 1981. Pour faire la stratification, on a choisi comme unité de stratification les secteurs de dénombrement du recensement de 1971, sauf au Québec et en Ontario. Pour ces deux provinces on a choisi les sous-divisions de recensement, car le grand nombre de S.D. dans certaines R.É. (jusqu'à 400) aurait occasionné des coûts d'exécution des programmes informatiques trop élevés.

Tableau 1

Options de stratification selon les variables

Variables	Option de stratification				
	1	2	3	4	5
Industries (7)[a]	x	x	x	x	x
Revenu		x	x	x	x
Employés		x	x		x
Chômeurs		x	x[b]		x
Démographie (2)[c]				x	x
Logement (4)[d]				x	x
Éducation (1)[e]				x	x

[a] nombre de personnes employées dans les secteurs agriculture, forêt et pêche, mines, manufacture, construction, transport, services.
[b] poids double pour nombre de chômeurs.
[c] population 15-24 ans, population 55 ans et plus.
[d] ménages à 1 personne, ménages à 2 personnes, logements possédés, loyer brut total.
[e] personnes ayant une éducation secondaire.

On a utilisé une filière de conversion entre les unités géographiques des deux recensements, pour faire l'évaluation basée sur le recensement de 1981. Les indices basés sur les données de 1981 ont été jugés plus au point pour l'évaluation, étant donné qu'en réalité l'âge moyen des données de stratification sera de 7 ou 8 ans pendant la vie du plan de sondage. On retrouve dans le tableau 2 les indices basés sur les deux recensements.

Pour cette étude on a choisi de former des strates contiguës et compactes, en utilisant les vecteurs de contiguité et les variables centroïdes avec des poids moyens de trois (voir sous-section 3.2.2). Le nombre de strates à former par R.É. était le même pour toutes les options.

On a tiré les conclusions suivantes des résultats du tableau 2:

Type de Stratification: La stratification rural/urbain était de beaucoup supérieure à la stratification globale dans le cas de la variable *agriculture*, ce qui n'est pas étonnant. Le même phénomène s'est produit pour la variable *manufacture*, quoique moins spectaculaire. Pour la variable *revenu* la stratification rural/urbain était également meilleure, au départ mais elle n'était pas très robuste (i.e., l'indice se détériore avec le temps). La stratification rural/urbain était meilleure pour la variable *chômeurs*, tandis qu'il n'y avait pas beaucoup de différence pour *employés*.

Variables de stratification: L'option 4, en combinaison avec la stratification rural/urbain était nettement supérieure pour la variable *chômeurs*. En ce qui concerne les autres variables, l'option 5 était légèrement meilleure que les autres pour *employés* et *revenu*.

Tableau 2
Indices de stratification pour les 5 options

Variables de stratification	Global		Rural/urbain	
	1971	1981	1971	1981
		Chômeurs		
7 industries	5.4	0.1	9.9	3.8
7 industries + revenu + employés + chômeurs	5.2	2.3	10.2	3.4
7 industries + revenu + employés + chômeurs × 2	7.4	2.2	10.2	5.3
17 variables	6.3	6.4	11.3	4.7
15 variables (sauf employés + chômeurs)	3.6	0.1	9.8	9.0
		Employés		
7 industries	2.9	0.5	8.9	4.8
7 industries + revenu + employés + chômeurs	8.8	2.7	8.6	3.2
7 industries + revenu + employés + chômeurs × 2	9.1	2.8	13.1	2.2
17 variables	14.1	7.8	12.2	6.4
15 variables (sauf employés + chômeurs)	6.3	1.6	11.4	3.7
		Revenu		
7 industries	7.4	5.7	18.9	9.5
7 industries + revenu + employés + chômeurs	11.2	6.8	22.1	5.9
7 industries + revenu + employés + chômeurs × 2	10.3	6.8	28.3	9.5
17 variables	10.5	9.4	24.4	11.9
15 variables (sauf employés + chômeurs)	21.0	5.3	28.9	4.5
		Agriculture		
7 industries	7.4	9.7	37.0	26.0
7 industries + revenu + employés + chômeurs	7.6	7.8	40.0	28.7
7 industries + revenu + employés + chômeurs × 2	8.6	7.9	43.2	31.0
17 variables	6.1	1.1	40.3	31.8
15 variables (sauf employés + chômeurs	7.0	0.4	42.7	29.0
		Manufacture		
7 industries	14.7	8.5	16.9	13.2
7 industries + revenu + employés + chômeurs	10.9	6.6	16.5	12.1
7 industries + revenu + employés + chômeurs × 2	5.5	4.3	14.8	16.1
17 variables	12.5	13.5	13.3	10.7
15 variables (sauf employés + chômeurs	7.2	1.4	14.1	16.4

3.2.2 Étude sur la contiguïté

Comme mentionné auparavant, on a décidé de retenir l'idée des strates contiguës pour l'E.P.A. Des telles strates devraient être meilleures pour la production d'estimations pour les petites régions, parce que l'échantillon sera bien dispersé au point de vue géographique. De plus on a pensé que les strates contiguës préserveraient davantage l'efficacité du plan d'échantillonnage sur une longue période de temps.

La question a été ensuite de savoir comment utiliser les centroïdes ou les vecteurs de contiguïté, ou les deux ensemble, pour obtenir des strates compactes et contiguës sans que les contraintes géographiques ne dominent trop la minimisation des autres variables.

L'étude a été réalisée avec les mêmes 11 régions économiques. Comme prévu, l'utilisation des seuls vecteurs de contiguïté a résulté en des strates contiguës, mais montrant des configurations souvent irrégulières. D'un autre côté, l'utilisation des centroïdes seulement, même si des poids très élevés leur ont été associés, n'a donné aucune garantie de contiguïté absolue. En faisant varier les poids des centroïdes, comparativement aux autres variables, on a trouvé que l'utilisation des centroïdes avec poids égal à 3 et des vecteurs de contiguïté ont fourni un bon compromis entre la compacité et l'optimisation non-géographique.

3.3 Nouvelle conception: stratification

À la lueur des résultats décrits précédemment et aussi en tenant compte des résultats supérieurs d'un plan d'échantillonnage utilisant une stratification rural/urbain tirés d'une étude sur les variances et coûts (Choudhry, Lee et Drew 1985), on a décidé que dans la mesure du possible, on devait procéder à une stratification séparée pour toutes les régions économiques. Des restrictions s'appliquent cependant lorsque dans une R.É. la population rurale, d'une part, ou la population urbaine, d'autre part, n'est pas suffisante pour former au moins une strate. On a déterminé qu'une strate devait pouvoir donner un échantillon d'au moins 90 logements, correspondant à la sélection de deux U.P.É. d'un rendement minimal de 45 logements chacune. Dans le cas où cette contrainte ne pouvait être respectée, on a choisi de procéder à une stratification globale et donc de former des strates mixtes formées de S.D. ruraux et urbains. Ce critère a mené à l'adoption des strates séparées dans plus de 2/3 des R.É.

En ce qui concerne les variables de stratification, on a choisi comme compromis une stratification basée sur les 15 variables de l'option 4 plus *employés*. On a ajouté *employés* parce que son inclusion dans l'option 4, comparativement à l'option 5, a amélioré la performance des deux caractéristiques *employés* et *revenu*. Suivant la même logique on a exclu *chômeurs* comme variable de stratification.

Pour les contraintes géographiques, on a décidé d'utiliser les vecteurs de contiguïté en combinaison avec un poids uniforme de centroïdes égal à 3 dans toutes les régions économiques.

On devait également prendre une décision à propos du nombre de strates par R.É.. En pratique, dans la plupart des cas on n'a pas eu le choix. Selon le plan d'échantillonnage, chaque U.P.É. correspond à une tâche d'interviewer, et on désirait sélectionner au moins deux U.P.É. par strate afin de permettre l'estimation de la variance sans biais. Étant donné ces contraintes, dans près de 2/3 des cas on n'a formé qu'une seule strate avec 2 ou 3 U.P.É. sélectionnées, dans les parties urbaines, rurales ou une combinaison des deux. Dans les autres cas, on a stratifié de façon à sélectionner également 2 ou 3 U.P.É. par strate. On s'est basé sur une autre étude montrant de légères réductions dans la variance pour cette façon de procéder, comparativement à l'ancien plan de sondage où on sélectionnait de 4 à 6 U.P.É. par strate (Choudhry, Lee, et Drew 1985).

3.4 Étude sur la robustesse des strates contiguës et non-contiguës

Des strates robustes sont des strates qui maintiennent l'efficacité du plan d'échantillonnage au fil du temps. Après le remaniement on a entrepris une étude pour voir si les strates contiguës seraient plus robustes, comme on en a fait l'hypothèse.

L'étude a porté sur trois régions économiques de l'Ontario, les R.É. 520, 540 et 580 (selon la numérotation de 1981). Pour chacune de ces régions, on a comparé les résultats de la nouvelle stratification (choisie pour le remaniement de l'E.P.A.) qui consiste en des strates contiguës, avec une stratification sans contrainte de contiguïté. On a stratifié en utilisant les données de 1981, et on a fait l'évaluation avec les données de 1971. Pour les strates contiguës on a utilisé les vecteurs de contiguïté avec centroïdes, tandis que pour les strates non-contiguës on a essayé deux options donnant des poids de 0 et 3 aux centroïdes, respectivement. Les variables de stratification utilisées ont été les mêmes 16 variables que l'on a déjà décrites (soit l'option 4 modifiée).

Les résultats se retrouvent dans le tableau 3. On voit qu'en général l'indice global calculé lors de la stratification est plus élevé pour les deux options où la contiguité n'est pas nécessaire, comme on pouvait s'y attendre (colonne 1981). Cependant, ces deux options donnent également des indices plus élevés au fil du temps (colonne 1971).

A-t-on vraiment besoin de strates contigües? Pour répondre à cette question, on devrait entreprendre une étude plus approfondie comportant des R.É. de plusieurs provinces. L'évaluation de la robustesse de la stratification poserait alors certains problèmes. Il est facile d'évaluer la robustesse en Ontario car la stratification y est faite au niveau des sous-divisions de recensement dont les frontières n'ont à peu près pas changé depuis 1971. À l'opposé, lorsque la stratification est faite au niveau des secteurs de dénombrement, qui ont des frontières changeantes d'un recensement à l'autre, il est très difficile d'obtenir des chiffres comparables entre recensements en ce qui concerne la robustesse, surtout lorsque les strates ne sont ni compactes ni contiguës.

S'il devait s'avérer que la stratification sans contiguïté est plus optimale, cela pourrait compenser pour les problèmes éventuels rencontrés dans l'estimation pour les petites régions. Cela pourrait également ouvrir de nouveaux horizons: en effet, une fois libérés des contraintes de contiguïté, pourquoi ne pas former tout d'abord les U.P.É., compactes mais pas nécessairement contiguës, pour les regrouper seulement ensuite en des strates? Encore une fois, on ne peut répondre qu'en procédant à de nouvelles études plus approfondies.

3.5 Formation des U.P.É.

On a modifié l'algorithme de classification pour effectuer la formation des U.P.É. dans les strates rurales et mixtes. Dans les strates rurales en particulier, la formation des U.P.É. ressemble beaucoup à la stratification, au point de vue conception. La seule différence tient au fait que dans la stratification, on veut minimiser la somme des carrés des variables géographiques et non-géographiques à l'intérieur de chaque strate, tandis que dans la formation des U.P.É. on veut minimiser la somme des carrés des variables géographiques (de façon à obtenir des U.P.É. compactes pour réduire les coûts) et maximiser celle des variables non-géographiques. Ce dernier critère permet d'obtenir les U.P.É. les plus hétérogènes possible en ce qui a trait aux caractéristiques, de façon à ce qu'elles soient toutes bien représentatives de la strate au moment de l'échantillonnage.

Tableau 3
Indices de stratification selon les contraintes géographiques

Région économique	No. de strates	Contraintes géographiques					
		Contiguïte et centroïdes (poids de 3)		Centroïdes (poids de 3)		Aucune	
		1981	1971	1981	1971	1981	1971
520	2	32.2	28.5	30.2	30.1	34.5	27.0
540	3	21.8	14.1	24.9	17.8	35.2	26.8
580	4	22.8	18.9	41.4	33.7	43.7	38.5

Il y a cependant un conflit entre la compacité désirée des U.P.É. et leur hétérogénéité, à cause de la tendance des unités adjacentes à posséder des caractéristiques similaires. Étant donné les coûts d'ordinateur peu élevés, on a fait 3 délimitations par strate avec des poids sur les centroïdes de 10, 15 et 20, relativement aux autres variables. On a ensuite montré les résultats de chaque délimitation sur un graphique dont les axes représentent les centroïdes (voir figure 1). On a fait un choix entre les 3 délimitations en tenant compte de la qualité de l'optimisation des variables, reflétée par les indices de stratification, et en consultant les graphiques. On tenait également compte d'un indice de compacité. En pratique, on a choisi la plupart du temps un poids de 10 ou de 15 sur les centroïdes.

La formation des U.P.É. dans les strates mixtes a amené une contrainte supplémentaire. On désirait en effet que la proportion de la population urbaine soit à peu près la même dans chaque U.P.É. Étant donné qu'on voulait de plus avoir des U.P.É. de populations totales approximativement égales, il a donc été nécessaire parfois de partager les grands centres urbains parmi plus d'une U.P.É. La solution retenue a été la suivante:

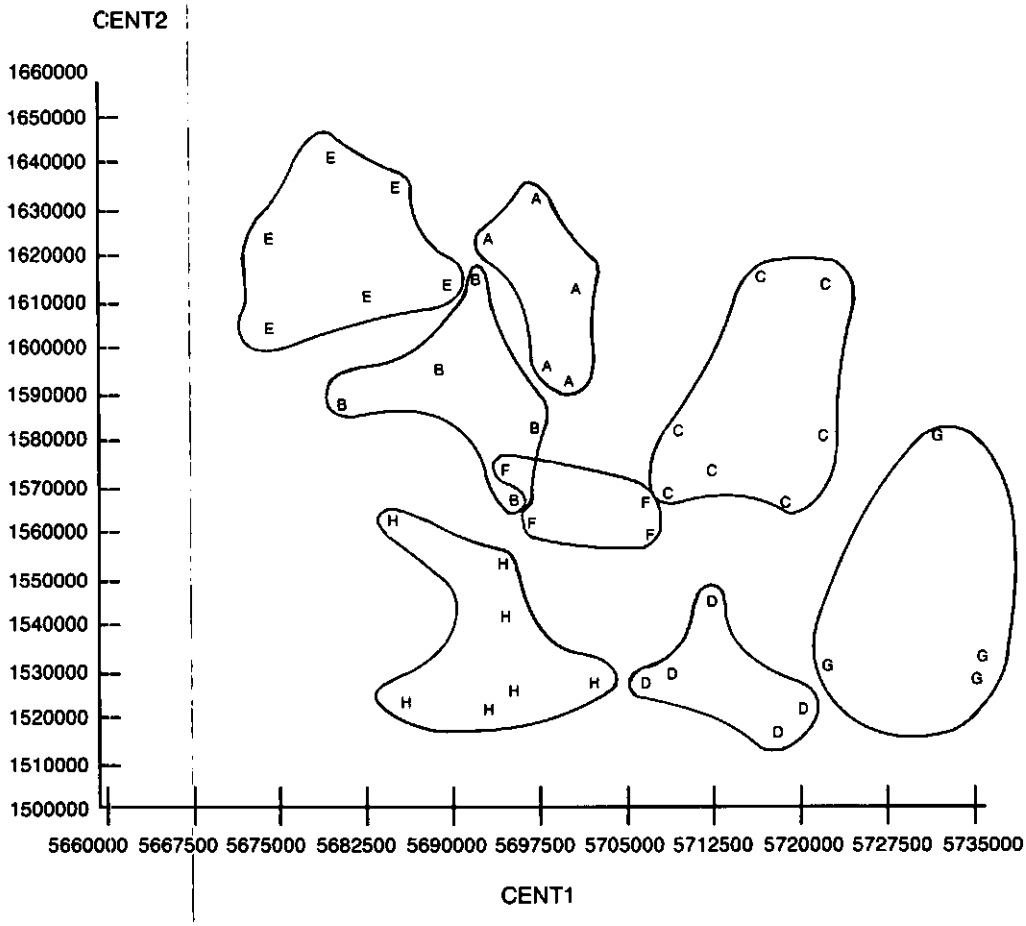


Figure 1. Exemple de délimitation d'U.P.É. Chaque unité de stratification est représentée par une lettre qui identifie l'U.P.É. à laquelle appartient l'unité. On a encerclé les U.P.É. pour mieux les différencier.

1. On détermine le nombre de parties de centres urbains que recevra en moyenne une U.P.É. (N). Ce nombre dépend de la proportion de la population urbaine dans la strate ainsi que du nombre d'unités urbaines. En pratique, il a été fixé à 1 ou 2. Certaines strates ne présentant pas une population ou un nombre d'unités urbaines suffisantes se sont vues reclassifiées comme étant des strates entièrement rurales.
2. On détermine le nombre de parties par lesquelles chaque centre urbain sera divisé. Le nombre total de parties doit égaler N fois le nombre d'U.P.É. et chaque centre urbain est divisé en un nombre de parties proportionnel à sa population.
3. On applique le programme de stratification optimale en considérant chaque partie de centre urbain comme une unité de stratification distincte et en ajoutant la variable "population urbaine" aux autres variables de stratification. On ajuste le poids accordé à cette variable pour obtenir une répartition rural/urbain la plus égale possible dans chaque U.P.É., en essayant de ne pas trop affecter la compacité et l'optimisation globale. Ceci est réalisé par tâtonnement seulement. On a trouvé en pratique qu'un poids de 10 ou 15 sur la population urbaine, relativement aux autres variables, menait à des résultats satisfaisants.

Tableau 4
Indices moyens de délimitation des U.P.É.

Variables	Type de strate		
	Rurale	Mixte	Urbaine
Agriculture	8.1	8.3	9.0
Foresterie	21.8	24.5	35.9
Mines	20.6	36.0	57.0
Manufacture	15.1	22.9	53.3
Construction	9.0	11.4	22.7
Transport	9.9	12.8	22.7
Services	9.4	12.8	29.1
Employés	7.7	10.2	23.6
Chômeurs[a]	13.6	14.2	18.6
Revenu	8.9	11.2	23.7
Population 15-24	9.4	13.4	29.8
Population 55+	7.4	13.9	34.5
Ménages 1 personne	5.1	7.4	13.0
Ménages 2 personnes	7.9	11.9	28.1
Logements possédés	6.8	12.5	29.4
Loyer brut total	5.1	7.7	14.4
Éducation secondaire	9.1	10.5	17.4
Population totale[a]	3.2	4.0	10.5
Logements[a]	5.9	8.9	18.6
Centroïde 1	91.6	92.7	99.2
Centroïde 2	90.5	91.7	97.2

[a] Non utilisé comme variable dans l'optimisation.

Dans les strates urbaines, les U.P.É. ont été formées de centres urbains. On a parfois combiné des centres petits et peu éloignés, mais sans considération d'optimalité des caractéristiques.

Le tableau 4 donne les indices de délimitation moyens pour les U.P.É. dans les strates rurales, mixtes et urbaines. Pour les variables non-géographiques, l'indice le plus bas représente la meilleure délimitation, tandis que c'est le contraire pour les centroïdes. Les résultats sont évidement meilleurs, au point de vue optimalité des caractéristiques, pour les strates rurales et mixtes, où on a utilisé l'algorithme de classification. Les indices élevés des centroïdes révèlent que les U.P.É. sont bien compactes.

4. STRATIFICATION DANS LES UNITÉS AUTOREPRÉSENTATIVES

4.1 Ancienne conception

Les unités autoreprésentatives de l'ancien plan de sondage correspondaient aux villes qui étaient assez grandes pour avoir comme rendement prévu un échantillon suffisant pour être traité par un interviewer. La limite inférieure de la taille des U.A.R. variait de 10,000 personnes dans les provinces de l'Atlantique à 29,000 personnes au Québec et en Ontario.

Les grandes U.A.R. étaient stratifiées géographiquement en regroupant de 3 à 5 secteurs de recensement (S.R.) contigus, sans recherche d'optimalité. Les S.R. sont des unités géostatistiques comprenant de 3,000 à 5,000 habitants, et dont la stabilité d'un recensement à l'autre en fait des unités opérationnelles pratiques. On s'attendait à ce que ces strates soient efficaces pour l'estimation des caractéristiques, et que leur petite taille (entre 10,000 et 15,000 personnes) permette la mise à jour de l'échantillon dans les secteurs expérimentant une croissance rapide, sans d'autre part déranger l'échantillon.

En plus de la base de sondage aréolaire, une base ouverte était tenue à jour pour les immeubles d'appartements dans les grandes villes.

4.2 Étude sur la stratification

On a considéré trois grandes U.A.R. dans cette étude, soient Québec, Ottawa et Toronto. L'unité de stratification choisie a été le secteur de recensement. À cause de contraintes opérationnelles imposées par le programme de stratification, on a dû séparer Toronto en six parties correspondant généralement aux grandes divisions naturelles de la ville. Toute stratification a été menée séparément dans chacune de ces parties. On a utilisé les mêmes 16 variables de stratification qui ont finalement été choisies dans la partie N.A.R.

Deux options principales ont été évaluées:

Option 1: Stratification à deux niveaux:
- strates primaires contiguës et compactes, avec un poids de 3 sur les centroïdes et une taille déterminée en vue d'un rendement d'environ 150 logements.
- strates secondaires – 4 ou 5 par strates primaires, formulées sans contraintes géographiques.

Option 2: Stratification compacte formulée avec l'emploi de centroïdes (poids de 3) sans utiliser les vecteurs de contiguïté, avec une taille comparable à celle des strates secondaires de l'option 1.

Le tableau 5 montre les résultats de la comparaison entre l'ancienne stratification et les deux options étudiées. Comme dans la partie N.A.R., la stratification a été effectuée avec les données du recensement de 1971, pour ensuite être évaluée avec celles du recensement de 1981.

On voit que les deux options étudiées donnent toujours de meilleurs indices que l'ancienne stratification, sauf peut-être pour les trois premières variables dont l'importance est de toute façon faible dans les villes. L'ancienne stratification est quand même très valable si l'on considère qu'elle a été faite sans souci d'optimalité.

Tableau 5
Comparaison de trois méthodes de stratification (U.A.R.)

Variables	Ancienne conception		Stratification à deux niveaux (option 1)		Stratification compacte (option 2)	
	1971	1981	1971	1981	1971	1981
Agriculture	5.5	2.9	3.2	1.8	3.4	1.8
Foresterie	2.2	2.3	2.1	1.7	2.2	2.3
Mines	7.6	4.9	8.5	4.1	7.6	4.0
Manufacture	34.7	35.0	36.6	34.1	39.1	35.0
Construction	32.5	29.6	39.7	30.1	42.4	33.4
Transport	9.2	6.8	18.1	11.6	20.0	11.6
Services	29.5	27.5	45.8	33.1	46.7	32.1
Employés	15.1	8.0	31.4	14.1	32.8	12.6
Chômeurs[a]	14.6	5.7	14.9	6.7	15.5	7.1
Revenu	39.4	38.6	51.8	29.8	53.6	48.0
Population 15-24	9.6	15.2	12.5	17.5	13.3	14.9
Population 55+	27.9	18.3	34.0	20.8	32.6	18.5
Ménages 1 personne	20.3	19.2	36.3	33.8	37.8	35.0
Ménages 2 personnes	21.9	20.3	40.3	30.9	40.1	30.2
Logements possédés	20.3	15.3	29.7	22.9	32.1	24.9
Éducation secondaire	32.6	42.4	50.3	47.9	51.6	49.1
Population 15+[a]	27.0	8.2	38.0	13.4	37.6	12.0
Logements[a]	21.8	18.5	41.7	33.8	42.1	34.3

[a] Non utilisé comme variable de stratification.

On remarque également que les trois méthodes donnent une stratification généralement robuste au fil du temps, telle que réflétée par la comparaison entre les indices 1981 et 1971. Des exceptions importantes à cette règle, malheureusement, semblent être les caractéristiques employés et chômeurs.

4.3 Nouvelle conception

Étant donnée la similitude des résultats entre les deux options étudiées, on a décidé d'adopter la stratification à deux niveaux (option 1) dans les grandes villes où l'échantillon est de 300 ménages ou plus, pour les raisons suivantes:

(i) le fait d'avoir la contiguïté au niveau des strates primaires nous donne une unité convenable pour la mise à jour de l'échantillon

(ii) les strates primaires pourront être utilisées pour la formation des tâches d'intervieweur. La taille des strates a été déterminée de manière à ce que l'échantillon à l'intérieur du secteur, soit l'échantillon aréolaire plus l'échantillon provenant des strates d'appartements, corresponde à deux tâches d'intervieweur (160 ménages dans le coeur de la ville et 120 ailleurs).

(iii) la stratification à deux niveaux mène à une meilleure représentation de la variance de réponse corrélative dans les estimations de la variance. Dans l'ancien plan de sondage, il n'y avait habituellement qu'un seul intervieweur par strate, ce qui occasionnait une sous-estimation de cette composante de la variance. Avec des strates secondaires non-géographiques, mais des tâches d'intervieweurs toujours géographiques, ce problème sera moins fréquent.

Les contraintes de coûts associées à l'exécution des programmes informatiques impliqués nous ont obligé à traiter certaines U.A.R. de façon particulière. En effet, la région de Montréal se voit diviser en sept parties indépendantes, le temps de la stratification. Il en est de même avec Toronto (5 parties), Winnipeg (2 parties), Calgary (2 parties), Edmonton (2 parties) et Vancouver (3 parties). Ces divisions se sont effectuées à partir de critères "naturels" tels que suggérés par la géographie de ces régions.

Dans les grandes U.A.R., les immeubles d'appartements existant au moment de la conception du plan de sondage ont été triés en fonction des strates primaires dans lesquelles ils étaient situés, pour atteindre une stratification implicite de cet échantillon.

Dans les U.A.R. de taille moyenne, où l'échantillon était insuffisant pour justifier la stratification à deux niveaux, on s'est contenté de construire des strates optimales à l'aide de l'algorithme de stratification, sans aucune contrainte géographique.

Les plus petites U.A.R., celle qui n'ont pas été décomposées en côtés d'îlot aux fins du recensement, ont été stratifiées manuellement, sans souci d'optimalité.

Mentionnons finalement que la période d'introduction du nouvel échantillon nous a amené une contrainte supplémentaire. Pour les grandes U.A.R., on a défini le secteur noyau comme étant constitué des strates complètes de l'ancienne conception qui n'ont pas été affectées par des changements de frontières. En s'assurant que les frontières des strates de la nouvelle conception respectent ces secteurs noyau, on s'est aussi assuré que le nouvel échantillon représente la même aire géographique que l'ancien échantillon durant la période d'introduction. Ceci a permis le remplacement graduel de l'ancien échantillon par le nouveau et par le fait même évité un coûteux chevauchement du nouvel échantillon et de l'ancien échantillon (Mayda, Drew, et Lindeyer 1985).

5. CONCLUSION

L'utilisation de l'algorithme de classification multi-variée nous a permis de développer une stratification très générale, renforçant ainsi l'E.P.A. dans son rôle d'enquête-ménages générale. De plus, l'automatisation des étapes de stratification dans les parties N.A.R. et A.R., et de la délimitation des U.P.É. dans les U.N.A.R., a mené à une réduction significative dans le coût et le temps nécessaire pour remanier l'échantillon.

Le système est documenté (Foy 1984) et il peut être utilisé pour la formation de strates ou de grappes dans les autres enquêtes. On pourrait également s'en servir dans des situations où on a à définir des régions statistiques ou administratives, en utilisant toute une gamme de variables.

Pour l'E.P.A., un aspect de recherche à approfondir serait le choix entre des strates contiguës et non-contiguës, et les implications des strates non-contiguës sur le plan d'échantillonnage.

REMERCIEMENTS

Les auteurs désirent remercier Sylvie Trudel et Marc Joncas pour leur aide apportée à la réalisation des études mentionnées dans ce rapport, ainsi que les membres du Comité du remaniement de l'E.P.A. pour leurs précieuses suggestions. Ils remercient aussi l'arbitre pour ses commentaires utiles.

BIBLIOGRAPHIE

CHOUDHRY, G.H., LEE, H., et DREW, J.D. (1985). Optimisation du coût et de la variance dans le cadre de l'enquête sur la population active du Canada. *Techniques d'enquête,* 11, 37-56.

DAHMSTRÖM, P., et HAGNELL, M. (1978) The formation of strata using cluster analysis. Document interne, Department of Statistics, University of Lund, Sweden.

FOY, P. (1984). Programme de stratification pour l'enquête sur la population active du Canada: Guide de l'usager. Document interne, division des méthodes de recensement et d'enquêtes ménages, Statistique Canada.

FRIEDMAN, H.P., et RUBIN, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

JUDKINS, D.R., et SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 274-284.

KOSTANICH, D., JUDKINS, D.R., SINGH, R.P., et SCHANTZ, M. (1981). Modification of Friedman-Rubin's clustering algorithm for use in stratified PPS sampling. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 285-290.

MAYDA, F., DREW, J.D., et LINDEYER, J. (1985). Phase-in of the redesigned Labour Force Survey. Document interne, division des méthodes de recensement et d'enquêtes ménages, Statistique Canada.

PLATEK, R., et SINGH, M.P. (1976). Méthodologie de l'enquête sur la population active du Canada. No. 71-526 au catalogue, Statistique Canada.

SINGH, M.P., DREW, J.D., et CHOUDHRY, G.H. (1984). Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981. *Techniques d'enquête*, 10, 139-154.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 125–133
Statistique Canada

Échantillonnage des questionnaires manuscrits de recensement reproduits sur microfilm

D.R. BELLHOUSE[1]

RÉSUMÉ

Dans la première partie du document, nous procédons à une revue rétrospective des travaux d'historiens sur les dossiers manuscrits de recensement microfilmés. Les historiens ont utilisé plusieurs types de plan d'échantillonnage qui vont, par ordre croissant de complexité, de l'échantillonnage aléatoire en grappes et stratifié jusqu'à l'échantillonnage en grappes à deux degrés stratifié. Dans la deuxième partie, nous proposons une méthode permettant de créer une bande-échantillon à grande diffusion qui contiendrait des données du recensement du Canada de 1881. Cette recherche faisait partie d'un projet pilote exécuté pour les Archives publiques du Canada et a été réalisée par le Centre de données sur les sciences sociales de l'Université Western Ontario. Le projet pilote avait pour but de déterminer s'il était avantageux et possible, du point de vue économique et technique, de construire une base de données ordinolingue à l'aide des questionnaires microfilmés du recensement du Canada de 1881.

MOTS CLÉS: Échantillonnage aléatoire informatisé; dossiers microfilmés; sondages à plusieurs degrés; échantillons à grande diffusion; stratification.

1. INTRODUCTION

Pour écrire l'histoire d'une personne ou d'un peuple, l'historien a besoin de sources. De nos jours, beaucoup d'historiens veulent retracer la vie de l'*homme de la rue*. Pour mener à bien leur recherche, ils peuvent se servir de documents comme les questionnaires de recensement, les titres de propriété et les annuaires commerciaux. Dans le présent document, nous nous penchons sur l'utilisation des questionnaires de recensement comme source. Le principal inconvénient que présente l'utilisation de données de recensement est l'abondance de ces données. L'historien qui dispose d'un budget de recherche normal n'a ni le temps ni les ressources financières ou humaines nécessaires pour passer en revue tous les questionnaires du recensement. Pour contourner cette difficulté, il doit former un échantillon aléatoire de questionnaires. La plupart des questionnaires de recensement que peut consulter l'historien sont reproduits sur microfilm. Au Canada, les questionnaires reproduits sur microfilm sont ceux des recensements coloniaux de 1841, de 1851 et de 1861 et ceux des recensements du Canada de 1871 et de 1881. Le problème se résume donc pour l'historien à déterminer le plan de sondage approprié pour prélever un échantillon de questionnaires microfilmés.

Dans la deuxième section du document, nous passons en revue les méthodes d'échantillonnage appliquées par les historiens. L'application de ces méthodes a donné des résultats très inégaux. Dans certains cas, les résultats ont été très satisfaisants, les historiens ayant su adapter l'application des méthodes à l'objet de la recherche. Dans d'autres cas, toutefois, il semble que les historiens aient utilisé des plans de sondage inutilement complexes. Un plan de sondage complexe peut avoir des effets qui s'écartent sensiblement de 1 et, par conséquent, compliquer l'analyse des données. Voir, par exemple, Rao et Scott (1981) et Holt *et coll.* (1980) pour des commentaires sur l'analyse de données qualitatives, et Scott et Holt (1982) pour des commentaires sur l'analyse par régression. Enfin, les rapports de plusieurs des enquêtes analysées ci-dessous ne contiennent pas assez de renseignements pour nous permettre d'évaluer les raisons qui ont justifié le choix d'un plan de sondage particulier.

[1] D.R. Bellhouse, Département des sciences statistiques et actuarielles, Université Western Ontario, London Ontario), Canada N6A 5B9.

Dans la troisième partie de l'exposé, nous proposons une méthode permettant de prélever des questionnaires du recensement du Canada de 1881 dans le but de créer une bande-échantillon à grande diffusion. Cette recherche faisait partie d'un projet exécuté pour les Archives publiques du Canada. Elle a été confiée par contrat au Centre de données sur les sciences sociales de l'Université Western Ontario. Nous donnons ici une description du plan d'échantillonnage. On trouvera un rapport complet sur le projet dans Mitchell *et coll.* (1982). Le plan d'échantillonnage utilisé ressemble à certains égards aux plans qui ont servi à la création des bandes-échantillons à grande diffusion pour les recensements de 1971 et de 1976. Les plans d'échantillonnage reposent tous sur la stratification; pour le recensement de 1881, toutefois, la stratification n'a pu se faire qu'en fonction d'une répartition géographique.

2. REVUE RÉTROSPECTIVE

Les travaux portant sur l'échantillonnage des documents manuscrits de recensement peuvent être classés suivant la méthode d'échantillonnage utilisée. Nous décrivons ci-dessous les diverses méthodes utilisées par ordre croissant de complexité du plan de sondage.

2.1 Échantillonnage en grappes

Ornstein et Darroch (1978) ont proposé une méthode simple et économique permettant d'échantillonner des dossiers de recensement et de les raccorder d'une période à une autre. Cette méthode consiste essentiellement à former des grappes de noms de famille et à constituer des échantillons à partir de ces grappes. Les grappes sont désignées par la première lettre du nom de famille. Si les mêmes grappes sont échantillonnées dans divers recensements à la fois, une personne dont le nom figure dans plus d'un recensement fera partie de l'échantillon choisi. Il y a donc moins de cas à analyser pour la liaison et le coût est par conséquent moins élevé. Ce plan de sondage se prête particulièrement bien aux études chronologiques de la migration ou de l'évolution démographique.

2.2 Échantillonnage stratifié

Aucun des plans de sondage avec stratification considérés dans la présente section ne prévoit une répartition optimale de l'échantillon. Cette situation s'explique par le fait qu'aucun des historiens n'était en mesure de connaître a priori les variations à l'intérieur des strates. Pour obtenir ce genre de renseignements, il aurait fallu supporter une hausse sensible du coût de chaque projet.

Hammarberg (1971) a utilisé une méthode de sondage à deux phases, ou méthode d'échantillonnage double, dans l'espoir de réduire le biais engendré par l'échantillonnage d'un ensemble incomplet de dossiers. Les dossiers échantillonnés dans la deuxième phase ont été les annuaires commerciaux de neuf comtés de l'Indiana. Dans la première phase du sondage, Hammarberg a prélevé un échantillon dans un ensemble de dossiers supposé complet, le recensement des États-Unis de 1870. Il a appliqué la méthode d'échantillonnage aléatoire stratifié avec répartition proportionnelle de sorte que l'échantillon soit autopondéré. Les strates correspondaient aux neuf comtés de l'Indiana. On trouve deux aspects de cette recherche dans des études ultérieures sur l'échantillonnage de documents anciens. Les strates correspondent à des régions géographiques et l'échantillon est autopondéré.

Hammarberg (1971) a également appliqué le test du khi carré classique à certaines variables pour vérifier dans quelle mesure la distribution des données de l'échantillon se rapprochait des distributions de population établies à partir des données du recensement. Dans beaucoup d'autres études, on ne s'est pas préoccupé de vérifier la *représentativité* de l'échantillon.

Soltow (1975) s'est servi d'échantillons des recensements américains de 1850, de 1860 et de 1870 pour faire une étude de la richesse aux États-Unis. Pour chaque année de recensement, il a prélevé un échantillon sur chaque bobine de microfilm de telle sorte que l'échantillon soit stratifié par bobine, ce qui équivalait à peu près à une stratification géographique. Le plan de sondage de Soltow semble correspondre à un échantillonnage systématique. Pour prélever un échantillon, il a déterminé un point sur l'écran de la visionneuse de microfilms puis a visionné le film. Il faisait avancer la pellicule par demi-tours successifs de manivelle jusqu'à ce qu'un questionnaire de recensement soit relativement centré sur le point désigné sur l'écran. Pour être sélectionné, le questionnaire devait concerner une personne de sexe masculin âgée de 20 ans ou plus. En outre, l'échantillon du recensement de 1860 comprenait 40 fois plus de personnes dont l'avoir s'élevait à $100,000 ou plus que de personnes dont l'avoir était inférieur à $100,000 (p. 5), de sorte que cet échantillon n'était pas autopondéré. Bien que Soltow n'en fasse pas mention, il peut avoir voulu *suréchantillonner* les personnes plus fortunées de manière à disposer d'un échantillon suffisamment grand pour lui permettre d'établir des comparaisons avec les classes moins aisées de la société. Il a également comparé les observations de ses échantillons aux distributions de fréquence publiées mais n'a effectué aucun test de validité de l'ajustement. Il a constaté que pour diverses variables, les données des échantillons se rapprochaient sensiblement de celles des recensements sur le plan des moyennes et des proportions. Ces observations s'appliquaient même à des variables comme le patrimoine moyen, chose surprenante compte tenu du suréchantillonnage de personnes plus fortunées et du fait que l'estimation de Soltow semble correspondre à la moyenne de l'échantillon.

Darroch et Ornstein (1980) ont utilisé un échantillon du recensement du Canada de 1871 pour étudier la relation entre l'origine ethnique et la profession. La méthode d'échantillonnage utilisée est décrite dans Ornstein (1978). Pour les besoins des deux études, il a fallu suréchantillonner certains groupes ethniques de manière à ne pas obtenir un échantillon autopondéré. Ne tenant pas compte de ce suréchantillonnage, les deux historiens ont appliqué un échantillonnage aléatoire stratifié, la stratification étant fondée sur la structure géographico-hiérarchique des dossiers du recensement: provinces, districts, sous-districts et divisions. La division correspond au secteur de dénombrement actuel et semble être le critère naturel de stratification. Toutefois, Ornstein (1978) l'a subdivisée en strates et a prélevé deux ménages dans chaque strate. Il ne dit pas comment il opère cette subdivision mais il la justifie en affirmant que l'échantillonnage de deux unités par strate réduit au minimum la variance des estimations de certaines valeurs d'une population. Bien qu'Ornstein (1978) ne le précise pas, il semble qu'il ait voulu accroître l'efficacité de la stratification en formant des strates à l'intérieur d'une division aussi homogène que possible. Il s'est ainsi trouvé à accroître le coût de l'échantillonnage. Enfin, sa méthode exigeait que l'on fasse défiler le film au moins deux fois dans la visionneuse, la première fois pour obtenir le nombre de ménages par division et la deuxième pour échantillonner les ménages.

Dans leurs ouvrages, Johnson (1978b) et Graham (1980) indiquent comment ils ont obtenu un échantillon à grande diffusion du recensement des États-Unis de 1900. Dans un autre ouvrage, Johnson (1978a) reprend à peu près les mêmes éléments pour décrire comment il a échantillonné les questionnaires du recensement du Rhode Island de 1860. Dans les trois cas, l'échantillon est formé en déterminant au hasard des lignes sur le microfilm et en cherchant par la suite ces lignes à l'aide d'une visionneuse munie d'un compteur. Compte tenu de la méthode d'échantillonnage, la taille globale de l'échantillon est aléatoire. Graham (1980, p. 41) donne un certain nombre de critères permettant d'accepter ou de rejeter les lignes échantillonnées. Il s'agit d'un échantillonnage aléatoire stratifié et les bobines de microfilms servent de strates. La stratification est fondée sur une répartition géographique dans la mesure où les questionnaires de recensement d'une même région se trouvent tous sur la même bobine de microfilms. Cette méthode a l'avantage de rendre l'exécution efficace et de ne nécessiter

qu'un seul visionnement. Elle élimine en outre le problème des strates vides ou des strates à une unité lorsque la fraction de sondage pour une strate est faible. En revanche, comme elle ne comporte qu'un seul visionnement, il faut pouvoir résoudre de façon ponctuelle les principaux problèmes qui peuvent survenir.

2.3 Échantillonnage en grappes stratifié

Bateman et Foust (1974) ont obtenu un échantillon des exploitations agricoles du nord des États-Unis à l'aide des données du recensement de 1860. Ils ont divisé le Nord en deux strates, est et ouest, et prélevé un échantillon aléatoire de comtés ruraux dans chaque strate. Ensuite, ils ont choisi au hasard, dans chaque comté, une commune rurale (grappe) et recueilli des données sur toutes les exploitations agricoles situées sur le territoire de la commune. Une des raisons du choix de l'échantillonnage en grappes est que cette méthode est économique. Comme les données sur les exploitations agricoles provenaient du recensement de l'agriculture et que les données démographiques sur les propriétaires de ces exploitations agricoles et ceux qui y travaillent provenaient du recensement de la population, il était plus facile de faire le lien entre les exploitations agricoles et leurs propriétaires respectifs en se limitant à une commune. Swierenga (1983) donne une deuxième raison pour justifier l'utilisation de l'échantillonnage en grappes. Il affirme que les données recueillies pour une commune ont permis d'estimer la productivité globale des facteurs en agriculture et de définir toute la main-d'oeuvre agricole, y compris les travailleurs agricoles qui demeurent à l'extérieur des 12,000 exploitations comprises dans l'échantillon (p. 793). Comme les grappes (communes) n'ont pas été choisies selon une probabilité proportionnelle à la taille, le plan de sondage n'a pas produit d'échantillons autopondérés.

Bateman and Foust (1974) ont aussi appliqué quelques tests pour vérifier la représentativité de leur échantillon. À l'instar d'Hammarberg (1971), ils ont appliqué le test du khi carré pour comparer les observations de l'échantillon aux observations probables de la population. Dans le cas des variables continues, ils ont utilisé le test t. Les estimations de la moyenne et de la variance étaient des estimations *simples*, c'est-à-dire qu'elles n'étaient pas fondées sur le plan de sondage.

2.4 Échantillonnage à deux degrés stratifié

Hammarberg (1977) a appliqué une méthode d'échantillonnage à deux degrés stratifié pour échantillonner des ménages dans le recensement du territoire de l'Utah de 1880. Les strates sont un amalgame assez complexe de cinq régions géographiques de l'Utah, de quelques comtés de régions populeuses et de quelques grandes municipalités. Hammarberg a prélevé dans chaque strate un échantillon de municipalités ou de wards. Les municipalités qui constituaient déjà des strates étaient automatiquement incluses dans l'échantillon. Les wards sont à l'Église mormone ce que les paroisses étaient à l'Église chrétienne médiévale. Un échantillon de ménages a ensuite été prélevé dans les municipalités ou les wards choisis. Cet échantillon était autopondéré. L'argument que fait valoir Hammarberg à la page 460 de son ouvrage pour justifier ce mode de stratification est convaincant:

> "Comme le mode d'organisation fondamental de la population repose sur une structure géographique et que la plupart des documents officiels - civils et religieux - correspondent à cette structure, on peut dire qu'un échantillonnage de la population suivant une répartition géographique équivaut dans une large mesure à un échantillonnage des dossiers produits pour cette population."

McInnis (1977) a aussi eu recours à l'échantillonnage à deux degrés stratifié pour obtenir un échantillon des dossiers du recensement colonial de 1861. Il étudiait alors la relation entre

le nombre d'enfants par famille et l'abondance des terres dans certaines régions. Il a commencé par répartir environ 300 cantons en strates selon l'année de colonisation. Il a ensuite prélevé un échantillon de cantons dans les strates et des échantillons de fermes dans les cantons. Il semble que McInnis ait choisi l'échantillonnage à deux degrés pour des raisons d'économie. En effet, comme les fermes échantillonnées étaient ensuite appariées au dossier correspondant dans le recensement de l'agriculture, il était moins long et, par conséquent, moins coûteux d'échantillonner quelques cantons et d'apparier les dossiers de plusieurs fermes d'un canton que de stratifier les cantons et d'apparier les dossiers d'un petit nombre de fermes dans chaque strate. Le même raisonnement s'applique aux travaux d'Hammarberg (1977). Lui aussi rattachait d'autres dossiers au ménage échantillonné.

2.5 Échantillonnage en grappes à deux degrés stratifié

Smith (1978) a appliqué une méthode d'échantillonnage en grappes à deux degrés stratifié pour étudier la population âgée dans le recensement des États-Unis de 1900. Les strates sont définies comme les régions de recensement, les comtés qui forment les régions étant les unités primaires d'échantillonnage. Celles-ci sont choisies selon une probabilité proportionnelle à la taille de leur population. Pour chaque comté, Smith a prélevé plusieurs pages de questionnaires du recensement. Il a ensuite relevé les noms des personnes de plus de 50 ans qui figuraient sur chacune des pages échantillonnées. L'échantillonnage en grappes était nécessaire du fait qu'il aurait été trop coûteux de déterminer toutes les personnes qui pouvaient être échantillonnées. Smith tente par la même occasion de comparer quelques distributions d'échantillons aux données publiées du recensement. Il se sert pour cela de la fonction des observations normalement utilisée dans les tests d'hypothèses portant sur une proportion simple même s'il s'agit de données multinomiales.

Foust (1968, chap. 2) décrit un deuxième cas d'échantillonnage en grappes à deux degrés stratifié. L'échantillon, dit échantillon Parker-Gallman, a été tiré du recensement des États-Unis de 1860 pour étudier les régions productrices de coton du sud du pays. Les strates étaient 405 *régions cotonnières* du Sud, où l'on avait produit au moins 1,000 balles de coton de 400 livres dans les douze mois qui précédaient le jour du recensement. Pour chaque région, on a prélevé un échantillon aléatoire systématique de pages de documents manuscrits du recensement, et un groupe de cinq plantations a été choisi au hasard sur une page donnée, le groupe étant considéré comme une grappe. On a recouru à l'échantillonnage en grappes car il fallait consulter trois questionnaires de recensement différents pour recueillir des données sur une plantation particulière. L'appariement des questionnaires a été qualifié de très laborieux. Fogel et Engerman (1974, p. 22-25) énumèrent plusieurs autres échantillons se rattachant à l'échantillon Parker-Gallman. Bode et Ginter (1984) formulent des critiques sur le contenu de l'échantillon.

De tous les échantillons considérés ici, l'échantillon Parker-Gallman et les échantillons de Bateman et Foust (1974) sont ceux qui ont été le plus étudiés. Swierenga (1983) a passé en revue une bonne partie des travaux fondés sur ces échantillons.

3. ÉCHANTILLONS À GRANDE DIFFUSION TIRÉS DU RECENSEMENT DU CANADA DE 1881

Au début des années 1980, les Archives publiques du Canada ont obtenu les copies du *questionnaire 1: Renseignements d'ordre général* du recensement du Canada de 1881. Les questionnaires ont été microfilmés, et on peut maintenant trouver des exemples de ces microfilms dans la plupart des bibliothèques de collège ou d'université et dans beaucoup de bibliothèques publiques. Après avoir produit les microfilms, les Archives publiques du

Canada ont voulu intégrer toutes les données du recensement dans une base ordinolingue ou créer des bandes-échantillons à grande diffusion semblables à celles qui avaient été produites pour les recensements de 1971 et de 1976 (voir Statistique Canada (1975, 1979) pour la documentation). Le Centre de données sur les sciences sociales de l'Université Western Ontario s'est donc vu accorder un contrat pour réaliser une étude de faisabilité, et on a demandé à l'auteur d'élaborer une méthode d'échantillonnage qui permettrait de former l'échantillon à grande diffusion. La présente section contient une description du plan de sondage proposé. On trouvera un rapport de l'étude de faisabilité dans Mitchell *et coll.* (1982).

Le questionnaire 1 contient des renseignements sur l'âge, le sexe, le pays de naissance, l'origine ethnique, la profession et l'état matrimonial de chaque personne et indique si cette personne souffre d'incapacité. Les sept autres questionnaires contiennent des renseignements sur l'industrie, l'agriculture, les forêts, la pêche et les mines. On trouvera une brève description des questionnaires dans *Recensement du Canada 1880-1881*, volume 1, p. v-xv.

Nous décrivons brièvement ci-dessous les conditions de base des échantillons à grande diffusion. Pour que ces échantillons soient conformes à ceux de 1971 et de 1976, il faudrait avoir deux échantillons indépendants, soit un échantillon de ménages et un échantillon de personnes. Si toutefois il n'était économiquement possible de produire qu'un échantillon, ce devrait être en priorité un échantillon de ménages. L'échantillon à grande diffusion tiré du recensement des États-Unis de 1900 et décrit par Johnson (1978b) et Graham (1980) est un échantillon de ménages. En outre, il semble que les historiens préfèrent avant tout le ménage comme unité d'échantillonnage. L'échantillon du recensement de 1900 indique également qu'une taille d'échantillon de l'ordre de 100,000 unités serait souhaitable pour l'échantillon de personnes ou l'échantillon de ménages. En ce qui concerne le recensement du Canada de 1881, cela impliquerait une fraction de sondage d'environ 2% pour l'un ou l'autre échantillon. Enfin, il est également souhaitable d'utiliser, pour l'un et l'autre échantillon, un échantillonnage stratifié avec répartition proportionnelle où les strates correspondent à des régions géographiques. Cette méthode est celle qui a été le plus couramment utilisée jusqu'à maintenant par les historiens, et elle produit un échantillon autopondéré. Le choix des unités dans une strate devrait se faire par échantillonnage aléatoire simple plutôt que par échantillonnage systématique. Johnson (1978a) soutient que l'échantillonnage systématique, bien que pratique, ne convient pas aux questionnaires manuscrits de recensement. Des voisins ont des caractéristiques similaires et ne pourraient jamais appartenir à un même échantillon systématique. Or les historiens pourraient vouloir étudier les personnes qui ont des caractéristiques similaires.

Compte tenu de ces conditions de base, nous proposons pour le prélèvement de l'échantillon de ménages un échantillonnage aléatoire stratifié où les strates correspondent, comme dans Ornstein (1978), aux divisions de recensement (secteurs de dénombrement actuels) plutôt qu'aux bobines de microfilm utilisées par Johnson (1978b) et Graham (1978). Les divisions de recensement sont des strates géographiques naturelles. En outre, les ménages sont numérotés successivement sur les listes des agents recenseurs, chaque page manuscrite comportant vingt-cinq noms. Il suffirait alors de faire un premier visionnement des microfilms pour connaître le nombre de ménages dans chaque strate. Avec une fraction de sondage de 2 à 2.5% et une répartition proportionnelle, on obtient des échantillons de moins de deux ménages dans des divisions (strates) comptant un peu moins de cent ménages. Dans ce cas, la division en question devrait être intégrée à des divisions contigües. Une stratification poussée au-delà de la division, comme celle d'Ornstein (1978), paraît inutile et ferait monter sensiblement le coût de l'échantillonnage.

L'échantillonnage peut être facilement informatisé. Pour un codeur assis à un terminal et utilisant une visionneuse de microfilms, l'échantillonnage est une opération simple. Lorsqu'un codeur échantillonne une division, il n'a qu'à taper le code de la division voulue et

le numéro du premier ménage à échantillonner apparaît à l'écran. Le codeur fait alors avancer le film jusqu'au numéro de ménage voulu. Une fois les données de ce ménage saisies, il appuie sur une touche appropriée et le deuxième numéro de ménage apparaît à l'écran. Lorsqu'il a fini d'échantillonner cette division, il appuie sur la même touche pour passer à une autre division. Il peut parfois y avoir des ménages manquants. Il se peut, par exemple, qu'une ou plus d'une feuille d'agent recenseur contenant vingt-cinq noms ait été perdue. Quand un codeur constate l'absence d'un ménage en échantillonnant une division, il introduit en mémoire le numéro correspondant en indiquant qu'il s'agit d'un ménage manquant et fait de même avec tous les autres numéros sous lesquels il ne voit aucun ménage d'inscrit. Il poursuit ensuite son échantillonnage jusqu'à ce qu'il ait couvert toute la division. Comme il manque au moins un ménage dans l'échantillon prévu, le codeur doit rembobiner le microfilm et poursuivre l'échantillonnage dans la même division. Le principal avantage de cette méthode est de permettre au codeur de faire défiler les microfilms dans une seule direction, sauf lorsqu'il y a des ménages manquants.

L'algorithme utilisé dans cette méthode d'échantillonnage est fondé sur un fichier contenant des données sur les divisions ou groupes de divisions de même que sur l'algorithme de Bebbington (1975) conçu pour le prélèvement d'un échantillon aléatoire simple sans remise. Après le premier visionnement des microfilms, on crée un fichier indiquant l'identificatif de la division et le nombre de ménages que contient la division. Si les divisions sont groupées, leur taille respective est enregistrée. Il suffit pour le codeur de taper l'identificatif de la division à échantillonner pour connaître la taille de cette division. Pour obtenir une répartition proportionnelle de l'échantillon de ménages, il faut que la taille de l'échantillon soit égale au produit de la taille de la division par la fraction de sondage s'appliquant à l'ensemble du recensement. Avec l'algorithme de Bebbington (1975), on effectue ensuite sondage progressif des unités de l'échantillon à partir d'une liste contenant dans l'ordre les numéros des ménages de la division donnée. Les numéros de ménage sont testés tour à tour, puis choisis ou rejetés. Lorsqu'un numéro de ménage est choisi, il est affiché sur l'écran et la sélection s'interrompt le temps de la saisie des données. Comme les numéros sont choisis dans l'ordre ascendant, le codeur n'a qu'à faire défiler le film dans une seule direction pour trouver les documents manuscrits voulus.

Cet algorithme permet aussi d'échantillonner des strates combinées ou divisions groupées. Supposons que L strates de taille $N_1, \ldots, N_L$ aient été combinées dans une seule strate de taille $N = N_1 + N_2 + \ldots + N_L$. Il suffit d'utiliser les tailles des strates pour connaître le ménage échantillonné dans chaque strate. Supposons,dans l'algorithme, que les unités $s(1), \ldots, s(n)$ aient été choisies pour former l'échantillon $1 \leq s(i) \leq N$. Si pour tout $i (i = 1, \ldots, n)$ $N_1 + \ldots + N_{h-1} < s(i) \leq N_1 + \ldots + N_{h-1} + N_h$, où $N_1 + \ldots + N_{h-1} = 0$ pour $h = 1$, l'unité $s(i)$ se trouve dans la strate h et le numéro de ménage correspondant est $s(i) - (N_1 + \ldots + N_{h-1})$.

On peut aussi modifier l'algorithme de sorte qu'il tienne compte des ménages manquants. La méthode décrite ci-dessous ne suppose pas l'énumération des ménages manquants avant l'échantillonnage. une fois terminé l'échantillonnage d'une strate au moyen de l'algorithme de Bebbington (1975), deux cas sont possibles: ou bien on a constaté l'absence d'un certain nombre de ménages, ou bien on n'a constaté aucune absence. Le deuxième cas ne pose aucun problème, l'échantillonnage étant complet. Dans le premier cas, la taille de l'échantillon formé, disons m, est inférieure à la taille voulue n. Pour obtenir un échantillon de ménages existants de taille n il faut échantillonner $n - m$ ménages additionnels. À cette fin, nous recommençons l'échantillonnage de la strate, mais cette fois avec une liste des ménages qui ont déjà été échantillonnés et des ménages manquants connus. Supposons que la liste contienne M ménages ($M \geq n$: le codeur peut avoir remarqué et enregistré des ménages manquants qui n'ont pas été échantillonnés). Définissons un vecteur v à N dimensions, où la valeur du

$u^{\text{ième}}$ élément est u, $v(u) = u$ pour $u = 1, \ldots, N$. Le $u^{\text{ième}}$ élément désigne le $u^{\text{ième}}$ ménage d'une division. Supprimons maintenant tous les éléments de v qui correspondent aux ménages qui ont déjà été échantillonnés ou à ceux qui ne figurent pas sur le microfilm et réduisons le vecteur v à un vecteur w à $(N - M)$ dimensions. Les valeurs $w(u), u = 1, \ldots, N - M$ correspondent aux numéros de ménages susceptibles d'être échantillonnés. Dans l'algorithme, il suffit de redéfinir la taille de la population comme $N - M$ et la taille de l'échantillon comme $n - m$.

Avec de légères modifications, la méthode d'échantillonnage des ménages peut facilement produire un échantillon distinct et indépendant de personnes. Cette flexibilité repose sur le fait que chaque page des listes de recensement est numérotée et contient vingt-cinq noms. Le premier visionnement des microfilms doit servir à connaître le numéro de la dernière page de chaque division et le nombre de lignes que contient cette page. Avec l'algorithme de Bebbington, l'ordinateur imprimera le numéro de la page et le numéro de la ligne où figure le nom de la personne échantillonnée.

Cette méthode d'échantillonnage a été programmée et testée avec succès par le Centre de données sur les sciences sociales. Selon l'étude de faisabilité, par exemple, la recherche des unités échantillonnées a représenté environ 6% de la durée totale estimée de la saisie de données pour l'échantillon de ménages et 18.5% de cette durée pour l'échantillon de personnes. Voir Mitchell *et coll.* (1982, p. 20-21).

BIBLIOGRAPHIE

BATEMAN, F., et FOUST, J.D. (1974). A sample of rural households selected from the 1860 manuscript censuses. *Agricultural History*, 48, 75-93.

BEBBINGTON, A.D. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 135.

BODE, F.A., et GINTER, D.E. (1984). A critique of land holding variables in the 1860 census and the Parker-Gallman sample. *Journal of Interdisciplinary History*, 15, 277-295.

DARROCH, A.G., et ORNSTEIN, M.D. (1980). Ethnicity and occupational structure in Canada in 1871: the vertical mosaic in historical perspective. *Canadian Historical Review*, 61, 305-333.

FOGEL, R.W., et ENGERMAN, S.L. (1974). *Time on the Cross: Evidence and Methods*. Boston: Little, Brown and Co.

FOUST, J.D. (1975). *The Yeoman Farmer and Westward Expansion of U.S. Cotton Production*. New York: Arno Press.

GRAHAM, S.N. (1980). *1900 Public Use Sample: User's Handbook*. Seattle: Centre for Studies in Demography and Ecology, University of Washington.

HAMMARBERG, M.A. (1971). Designing a sample from incomplete historical lists. *American Quarterly*, 23, 542-561.

HAMMARBERG, M.A. (1977). A sampling design for Mormon Utah, 1880, *Journal of Interdisciplinary History*, 7, 453-476.

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

JOHNSON, R.C. (1978a). A procedure for sampling manuscript census schedules. *Journal of Interdisciplinary History*, 8, 513-530.

JOHNSON, R.C. (1978b). The 1900 census sampling project: methods and procedures for sampling and data entry. *Historical Methods*, 11, 147-151.

McINNIS, R.M. (1977). Childbearing and land availability: some evidence from individual household data. *Population Patterns in the past*, (R.D. Lee ed.), New York: Academic Press, 201-227.

MITCHELL, S.P., LING, D.G., et HANIS, E.H. (1982). *Final Report: Determination of Procedures and Costs for the Production of a Machine Readable Edition of the 1881 Census of Canada.* MAS contrat n° OSU80-00326.

ORNSTEIN, M.D. (1978). The design of a sample of households from the 1871 census of Canada. Manuscrit non publié, Université York, Toronto.

ORNSTEIN, M.D., et DARROCH, G.O. (1978). National mobility studies in past time: a sample strategy. *Historical Methods*, 11, 152-161.

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

SCOTT, A.J., et HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 17, 848-854.

SMITH, D.S. (1978). A community-based sample of the older population from the 1880 and 1900 United States manuscript census. *Historical Methods*, 11, 67-74.

SOLTOW, L. (1975). *Men and Wealth in the United States 1850-1870.* New Haven: Yale University Press.

STATISTIQUE CANADA (1975). *Recensement du Canada de 1971: Bandes-échantillon à grande diffusion, Documentation des utilisateurs*, Ottawa.

STATISTIQUE CANADA (1979). *Recensement du Canada de 1976: Bandes-échantillon à grande diffusion, Documentation des utilisateurs*, Ottawa.

SWEIRENGA, R.P. (1983). Quantitative methods in rural landholding. *Journal of Interdisciplinary History*, 13, 787-808.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 135–148
Statistique Canada

Estimation du total pour deux caractères dans les enquêtes à bases de sondage multiples

B.C. SAXENA, P. NARAIN, et A.K. SRIVASTAVA[1]

RÉSUMÉ

Le présent document traite de l'estimation de caractères multiples dans les enquêtes à bases de sondage multiples. Le gain attribuable à l'étude de deux caractères dans une même enquête par rapport à la tenue d'enquêtes distinctes pour chaque caractère est analysé. On fait également des comparaisons de coûts entre une enquête à bases multiples portant sur deux caractères et une enquête à base simple portant sur deux caractères.

MOTS CLÉS: Enquête portant sur plusieurs caractères; estimation par stratification a posteriori; optimisation; comparaison de coûts.

1. INTRODUCTION

La technique des enquêtes à bases multiples a été présentée pour la première fois par Hartley (1962) et examinée en détail par la suite par Lund (1968), Hartley (1974), Vogel (1975), Armstrong (1979), etc. Lund a proposé une solution de rechange à l'estimateur de Hartley, qui comporte la division de l'échantillon entre divers domaines. Hartley (1974) a poursuivi avec une approche plus générale pouvant s'appliquer à divers plans de sondage. Il a constaté que la plupart des cas possibles de bases multiples utilisaient des types d'unités différents dans leurs bases respectives. Bosecker et Ford (1976) ont développé l'estimateur de Hartley pour tirer avantage de la stratification dans le domaine de chevauchement. Serrurier et Phillips (1976) et Armstrong (1978) ont expérimenté les techniques comportant des bases de sondage multiples dans des enquêtes agricoles. L'utilité d'une enquête à bases de sondage multiples a été démontrée dans des cas très divers. Dans des enquêtes par sondage, il arrive parfois que l'intérêt ne réside pas seulement dans l'estimation d'un seul caractère, mais dans l'estimation de plusieurs caractères qui doivent être étudiés simultanément. Pour en arriver à une utilisation efficace des ressources, cela est souvent réalisé par la truchement d'enquêtes intégrées. Par exemple, dans le but d'estimer la production de récoltes de légumes, on peut concevoir une enquête unique qui a pour objet d'estimer la production de plusieurs récoltes de légumes. Également, au lieu d'utiliser une base composée de tous les producteurs de légumes, on peut utiliser une base incomplète mais relativement facilement, accessible notamment une base composée des principaux producteurs de légumes. L'estimation du total pour deux caractères dans les enquêtes à bases multiples a été examinée dans le présent document. L'avantage d'étudier plusieurs caractères dans une seule enquête par rapport au recours à des enquêtes indépendantes portant chacune sur une caractère est également analysé.

2. ESTIMATEUR

Soit A et B deux bases de sondage empiétant l'une sur l'autre de taille N_A et N_B respectivement. Dans des enquêtes à bases multiples, deux échantillons de taille n_A et n_B sont tirés de façon indépendante par échantillonnage aléatoire simple à partir des bases de sondage A et B respectivement. Les bases chevauchantes génèrent des domaines *a, b* et *ab* définis comme suit:

a: domaine composé des unités appartenant à la base A uniquement,
b: domaine composé des unités appartenent à la base B uniquement,
ab: domaine composé des unités appart;enant à la fois à la base A et à la base B.

[1] B.C. Saxena, P. Narain, et A.K. Srivastava, Indian Agricultural Statistics Research Institute, Nouvelle-Delhi, Inde.

Les échantillons de taille n_A et n_B sont divisés en échantillons de taille n_a *et* n_{ab} et de taille n_b *et* n_{ba} de sorte que n_a et n_{ab} correspondent aux nombre d'unités de l'échantillon de taille n_A appartenant aux domaines a et ab respectivement. De la même façon, les échantillons de taille n_b et n_{ba} sont les échantillons provenant de la division de l'échantillon de taille n_B en échantillons composés d'unités appartenant aux domaines b et ab respectivement. Dans une étude à plusieurs caractères, ces domaines seront une fois de plus divisés, générant des sous-domaines comme suit:

Soit $y_{(1)}$ et $y_{(2)}$ deux caractères devant faire l'objet d'une étude. Chacun des domaines habituels a, ab et b sera une fois de plus divisé en $a(1)$, $a(12)$, $a(2)$, $ab(1)$, $ab(12)$, $ab(2)$ et $b(1)$, $b(12)$ et $b(2)$ respectivement. Dans ce cas-ci, $a(1)$, $a(12)$ et $a(2)$ sont les sous-domaines composés des unités du domaine a ayant le caractère $y_{(1)}$, les deux caractères $y_{(1)}$ et $y_{(2)}$ et le caractère $y_{(2)}$ respectivement. La même explication s'applique aux autres sous-domaines $ab(1)$, $ab(12)$, etc. Par conséquent, la division des deux échantillons dans l'étude de deux caractères se fera de la façon suivante:

$$n_A = n_a + n_{ab}$$

où

$$n_a = n_{a(1)} + n_{a(2)} + n_{a(12)} \quad \text{et} \quad n_{ab} = n_{ab(1)} + n_{ab(2)} + n_{ab(12)},$$

et

$$n_B = n_b + n_{ba}$$

où

$$n_b = n_{b(1)} + n_{b(2)} + n_{b(12)} \quad \text{et} \quad n_{ba} = n_{ba(1)} + n_{ba(2)} + n_{ba(12)}.$$

Dans ce cas-ci, les échantillons de taille $n_{a(1)}$, $n_{a(2)}$, etc. sont les échantillons composés des unités de l'échantillon de taille n_a appartenant respectivement aux sous-domaines $a(1)$, $a(2)$, etc. Si on se limite à un caractère, on peut alors définir

$$n_{A(1)} = n_{a(1)} + n_{a(12)} + n_{ab(1)} + n_{ab(12)},$$

$$n_{B(1)} = n_{b(1)} + n_{b(12)} + n_{ba(1)} + n_{ba(12)}.$$

De la même façon, les échantillons de taille $n_{A(2)}$ et $n_{B(2)}$ sont définis pour le deuxième caractère. L'estimation du total pour le premier caractère est donnée par

$$\hat{Y}^{(1)} = \hat{Y}_{a(1)} + \hat{Y}^{(1)}_{a(12)} + p_1 \hat{Y}_{ab(1)} + q_1 \hat{Y}_{ba(1)} + p_2 \hat{Y}^{(1)}_{ab(12)} + \\ + q_2 \hat{Y}^{(1)}_{ba(12)} + \hat{Y}_{b(1)} + \hat{Y}^{(1)}_{b(12)} \qquad (1)$$

où $\hat{Y}_{a(1)}$, $\hat{Y}^{(1)}_{a(12)}$, etc . sont les totaux estimatifs pour le caractère $y_{(1)}$ des sous-domaines respectifs. Dans l'analyse qui suit, pour ce qui a trait aux domaines ayant les deux caractères, l'indice supérieur correspond au caractère examiné. Dans le cas des domaines ayant seulement un caractère, l'indice supérieur n'est pas utilisé parce que, de toute évidence, le domaine correspond au caractère.

Aussi, $p_1 + q_1 = 1$ et $p_2 + q_2 = 1$. Définissons $\bar{y}_{a(1)}$, $\bar{y}_{a(2)}$, etc. comme les moyennes des échantillons pour les sous-domaines respectifs ayant les caractères $y_{(1)}$ et $y_{(2)}$ respectivement.

Par conséquent,

$$\hat{Y}^{(1)} = N_{a(1)}\bar{y}_{a(1)} + N_{a(12)}\bar{y}^{(1)}_{a(12)} + N_{ab(1)}\left(p_1\bar{y}_{ab(1)} + q_1\bar{y}_{ba(1)}\right)$$
$$+ N_{ab(12)}\left(p_2\bar{y}^{(1)}_{ab(12)} + q_2\bar{y}^{(1)}_{ba(12)}\right)$$
$$+ N_{b(12)}\bar{y}^{(1)}_{b(12)} + N_{b(1)}\bar{y}_{b(1)}. \qquad (2)$$

De la même façon, pour le deuxième caractère, on a

$$\hat{Y}^{(2)} = N_{a(2)}\bar{y}_{a(2)} + N_{a(12)}\bar{y}^{(2)}_{a(12)} + N_{ab(2)}\left(p_3\bar{y}_{ab(2)} + q_3\bar{y}_{ba(2)}\right)$$
$$+ N_{ab(12)}\left(p_4\bar{y}^{(2)}_{ab(12)} + q_4\bar{y}^{(2)}_{ba(12)}\right) + N_{b(12)}\bar{y}^{(2)}_{b(12)}$$
$$+ N_{b(2)}\bar{y}_{b(2)} \qquad (3)$$

où

$$p_3 + q_3 = 1 \text{ et } p_4 + q_4 = 1.$$

2.1 Variance de l'estimateur

La variance conditionnelle des estimations $\hat{Y}^{(1)}$ et $\hat{Y}^{(2)}$ fondées sur la stratification a posteriori pour un sous-domaine donné qui ne tient pas compte de la correction d'échantillonnage pour population finie peut s'écrire de la façon suivante

$$V(\hat{Y}^{(1)} \mid n_{a(1)}, n_{a(12)}, \text{ etc.}) = N^2_{a(1)}\frac{\sigma^2_{a(1)}}{n_{a(1)}} + N^2_{a(12)}\frac{\sigma^{(1)2}_{a(12)}}{n_{a(12)}}$$
$$+ N^{(2)}_{ab(1)}\left(p_1^2\frac{\sigma^2_{ab(1)}}{n_{ab(1)}} + q_1^2\frac{\sigma^2_{ab(1)}}{n_{ba(1)}}\right)$$
$$+ N^2_{ab(12)}\left(p_2^2\frac{\sigma^{(1)2}_{ab(12)}}{n_{ab(12)}} + q_2^2\frac{\sigma^{(1)2}_{ab(12)}}{n_{ba(12)}}\right) + N^2_{b(1)}\frac{\sigma^2_{b(1)}}{n_{b(1)}}$$
$$+ N^2_{b(12)}\frac{\sigma^{(1)2}_{b(12)}}{n_{b(12)}} \qquad (4)$$

La variance non conditionnelle de $\hat{Y}^{(1)}$ est donnée approximativement par

$$V(\hat{Y}^{(1)}) = \frac{N_A}{n_A}\left\{N_{a(1)}\sigma^2_{a(1)} + N_{a(12)}\sigma^{(1)2}_{a(12)} + p_1^2N_{ab(1)}\sigma^2_{ab(1)}\right.$$
$$\left. + p_2^2N_{ab(12)}\sigma^{(1)2}_{ab(12)}\right\} + \frac{N_B}{n_B}\left\{N_{b(1)}\sigma^2_{b(1)} + N_{b(12)}\sigma^{(1)2}_{b(12)}\right. \qquad (5)$$
$$\left. + q_1^2N_{ab(1)}\sigma^2_{ab(1)} + q_2^2N_{ab(12)}\sigma^{(1)2}_{ab(12)}\right\}$$

qui est égale à la variance obtenue dans le cas d'un échantillonnage stratifié à répartition proportionnelle.

Pareillement,

$$V(\hat{Y}^{(2)}) = \frac{N_A}{n_A}\left\{(N_{a(2)}\sigma^2_{a(2)} + N_{a(12)}\sigma^{(2)^2}_{a(12)} + p_3^2 N_{ab(2)}\sigma^2_{ab(2)}\right.$$

$$\left. + p_4^2 N_{ab(12)}\sigma^{(2)^2}_{ab(12)}\right\} + \frac{N_B}{n_B}\left\{N_{b(2)}\sigma^2_{b(2)} + N_{b(12)}\sigma^{(2)^2}_{b(12)}\right.$$

$$\left. + q_3^2 N_{ab(2)}\sigma^2_{ab(2)} + q_4^2 N_{ab(12)}\sigma^{(2)^2}_{ab(12)}\right\} \tag{6}$$

où $\sigma^2_{a(1)}$, $\sigma^2_{a(2)}$, etc. sont les variances pour les deux caractères dans les sous-domaines respectifs.

Pour optimiser les p_i's ($i = 1, 2, 3, 4$) dans une enquête commune, il faut minimiser une combinaison de variances individuelles sous réserve du maintien du coût fixe total de l'enquête combinée. Prenons la combinaison linéaire la plus simple possible

$$F = V(\hat{Y}^{(1)}) + V(\hat{Y}^{(2)}).$$

Dans le cas de l'enquête commune, une fonction de coût qui convienne pourrait être la suivante:

$$C' = C_1(n_{a(1)} + n_{ab(1)}) + C_2(n_{a(12)} + n_{ab(12)}) + C_3(n_{a(2)} + n_{ab(2)})$$

$$+ C_4(n_{b(1)} + n_{ba(1)}) + C_5(n_{b(12)} + n_{ba(12)}) + C_6(n_{b(2)} + n_{ba(2)}) \tag{7}$$

où C_1 est le coût par unité dans le sous domaine $a(1)$, $ab(1)$; C_2 dans le sous-domaine $a(12)$, $ab(12)$; C_3 dans le sous-domaine $a(2)$, $ab(2)$ de la base A. De même C_4, C_5 et C_6 sont les coûts par unité dans les sous-domaines de la base B. Dans la fonction de coût ci-haut, la taille des échantillons pris au hasard entre en jeu. Soit le coût prévu

$$C = E(C') = n_A(C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3) + n_B(C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6) \tag{8}$$

où

$$\Phi_1 = \frac{N_{a(1)} + N_{ab(1)}}{N_A}, \quad \Phi_2 = \frac{N_{a(12)} + N_{ab(12)}}{N_A},$$

$$\Phi_3 = \frac{N_{a(2)} + N_{ab(2)}}{N_A}, \quad \Phi_4 = \frac{N_{b(1)} + N_{ba(1)}}{N_B},$$

$$\Phi_5 = \frac{N_{b(12)} + N_{ba(12)}}{N_B}, \quad \Phi_6 = \frac{N_{b(2)} + N_{ba(2)}}{N_B}.$$

Où

$$C = n_A C_A + n_B C_B \tag{9}$$

où

$$C_A = C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3 \quad \text{et} \quad C_B = C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6.$$

Pour obtenir les p_i optimums, tout comme n_A et n_B optimums, ilfaut minimiser la fonction F sous réserve de l'existence de la fonction de coût prévu comme celle calculée dans (9). Les variables de pondération p_i et les tailles des échantillons sont obtenues comme suit à l'aide du multiplicateur de Lagrange:

$$\frac{P_1}{q_1} = \frac{P_2}{q_2} = \frac{P_3}{q_3} = \frac{P_4}{q_4} = \frac{N_B}{n_B}\frac{n_A}{N_A} = \frac{P}{q} \text{ (disons)}, \tag{10}$$

et

$$\frac{n_A^2}{N_A} = \gamma\frac{K_5 + K_1p_1^2 + K_2p_2^2 + K_3p_3^2 + K_4p_4^2}{C_A},$$

$$\frac{n_B^2}{N_B} = \gamma\frac{K_6 + K_1q_1^2 + K_2q_2^2 + K_3q_3^2 + K_4q_4^2}{C_B}, \tag{11}$$

où γ est déterminé pour satisfaire au coût prévu et où

$$K_1 = N_{ab(1)}\sigma^2_{ab(1)},\ K_2 = N_{ab(12)}\sigma^{(1)2}_{ab(12)},$$

$$K_3 = N_{ab(2)}\sigma^2_{ab(2)},\ K_4 = N_{ab(12)}\sigma^{(2)2}_{ab(12)},$$

$$K_5 = N_{a(1)}\sigma^2_{a(1)} + N_{a(2)}\sigma^2_{a(2)} + N_{a(12)}(\sigma^{(1)2}_{a(12)} + \sigma^{(2)2}_{a(12)}),$$

$$K_6 = N_{b(1)}\sigma^2_{b(1)} + N_{b(2)}\sigma^2_{b(2)} + N_{b(12)}(\sigma^{(1)2}_{b(12)} + \sigma^{(2)2}_{b(12)}). \tag{12}$$

À partir de (10) et (11), on obtient

$$\frac{q^2}{p^2}\frac{N_B}{N_A}\frac{C_B}{C_A} = \frac{K_6 + (K_1 + K_2 + K_3 + K_4)q^2}{K_5 + (K_1 + K_2 + K_3 + K_4)p^2} \tag{13}$$

Il s'agit d'une fonction bi-quadratique en p qui peut être résolue pour p. Les fractions de sondage optimums peuvent être calculées à l'aide de (11). Un des cas qui se présente souvent dans les sondages à bases multiples est celui où les bases englobent la population totale. Prenons le cas où la base A englobe la population totale, alors $N_{b(1)} = N_{b(2)} = N_{b(12)} = 0$.

Dans ce cas, (13) se ramène à

$$p^2 = \frac{\alpha}{\varrho - \alpha}\frac{K_5}{K_1 + K_2 + K_3 + K_4} \tag{14}$$

où

$$\varrho = \frac{C_A}{C_B} \quad \text{and} \quad \alpha = \frac{N_B}{N_A}.$$

Supposons que

$$\sigma^2_{a(1)} = \sigma^{(1)2}_{a(12)},\ \sigma^2_{a(2)} = \sigma^{(2)2}_{a(12)},\ \sigma^2_{ab(1)} = \sigma^{(1)2}_{ab(12)},\ \sigma^2_{ab(2)} = \sigma^{(2)2}_{ab(12)}. \tag{15}$$

Ces hypothèses semblent plausibles étant donné qu'il est peu probable que la variabilité d'un caractère soit touchée par la présence ou l'absence de l'autre caractère. Alors, p^2 se ramène à

$$p^2 = \frac{\alpha}{\varrho - \alpha}\left\{\frac{\sigma^2_{a(1)}(N_{a(1)} + N_{a(12)}) + \sigma^2_{a(2)}(N_{a(2)} + N_{a(12)})}{\sigma^2_{ab(1)}(N_{ab(1)} + N_{ab(12)}) + \sigma^2_{ab(2)}(N_{ab(2)} + N_{ab(12)})}\right\}$$

ou enfin à

$$p^2 = \frac{(1 - \alpha)\Phi'_2}{(\varrho - \alpha)}\left\{\frac{\Phi'_3(\xi_1 + \xi_2) + (1 - \xi_1)}{\Phi'_4(\xi_3 + \xi_4) + (1 - \xi_3)}\right\} \tag{16}$$

où

$$\Phi'_1 = \frac{\sigma^2_{a(1)}}{\sigma^2_{ab(1)}},\ \Phi'_2 = \frac{\sigma^2_{a(2)}}{\sigma^2_{ab(2)}},\ \Phi'_3 = \frac{\sigma^2_{a(1)}}{\sigma^2_{a(2)}},\ \Phi'_4 = \frac{\sigma^2_{ab(1)}}{\sigma^2_{ab(2)}}$$

et

$$\xi_1 = \frac{N_{a(1)}}{N_a},\ \xi_2 = \frac{N_{a(12)}}{N_a},\ \xi_3 = \frac{N_{ab(1)}}{N_{ab}},\ \xi_4 = \frac{N_{ab(12)}}{N_{ab}}.$$

En se servant du fait que $N_{ab} = N_B$, $N_{a(2)} + N_{a(12)} = N_a - N_{a(1)}$ et $N_{ab(2)} + N_{ab(12)} = N_{ab} - N_{ab(1)}$, on peut voir que l'expression ci-dessus se ramène à la forme habituelle du cas à un seul caractère étant donné que $\xi_1 = \xi_3 = 1$ et $\xi_2 = \xi_4 = 0$. Il convient de noter que les variances dans des domaines ne sont en général pas connues étant donné que de telles valeurs sont fondées sur des renseignements antérieurs ou sur des valeurs estimées. L'optimalité de p^2 est affecté dans cette mesure.

3. COMPARISON D'UNE ENQUÊTE PORTANT SUR PLUSIEURS CARACTÈRES ET DES ENQUÊTES INDÉPENDANTES PORTANT SUR UN SEUL CARACTÈRE DANS DES SITUATIONS À BASES MULTIPLES

Les enquêtes portant sur plusieurs caractères sont conçues dans le but d'économiser les ressources disponibles, et on peut s'attendre qu'une enquête commune présente des avantages supérieurs à ceux des enquêtes indépendantes portant sur un seul caractère, sur le plan des coûts et de l'efficacité. Nous allons donc examiner l'importance des avantages (gain) attribuables à l'utilisation d'une enquête commune à bases de sondage multiples.

Soit, dans une étude portant sur un seul caractère, disons $y_{(1)}$, des échantillons aléatoires simples de taille n_A et n_B tirés des bases de sondage A et B respectivement. Nous supposons ici que les seules bases disponibles sont les bases utilisées précédemment, et non pas les bases réduites pour chaque caractère. Définissions N_{A1}, N_{B1}, n^*_{A1}, et n^*_{B1} comme les tailles de la population et les tailles des échantillons portant sur le caractère $y_{(1)}$. Ici, n^*_{A1} and n^*_{b1}

sont les tailles des échantillons aléatoires et $E(n_{A1}^*) = n_A N_{A1}/N_A$ et $E(n_{B1}^*) = n_B N_{B1}/N_B$. Dans ce cas, l'estimateur $\hat{Y}^{(1)*}$ et sa variance sont donnés par:

$$\hat{Y}^{(1)*} = (N_{a(1)} + N_{a(12)})\bar{y}_{(a(1),\, a(12))}$$

$$+ (N_{ab(1)} + N_{ab(12)})(p'\bar{y}_{(ab(1),\, ab(12))} + q'\bar{y}_{(ba(1),\, ba(12))})$$

$$+ (N_{b(1)} + N_{b(12)})\bar{y}_{(b(1),\, b(12))}$$

où p' et q' sont les variables de pondération de sorte que $p' + q' = 1$ et où $\bar{y}_{(a(1),\, a(12))}, \bar{y}_{(ab(1),\, ab(12))}$, etc. sont les moyennes des échantillons tirés des domaines combinés respectifs, c'est-à-dire, par exemple, que $\bar{y}_{(a(1),\, a(12))}$ est la moyenne des unités de l'échantillon tiré des domaines $a(1)$ and $a(12)$.

$$V(\hat{Y}^{(1)*}) = \frac{N_A}{n_A}(N_{a(1)}\sigma^2_{a(1)} + N_{a(12)}\sigma^{(1)2}_{a(12)}$$

$$+ (p'^2\frac{N_A}{n_A} + q'^2\frac{N_B}{n_B})(N_{ab(1)}\sigma^2_{ab(1)} + N_{ab(12)}\sigma^{(1)2}_{ab(12)})$$

$$+ \frac{N_B}{n_B}(N_{b(1)}\sigma^2_{b(1)} + N_{b(12)}\sigma^{(1)2}_{b(12)}). \qquad (17)$$

Dans ce cas, la fonction de coût prend la forme

$$C = C_1 n_{A1}^* + C_4 n_{B1}^*$$

et le coût prévu est représenté par C^*

$$C^* = C_1\frac{n_A}{N_A}N_{A1} + C_4\frac{n_B}{N_B}N_{B1} = C_A' n_A + C_B' n_B \qquad (18)$$

où $C_A' = C_1 N_{A1}/N_A$ et $C_B' = C_4 N_{B1}/N_B$.

Pour simplifier, nous supposons un champ d'observation de 100% par la base A et l'égalité des variances comme dans (15), et nous supposons que $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$. Suivant ces hypothèses, le coût C^* avec n_A et n_B qui minimisent la variance (17) se calcule comme suit (voir annexe pour le détail du calcul).

$$C^* = \frac{(\xi_1 + \xi_2)[\{C_1(1 + \alpha_1^*)(\Phi_1' + \alpha_1^* p'^2)\}^{1/2} + \alpha_1^*(C_4 q'^2)^{1/2}]^2}{\frac{1}{1 - \alpha}\left\{\frac{(\Phi_1' + \alpha_1^* p^2)}{n_A} + \frac{\alpha\alpha_1^* q^2}{n_B}\right\}}$$

où

$$\alpha_1^* = \frac{\alpha}{1-\alpha}\frac{\xi_3 + \xi_4}{\xi_1 + \xi_2}.$$

De la même façon, pour l'enquête distincte portant sur le deuxième caractère, le coût se calcule somme suit:

$$C^{**} = \frac{(1-\xi_1)\left[\left\{C_3(1+\alpha_2^*)(\Phi_2' + \alpha_2^* p''^2)\right\}^{1/2} + \alpha_2^*(C_6 q''^2)^{1/2}\right]^2}{\frac{1}{1-\alpha}\left\{\frac{(\Phi_2' + \alpha_2^* p^2)}{n_A} + \frac{\alpha\alpha_2^* q^2}{n_B}\right\}} \tag{19}$$

où

$$p''^2 = \frac{K\Phi_2'}{1 + \alpha_2^*(1-K)}, \quad \alpha_2^* = \frac{\alpha}{1-\alpha}\frac{1-\xi_3}{1-\xi_1}.$$

Pour l'étude combinée portant sur les deux caractères, le coût total C dans le cas d'un champ d'observation de 100% de la base A est donné par (8).

Par conséquent

$$C = \frac{n_A}{N_A}\left[C_1(N_{a(1)} + N_{ab(1)}) + C_2(N_{a(12)} + N_{ab(12)}) + C_3(N_{a(2)} + N_{ab(2)})\right]$$

$$+ \frac{n_B}{N_B}\left[C_4 N_{ab(1)} + C_5 N_{ab(12)} + C_6 N_{ab(2)}\right].$$

En utilisant les hypothèses formulées au sujet des coûts (c'est-à-dire, en supposant que $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$) on obtient

$$C = C_2 n_A\left[(1-\alpha)\left\{\varrho_1\xi_1 + \xi_2 + \varrho_3(1-\xi_1-\xi_2)\right\} +\right.$$

$$\left.\alpha\left\{\varrho_1\xi_3 + \xi_4 + \varrho_3(1-\xi_3-\xi_4)\right\} + \frac{K}{r}\left\{\varrho_1\xi_3 + \xi_4 + \varrho_3(1-\xi_3-\xi_4)\right\}\right] \tag{20}$$

où $r = n_A/n_B$, $\varrho_1 = C_1/C_2$ *and* $\varrho_3 = C_3/C_2$.

Mais dansl'étude combinée portant sur les deux caractères (n_A/n_B) optimal $= p/\alpha q$ où p est donné par (16). Donc, la valeur du gain peut être obtenue à partir du ratio.

$$\frac{C^* + C^{**}}{C} =$$

$$\frac{\frac{(\xi_1+\xi_2)\varrho_1 T_1^2}{(\Phi_1' + \alpha_1^* p)} + \frac{(1-\xi_1)\varrho_3 T_2^2}{(\Phi_3' + \alpha_2^* p)}}{\left\{\varrho_1\xi_1 + \xi_2 + \varrho_3(1-\xi_1-\xi_2)\right\} + \left\{\varrho_1\xi_3 + \xi_4 + \varrho_3(1-\xi_3-\xi_4)\right\}\left\{\frac{r\alpha + K}{r(1-\alpha)}\right\}} \tag{21}$$

où

$$T_1 = \left\{(\Phi_1' + \alpha_1^* p'^2)(1 + \alpha_1^*)\right\}^{1/2} + \alpha_1^* q' \sqrt{K}$$

$$T_2 = \left\{(\Phi_2' + \alpha_2^* p''^2)(1 + \alpha_2^*)\right\}^{1/2} + \alpha_2^* q'' \sqrt{K}.$$

K peut être calculé de la façon suivante: à l'aide de la définition de C_A, C_B, Φ_i et de ($i = 1, \ldots, 6$) et de l'équation (A.1), on obtient

$$\frac{C_A}{C_B} = \frac{1}{K} \frac{\varrho_1 \Phi_1 + \Phi_2 + \varrho_3 \Phi_3}{\varrho_1 \Phi_4 + \Phi_5 + \varrho_3 \Phi_6} = \varrho_3,$$

et donc

$$K = \varrho^{-1}\left\{\alpha + (1 - \alpha)\frac{\varrho_1 \xi_1 + \xi_2 + \varrho_3(1 - \xi_1 - \xi_2)}{\varrho_1 \xi_3 + \xi_4 + \varrho_3(1 - \xi_3 - \xi_4)}\right\}. \tag{22}$$

On peut utiliser l'expression de l'équation (21) pour calculer le gain au niveau des coûts que produit une étude portant sur les deux caractères simultanément par rapport à des enquêtes individuelles indépendantes. Le gain G exprimé en pourcentage est donc donné par

$$G = \left(\frac{C^* + C^{**}}{C} - 1\right) \times 100$$

Dans la comparaison des coûts ci-dessus, les coûts prévus, C, C^* et C^{**} n'incluent pas les frais généraux afférents à la réalisation d'une enquête combinée ou à la réalisation d'enquêtes distinctes. On s'attend toutefois que le total des frais généraux d'enquêtes dinstinctes soit beaucoup plus élevé que le total correspondant se rapportant à la réalisation d'une enquête combinée. Par conséquent, le gain réel attribuable à la réalisation d'enquêtes indépendantes sera supérieur au gain G exprimé en pourcentage, tel qu'il est défini plus haut.

L'expression (21) se simplifie de façon appréciable en fonction des hypothèses $\Phi_1' = \Phi_2' = \Phi$ et $\xi_1 = \xi_2 = \xi_3 = \xi_4 = \xi$ (disons).

À partir de (22) $\varrho = 1/K$ et de (16) puisque $\Phi_1'/\Phi_2' = \Phi_3'/\Phi_4'$, le p^2 se simplifie comme suit:

$$p^2 = \frac{K(1 - \alpha)}{1 - k\alpha} \Phi.$$

En outre $\alpha_1^* = \alpha_2^* = \alpha/(1 - \alpha)$.

Par conséquent, à partir de (A.1)

$$p' = p'' = \left\{\frac{K(1 - \alpha)\Phi}{1 - K\alpha}\right\}^{1/2}.$$

D'où

$$T_1 = T_2 = \left\{\Phi \frac{1 - K\alpha}{1 - \alpha}\right\}^{1/2} + \frac{\alpha \sqrt{K}}{1 - \alpha}.$$

Avec toutes ces substitutions apportées à l'expression (21) $(C^* + C^{**})/C$ se simplifie comme suit:

$$\frac{C^* + C^{**}}{C} = \frac{T_2^2}{\Phi + \alpha_1^* p} \times \frac{2\xi\varrho_1 + (1 - \xi)\varrho_3}{\left\{\frac{r\alpha + K}{K(1 - \alpha)} + 1\right\}\{\varrho_1\xi + \xi + \varrho_3(1 - 2\xi)\}}$$

$$= \frac{T_1^2}{(\Phi + \alpha_1^* p)\left\{\frac{r\alpha + K}{K(1 - \alpha)} + 1\right\}} \times \frac{\varrho_3 + \xi(2\varrho_1 - \varrho_3)}{\varrho_3 + \xi(\varrho_1 + 1 - 2\varrho_3)}$$

$$= \frac{\varrho_3 + \xi(2\varrho_1 - \varrho_3)}{\varrho_3 + \xi(1 + \varrho_1 - 2\varrho_3)}$$

où $r = (n_A/n_B)$ optimal $= p/\alpha q$ de (10).

D'où,

$$G = \frac{\xi(\varrho_1 + \varrho_3 - 1)}{\varrho_3 + \xi(1 + \varrho_1 - 2\varrho_3)} \times 100.$$

L'équalité des ξ_i ne semble pas être une hypothèse réaliste. La valeur de G a donc été calculée à l'aide de (19) pour des combinaisons réalistes et représentatives de paramètres; les résultats obtenus figurent dans le tableau 1.

Ce tableau révèle qu'il y a manifestement un gain attribuable à l'intégration d'enquêtes à bases multiples portant sur deux caractères par rapport à la réalisation d'enquêtes distinctes. Le gain augment à mesure qu'augmentent les valeurs de ϱ_1 et ϱ_3.

4. COMPARISON D'ENQUÊTES À BASE MULTIPLES PORTANT SUR DEUX CARACTÈRES ET D'UNE ENQUÊTE À BASE UNIQUE

La comparaison d'une enquête à deux bases de sondage et des enquêtes à base unique pour l'étude de deux caractères présente de l'intérêt sur le plan pratique. Une étude analogue a été faite par Hartley (1962) pour le cas d'un seul caractère. La réduction relative des coûts calculée suivant une méthode analogue s'exprime

$$R = \left(1 + \frac{\alpha q}{p\varrho}\right)^2 \Big/ \left\{1 + \frac{\alpha q(1 + p)}{p^2\varrho}\right\}$$

où p^2 is given by (16), $\varrho = C_A/C_B$ et $\alpha = N_{ab}/N_A$

La réduction des coûts attribuable à l'utilisation d'une enquête à bases multiples plutôt qu'à l'utilisation d'une enquête à base unique a été calculée pour un ensemble donné de valeurs des paramètres et les résultats figurent dans le tableau 2. D'après ce tableau, la réduction des coûts est considérable.

Tableau 1

Gain en pourcentage au niveau des coûts attribuable à l'utilisation d'une enquête commune à bases multiples portant sur deux caractères plutôt qu'à l'utilisation d'enquêtes distinctes

Lorsque $\varrho = 10$, $\Phi_1' = 0.25$, $\Phi_2' = 0.5$, $\Phi_3' = 1$, $\alpha = 0.5$.

	ϱ_3						
ϱ_1	0.3	0.4	0.5	0.6	0.7	0.8	0.9
			$\xi_1 = 0.2, \xi_2 = 0.2, \xi_3 = 0.4, \xi_4 = 0.2$				
0.3						1.5	3.9
0.4					1.7	4.2	6.4
0.5				1.8	4.4	6.7	8.7
0.6			1.8	4.5	6.9	8.9	10.7
0.7		1.7	4.6	7.0	9.1	10.9	12.6
0.8	1.7	4.6	7.1	9.3	11.2	12.8	14.3
0.9	4.5	7.1	9.4	11.3	13.0	14.5	15.9
			$\xi_1 = 0.2, \xi_2 = 0.4, \xi_3 = 0.2, \xi_4 = 0.4$				
0.3						4.5	9.1
0.4					4.6	9.3	13.6
0.5				4.8	9.6	14.0	17.9
0.6			4.9	9.9	14.3	18.3	22.0
0.7		5.1	10.1	14.7	18.8	22.5	25.9
0.8	5.2	10.4	15.1	19.3	23.1	26.5	29.7
0.9	10.8	15.5	19.8	23.6	27.1	30.3	33.2

Tableau 2

Réduction des coûts pour des variances constantes

Lorsque $\Phi_1' = 0.25$, $\Phi_2' = 0.5$, $\Phi_3' = 1$, et $\xi_1 = 0.2$, $\xi_2 = 0.3$, $\xi_4 = 0.4$.

	α					
ϱ	0.5	0.6	0.7	0.8	0.9	0.95
100	.227	.175	.132	.094	.059	.040
20	.304	.254	.200	.169	.127	.101
10	.367	.321	.279	.238	.193	.164
5	.462	.423	.387	.351	.308	.277
2	.661	.646	.634	.621	.599	.578
1	.876	.895	.918	.943	.971	.985

ANNEXE

La solution optimale pour p' qui minimise la variance (17) sous la contrainte de C^* selon les hypothèses d'un champ d'observation de 100% par la base A et l'égalité des variances est donnée par

$$p'^2 = \frac{1 - \alpha}{\varrho' - \alpha}\left\{\frac{\sigma^2_{a(1)}(\xi_1 + \xi_2)}{\sigma^2_{ab(1)}(\xi_3 + \xi_4)}\right\}$$

avec

$$\varrho' = \frac{C'_A}{C'_B}.$$

En se servant de ce que $N_{A1} = N_{a(1)} + N_{a(12)} + N_{ab(1)} + N_{ab(12)}$ et de ce que $N_{B1} = N_{ab(1)} + N_{ab(12)}$, on peut écrire ϱ' de la façon suivante

$$\varrho' = \frac{C_1}{C_4}\alpha\frac{N_a(\xi_1 + \xi_2) + N_{ab}(\xi_3 + \xi_4)}{N_{ab}(\xi_3 + \xi_4)}$$

$$= \frac{C_1}{C_4}\alpha\left(\frac{1 - \alpha}{\alpha}\frac{\xi_1 + \xi_2}{\xi_3 + \xi_4} + 1\right)$$

$$= \frac{\alpha}{K}\left(\frac{1}{\alpha_1^*} + 1\right)$$

où

$$\alpha_1^* = \frac{\alpha}{1 - \alpha}\frac{\xi_3 + \xi_4}{\xi_1 + \xi_2}.$$

D'où

$$p'^2 = \frac{1 - \alpha}{\frac{\alpha}{K}\left(\frac{1}{\alpha_1^*} + 1\right) - \alpha}\frac{\xi_1 + \xi_2}{\xi_3 + \xi_4}\Phi'_1$$

$$= \frac{K\Phi'_1}{1 + \alpha_1^*(1 - K)} \qquad \text{(A.1)}$$

Définissons

$$\lambda_1 = (N_{a(1)} + N_{a(12)})\sigma^2_{a(1)} + p^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)},$$

$$\lambda_2 = q^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)},$$

$$\lambda_3 = (N_{a(1)} + N_{a(12)})\sigma^2_{a(1)} + p'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)},$$

$$\lambda_4 = q'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)}.$$

Étant donné le p' de l'équation (A.1), les tailles optimales des échantillons seront

$$\frac{n^2_{AO}}{N_A} = \gamma' \frac{(N_{a(1)} + N_{a(12)})\sigma^2_{ab(1)} + p'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(12)}}{C'_A}$$

$$= \gamma' \frac{\lambda_3}{C'_A}$$

$$\frac{n^2_{BO}}{N_B} = \gamma' \frac{q'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)}}{C'_B} = \gamma' \frac{\lambda'_4}{C'_B}$$

avec γ' déterminé en fonction de (18). À partir de cela, on obtient

$$\frac{n_{BO}}{n_{AO}} = \frac{N_B}{N_A}\left(\frac{C_1 N_{A1}\lambda_4}{C_4 N_{B1}\lambda_3}\right)^{1/2}. \qquad \text{(A.2)}$$

De plus, les variances calculées à l'aide de (5) et (17) pour des tailles d'échantillons optimales peuvent s'écrire

$$V(\hat{Y}^{(1)}) = \frac{N_A}{n_A}\lambda_1 + \frac{N_B}{n_B}\lambda_2 \qquad \text{(A.3)}$$

$$V(\hat{Y}^{*(1)}) = \frac{N_A}{n_{AO}}\lambda_3 + \frac{N_B}{n_{BO}}\lambda_4.$$

En égalant les variances ci-dessus et en utilisant l'équation (A.2), on obtient l'expression pour n_{AO} et n_{BO} en termes de n_A et n_B de la façon suivante:

$$\frac{n_{AO}}{N_A} = \frac{\lambda_3 + \left(\dfrac{C_4\lambda_3\lambda_4 N_{B1}}{C_1 N_{A1}}\right)^{1/2}}{\dfrac{N_A}{n_A}\lambda_1 + \dfrac{N_B}{n_B}\lambda_2}$$

et

$$\frac{n_{BO}}{N_B} = \frac{\lambda_4 + \left(\dfrac{C_1\lambda_3\lambda_4 N_{A1}}{C_4 N_{B1}}\right)^{1/2}}{\dfrac{N_A}{n_A}\lambda_1 + \dfrac{N_B}{n_B}\lambda_2}.$$

En utilisant ces relations, on peut calculer le coût C^* de la façon suivante:

$$C^* = \frac{(\xi_1 + \xi_2)\left[\left\{C_1(1 + \alpha_1^*)(\Phi_1' + \alpha_1^* p'^2)\right\}^{1/2} + \alpha_1^*\left(C_4 q'^2\right)^{1/2}\right]^2}{\dfrac{1}{1 - \alpha}\left\{\dfrac{(\Phi_1' + \alpha_1^* p^2)}{n_A} + \dfrac{\alpha\alpha_1^* q^2}{n_B}\right\}}. \qquad \text{(A.4)}$$

BIBLIOGRAPHIE

ARMSTRONG, B. (1979). Test for multiple frames sampling technique for agricultural survey: Nouveau-Brunswick, 1978. *Techniques d'enquête*, 5, 178-199.

BOSECKER, R.R., et FORD, B.L. (1976). Multiple frame estimation with stratified overlap domain. *American Statistical Association Proceedings of the Social Statistics Section*, 219-224.

HARTLEY, H.O. (1962). Multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 203-206.

HARTLEY, H.O. (1974). Multiple frame methodology and selected application. *Sankhya*, Series C, 36, 99-118.

LUND, R.E. (1968). Estimation in multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 282-288.

SERRURIER, D., and PHILLIPS, J. (1976). Double frame Ontario pilot hog surveys. *Survey Methodology*, 2, 138-170.

VOGEL, F.A. (1975). Surveys with overlapping frames, problems in application. *American Statistical Association Proceedings of the Social Statistics Section*, 695-699.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 149–161
Statistique Canada

Désaisonnalisation des séries pour la population active en périodes de récession et de non-récession

ESTELA BEE DAGUM et MARIETTA MORRY[1]

RÉSUMÉ

Les auteures analysent les révisions de huit séries sur la population active désaisonnalisées en périodes de récession et de non-récession. Elles utilisent à cette fin les méthodes de désaisonnalisation X-11 et X-11-ARMMI appliquées avec des facteurs saisonniers courants ou avec des facteurs saisonniers prévus. La désaisonnalisation se fait suivant un modèle de décomposition additif et un modèle de décomposition multiplicatif. Les résultats indiquent que la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne les révisions les plus faibles tant en période de récession qu'en période de non-récession, et ce peu importe le modèle de décomposition utilisé.

MOTS CLÉS: X-11; X-11-ARMMI; désaisonnalisation courante; récession/non-récession.

1. INTRODUCTION

Le mouvement saisonnier de certaines séries sur la population active peut connaître des fluctuations soudaines à cause des changements majeurs que subit la composition de ces séries au cours des diverses étapes du cycle économique. La série sur le nombre total de chômeurs en est un exemple typique. Durant les années relativement prospères, cette série comprend surtout les personnes qui changent d'emploi, les nouveaux venus sur le marché du travail, les travailleurs du secteur primaire (agriculture, forêt, pêche, piégeage, etc.) et de la construction (en hiver) et les étudiants en quête d'un emploi (en été). En période de récession, toutefois, le nombre de chômeurs augmente rapidement et les nouveaux chômeurs sont surtout des travailleurs réguliers des industries lourdes et des secteurs connexes caractérisés par des variations saisonnières moins fortes et des structures saisonnières de l'emploi différentes de celles observées en période *normale*. Ce phénomène s'est produit au Canada en 1981-1982, lorsque le total non désaisonnalisé des chômeurs est passé de 790,000 en août 1981 à 1,494,000 en décembre 1982, les nouveaux chômeurs provenant principalement du secteur manufacturier et du secteur des services.

Compte tenu des variations subites de la taille et de la composition de la population des chômeurs durant les périodes de crise, il y a lieu de se demander si la méthode d'estimation des facteurs saisonniers, qui est appliquée aux années où le chômage est faible et surtout frictionnel et saisonnier, peut aussi s'appliquer aux années où le chômage est élevé et alimenté en grande partie par des pertes d'emploi dans les secteurs secondaire et tertiaire.

Au terme d'une recherche empirique effectuée à Statistique Canada en 1974, on a réussi à désaisonnaliser les séries sur la population active au moyen de la méthode X-11-ARMMI avec facteurs saisonniers courants. Dans les sections qui suivent, nous désignons cette méthode de désaisonnalisation la méthode *officielle*. En 1980, le U.S. Bureau of Labor Statistics a adopté officiellement la méthode X-11-ARMMI avec facteurs saisonniers prévus six mois à l'avance. Cet organisme applique également la méthode X-11-ARMMI avec facteurs saisonniers courants pour calculer le taux de chômage qu'il publie mensuellement. Les facteurs saisonniers courants s'obtiennent en désaisonnalisant chaque mois toutes les données connues jusqu'à ce jour, tandis que les facteurs saisonniers prévus proviennent de données qui remontent habituellement à un an (à six mois pour ce qui est du Bureau of Labor Statistics).

[1] Estela Bee Dagum et Marietta Morry, Division des séries chronologique - recherche et analyse, Statistique Canada, 13e étage, Immeuble R.H. Coats, parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

La présente étude vise principalement à vérifier si la méthode X-11-ARMMI utilisée avec des facteurs saisonniers courants entraîne toujours les plus faibles révisions en période de récession, comparativement à trois autres méthodes de désaisonnalisation.

Dans la section 2, on présente les erreurs absolues moyennes (EAM) des facteurs saisonniers courants et des facteurs saisonniers prévus une année à l'avance pour huit séries sur la population active du Canada désaisonnalisées au moyen des méthodes X-11-ARMMI et X-11 avec modèle de décomposition multiplicatif. On analyse les facteurs prévus une année à l'avance plutôt que ceux prévus six mois à l'avance parce que de nombreux organismes statistiques publics les utilisent. En outre, les EAM des facteurs prévus six mois à l'avance se situent entre celles des facteurs courants et celles des facteurs prévus une année à l'avance.

Dans la section 3, on calcule les erreurs absolues moyennes engendrées par les quatre méthodes de désaisonnalisation courante appliquées à un modèle additif et on les compare à celles obtenues avec un modèle multiplicatif.

Enfin, la section 4 contient les conclusions de l'étude.

2. RÉVISION DES FACTEURS SAISONNIERS SERVANT À LA DÉSAISONNALISATION COURANTE DES SÉRIES SUR LA POPULATION ACTIVE EN PÉRIODES DE RÉCESSION ET DE NON-RÉCESSION

La plupart des méthodes de désaisonnalisation appliquées par les organismes statistiques publics sont fondées sur les filtres de lissage linéaires, appelés couramment moyennes mobiles. De par la nature même de ces méthodes, les estimations des observations des dernières années sont moins précises que celles des observations centrales à cause de l'asymétrie des filtres appliqués aux extrémités de la série. Les méthodes de désaisonnalisation les plus connues sont la variante X-11 de la Census Method II, mise au point par Shiskin, Young et Musgrave (1967), et la X-11-ARMMI, mise au point par Dagum (1980). La X-11-ARMMI est une version modifiée de la variante X-11 et comporte essentiellement deux étapes. Elle permet premièrement d'extrapoler la série originale au moyen de modèles ARMMI (modèles autorégressifs à moyennes mobiles intégrés) du type Box et Jenkins (1970) et de désaisonnaliser la série ainsi prolongée au moyen d'un ensemble de moyennes mobiles qui sont une combinaison des filtres saisonniers de la variante X-11 et des filtres d'extrapolation de l'ARMMI. Les filtres de désaisonnalisation de la X-11-ARMMI et de la X-11 ne sont donc pas les mêmes pour ce qui a trait aux observations des dernières années. En ce qui concerne les observations centrales toutefois, les deux méthodes prévoient l'application du même filtre symétrique. Lorsque le modèle ARMMI n'est pas utilisé, la X-11-ARMMI équivaut à la méthode X-11.

À mesure que des données s'ajoutent, on révise l'estimation désaisonnalisée se rapportant à une période donnée jusqu'à ce qu'il y ait un écart de trois ans entre celle-ci et la fin de la série et que les filtres symétriques puissent s'appliquer. L'estimation devient alors pratiquement fixe et est considérée comme une estimation désaisonnalisée finale. La différence entre la toute première estimation et l'estimation désaisonnalisée finale est définie comme la révision totale. L'application des méthodes X-11-ARMMI et X-11 entraîne une révision des valeurs désaisonnalisées courantes à cause des différences qui surgissent entre les filtres de lissage linéaires appliqués aux mêmes observations à mesure que de nouvelles données s'ajoutent et à cause des modifications que les nouvelles observations introduisent dans une série. Il serait souhaitable de supprimer ou de réduire au minimum les révisions causées par le premier facteur.

Des études théoriques réalisées par l'une des auteures (Dagum 1982a, 1982b) ont montré qu'il était possible de réduire sensiblement ces révisions en prolongeant la série d'origine avec des valeurs extrapolées ARMMI (c'est-à-dire, en appliquant la X-11-ARMMI) et en utilisant des facteurs saisonniers courants au lieu de facteurs saisonniers prévus une année à l'avance.

Les conclusions de ces deux études théoriques confirment les résultats présentés dans de nombreux travaux théoriques et empiriques (voir, par exemple, Dagum 1978; Dagum et Morry 1982; Kuiper 1978, 1981; Pierce 1980; Kenny et Durbin 1982; McKenzie 1982; Wallis 1982; Pierce et McKenzie 1985; Otto 1985).

Nous allons maintenant comparer l'efficacité de la méthode X-11-ARMMI utilisée avec des facteurs saisonniers courants et celle de trois autres méthodes de désaisonnalisation en périodes de récession et de non-récession. La méthode la plus efficace sera celle qui entraîne les plus faibles révisions.

2.1 Comparaison de quatre méthodes de désaisonnalisation courante des séries sur la population active

Quatre méthodes de désaisonnalisation sont généralement utilisées pour produire des valeurs désaisonnalisées courantes; ce sont:

1) la méthode X-11-ARMMI avec facteurs saisonniers courants;
2) la méthode X-11 avec facteurs saisonniers courants;
3) la méthode X-11-ARMMI avec facteurs saisonniers prévus une année à l'avance;
4) la méthode X-11 avec facteurs saisonniers prévus une année à l'avance.

La mesure de la révision qui sert, en l'occurrence, à évaluer ces quatre méthodes est l'erreur absolue moyenne (EAM) des facteurs saisonniers pour la désaisonnalisation courante:

$$MAE(N) = \sum_{t=1}^{N} | \hat{S}_t^C - \hat{S}_t^F | / N \qquad (1)$$

Dans l'équation ci-dessus, N représente le nombre d'éléments inclus dans la moyenne et $\hat{S}_t^C$ désigne la valeur courante du facteur saisonnier, lequel peut être un facteur saisonnier courant ou un facteur saisonnier prévu une année à l'avance (l'un ou l'autre utilisé avec les méthodes X-11 ou X-11-ARMMI). $\hat{S}_t^F$ désigne la valeur "finale" du facteur saisonnier, c'est-à-dire que cette valeur ne variera pas beaucoup lorsque de nouvelles données s'ajouteront à la série. Pour la X-11 comme pour la X-11-ARMMI, la valeur courante d'un facteur saisonnier devient finale lorsque les observations d'au moins trois années sont ajoutées à la série (Young 1968; Wallis 1974).

Dans le présent exposé, on examine les révisions des facteurs saisonniers (ou des facteurs saisonniers implicites s'il s'agit du modèle additif) plutôt que celles des estimations désaisonnalisées, et ce pour plusieurs raisons. Premièrement, l'utilisation des facteurs saisonniers permet d'avoir une idée de l'importance des révisions par rapport à l'ordre de grandeur de la série (puisqu'il s'agit d'un pourcentage); deuxièmement, elle uniformise la valeur des révisions à l'intérieur des séries dont l'ordre de grandeur est soumis à de fortes variations (par exemple, les séries sur le chômage); troisièmement, elle permet de faire des recoupements entre les séries.

Contrairement à une étude antérieure faite par les mêmes auteures (Dagum et Morry 1982), le présent exposé ne tient pas compte des révisions des variations mensuelles des données désaisonnalisées puisque ces révisions ne présentent pas un très grand intérêt lorsqu'on analyse des données sur la population active (Statistique Canada, par exemple, ne publie pas de révisions annuelles du taux de croissance pour ces séries). Cette étude porte donc sur les révisions de l'ordre de grandeur plutôt que sur les révisions de ses variations.

Les huit séries canadiennes sur l'emploi et le chômage analysées ici débutent en janvier 1966 et se terminent en octobre 1982. Pour pouvoir utiliser l'option des extrapolations ARMMI du programme X-11-ARMMI, il faut disposer de données remontant à au moins cinq

années pour produire une série désaisonnalisée. Par conséquent, 1971 est la première année pour laquelle nous pouvons calculer des mesures de la révision totale. L'année 1977 est la dernière année complète pour laquelle une EAM a été calculée. Les sept années se prêtant ainsi au calcul des révisions se caractérisent par deux années de récession et cinq années de non-récession.

La période de récession englobe les données d'août 1974 à juillet 1975 et de juin 1976 à mai 1977. Ces deux années sont considérées comme des années de récession parce qu'on y a observé une forte augmentation (supérieure à 25%) du nombre annuel de chômeurs imputable surtout au grand nombre de mises à pied.

Un autre aspect important dont les auteures ont tenu compte a trait au genre de modèle de décomposition utilisé pour la désaisonnalisation de chaque série. La variante X-11 et le programme X-11-ARMMI produisent aussi bien une version additive qu'une version multiplicative. Il n'y a aucune raison théorique qui justifie la supériorité d'un modèle par rapport à l'autre. Les deux reposent sur des hypothèses différentes concernant le mécanisme de production de la composante saisonnière.

Dans un modèle additif, les éléments d'une série chronologique (tendance-cycle, variations saisonnières et aléas) sont supposés être indépendants. L'effet saisonnier n'est donc pas influencé par le niveau d'activité économique, lequel est déterminé par les étapes du cycle économique.

Dans un modèle multiplicatif, par ailleurs, l'effet saisonnier est proportionnel à la tendance-cycle. Si les facteurs saisonniers sont constants, cette relation signifie que plus l'ordre de grandeur de la série désaisonnalisée est élevé, plus l'effet saisonnier est prononcé.

Le choix du modèle de décomposition n'est pas déterminant pour l'estimation des valeurs désaisonnalisées (finales) puisque, la plupart du temps, les chiffres diffèrent peu d'un modèle à l'autre. Cependant, la question prend une toute autre importance lorsqu'il s'agit d'estimer la composante saisonnière des premières et des dernières années d'une série, surtout une série dont la tendance-cycle croît rapidement. Les filtres asymétriques utilisés pour l'estimation des valeurs extrêmes d'une série, particulièrement ceux de la méthode X-11, introduisent d'importantes erreurs systématiques lorsque les estimations saisonnières fluctuent rapidement (Dagum 1978). En fait, si le modèle de décomposition considéré est un modèle multiplicatif caractérisé par un mouvement saisonnier plutôt stable, la désaisonnalisation d'une série de composition additive produira des estimations saisonnières qui sembleront varier en fonction de la tendance-cycle. Réciproquement, s'il s'agit d'un modèle de décomposition additif caractérisé par un mouvement saisonnier stable, la désaisonnalisation d'une série de composition multiplicative produira des facteurs saisonniers dont le mouvement semblera imprévisible.

Du point de vue de la désaisonnalisation, il est donc préférable de choisir le modèle de décomposition qui produit les estimations saisonnières les plus stables. Les tests mis au point par Morry (1975) et Higginson (1977) ont été appliqués aux huit séries pour déterminer quel modèle de décomposition convient le mieux.

Ces tests ont révélé que deux séries seulement (chômage chez les femmes de 25 ans et plus et celles de 15 à 24 ans) étaient de composition additive tandis que les autres étaient de composition multiplicative.

Dans le présent exposé, toutefois, nous analysons les révisions absolues moyennes en fonction des deux possibilités, c'est-à-dire une série de composition multiplicative ou une série de composition additive. Nous appliquons les deux genres de modèles de décomposition à des données couvrant à la fois des périodes de récession et des périodes de non-récession afin de déterminer lequel de ces modèles, du point de vue de la révision, est le plus sensible aux variations subites d'ordre de grandeur.

Les chiffres figurant dans les tableaux qui suivent ont été obtenus par la désaisonnalisation de séries de composition multiplicative. Les résultats de la désaisonnalisation de séries de composition additive sont analysés à la section 3.

Le tableau 1 donne l'erreur absolue moyenne (EAM) des facteurs saisonniers des méthodes X-11-ARMMI et X-11 utilisées pour la désaisonnalisation courante en période de récession. Il est clair d'après ce tableau que la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne les plus faibles révisions. Cela confirme les résultats des études théoriques qui ont été commentées précédemment et selon lesquelles l'utilisation des extrapolations ARMMI avec des facteurs saisonniers courants avait pour effet de réduire sensiblement les révisions attribuables aux filtres.

La méthode X-11 avec facteurs saisonniers courants produit les plus faibles EAM après celles de la X-11-ARMMI avec facteurs saisonniers courants dans six séries sur huit. Pour ce qui a trait aux deux autres séries (chômage–hommes de 25 ans et plus et emploi – hommes de 25 ans et plus), la X-11 avec facteurs saisonniers courants produit les mêmes EAM que la méthode X-11-ARMMI avec facteurs saisonniers prévus une année à l'avance. Enfin, la méthode X-11 avec facteurs saisonniers prévus une année à l'avance est celle qui produit les estimations les moins précises.

Le tableau 2 présente une comparaison de l'efficacité relative de la méthode X-11-ARMMI avec facteurs saisonniers courants et de celle des trois autres méthodes, en période de récession. Les rapports, tous supérieurs à l'unité, indiquent que la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne effectivement les plus faibles révisions.

La période de non-récession s'étend de janvier 1971 à décembre 1977, hors les années de récession. Le tableau 3 donne les EAM des séries désaisonnalisées courantes par suite de l'application des quatre méthodes en période de non-récession.

Comme au tableau 1, la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne les plus faibles révisions dans toutes les séries puisque, comme il a été mentionné plus haut, elle réduit au minimum les révisions attribuables aux filtres. La méthode X-11 avec facteurs saisonniers courants produit les deuxièmes meilleurs résultats dans sept séries sur huit, ces résultats étant relativement comparables à ceux obtenus par la méthode X-11-ARMMI avec facteurs saisonniers prévus une année à l'avance. Enfin, la méthode X-11 avec facteurs saisonniers prévus une année à l'avance est celle qui entraîne les révisions les plus importantes.

Tableau 1

Erreurs absolues moyennes (EAM(N)) des facteurs saisonniers des méthodes X-11-ARMMI et X-11 durant les années de récession[a] (N = 24)

Séries	Facteurs saisonniers courants		Facteurs saisonniers prévus une année à l'avance	
	X-11-ARMMI (1)	X-11 (2)	X-11-ARMMI (3)	X-11 (4)
Chômage				
Hommes, 25 ans et plus	1.95	2.75	2.74	3.35
Femmes, 25 ans et plus	1.94	2.94	3.43	4.70
Hommes, 15 à 24 ans	2.16	3.02	3.49	4.33
Femmes, 15 à 24 ans	1.25	1.73	2.48	3.44
Emploi				
Hommes, 25 ans et plus	0.08	0.12	0.12	0.16
Femmes, 25 ans et plus	0.23	0.29	0.33	0.42
Hommes, 15 à 24 ans	0.41	0.53	0.66	0.76
Femmes, 15 à 24 ans	0.50	0.70	0.81	0.97

[a] D'août 1974 à juillet 1975 et de juin 1976 à mai 1977.

Tableau 2

Comparaison de l'EAM(N) produites par la méthode X-11-ARMMI avec facteurs saisonniers courants et des EAM(N) produites par trois autres méthodes pour la désaisonnalisation des séries de composition multiplicative sur l'emploi et le chômage en période de récession (N = 24)

Séries	X-11 Facteurs saisonniers courants	X-11-ARMMI Facteurs saisonniers prévus	X-11 Facteurs saisonniers prévus
	X-11-ARMMI Facteurs saisonniers courants (1)[a]	X-11-ARMMI Facteurs saisonniers courants (2)[b]	X-11-ARMMI Facteurs saisonniers courants (3)[c]
Chômage			
Hommes, 25 ans et plus	1.41	1.40	1.72
Femmes, 25 ans et plus	1.52	1.77	2.41
Hommes, 15 à 24 ans	1.40	1.61	2.00
Femmes, 15 à 24 ans	1.38	1.98	2.75
Emploi			
Hommes, 25 ans et plus	1.50	1.50	1.50
Femmes, 25 ans et plus	1.26	1.43	1.83
Hommes, 15 à 24 ans	1.29	1.61	1.85
Femmes, 15 à 24 ans	1.40	1.62	1.94

[a] La colonne (1) est égale au quotient des colonnes (2) et (1) du tableau 1.
[b] La colonne (2) est égale au quotient des colonnes (3) et (1) du tableau 1.
[c] La colonne (3) est égale au quotient des colonnes (4) et (1) du tableau 1.

Tableau 3

Erreurs absolues moyennes (EAM(N)) des facteurs saisonniers des méthodes X-11-ARMMI et X-11 durant les années de non-récession[a] (N = 60)

Séries	Facteurs saisonniers courants		Facteurs saisonniers prévus une année à l'avance	
	X-11-ARMMI (1)	X-11 (2)	X-11-ARMMI (3)	X-11 (4)
Chômage				
Hommes, 25 ans et plus	1.37	1.73	2.22	2.73
Femmes, 25 ans et plus	1.84	2.41	2.92	3.55
Hommes, 15 à 24 ans	1.97	2.66	3.17	3.96
Femmes, 15 à 24 ans	1.93	2.87	2.59	3.18
Emploi				
Hommes, 25 ans et plus	0.08	0.10	0.12	0.13
Femmes, 25 ans et plus	0.23	0.27	0.33	0.34
Hommes, 15 à 24 ans	0.39	0.46	0.58	0.69
Femmes, 15 à 24 ans	0.43	0.49	0.68	0.80

[a] De janvier 1971 à décembre 1977, à l'exclusion des périodes de récession définies à la note (*a*) du tableau 1.

Le tableau 4 présente une comparaison de l'efficacité relative de la méthode X-11-ARMMI avec facteurs saisonniers courants et de celle des trois autres méthodes pour les années de non-récession. On constate que les chiffres de la première colonne, sauf pour ce qui a trait à une série, sont inférieurs à ceux de la première colonne du tableau 2. On peut en déduire que l'utilisation des extrapolations ARMMI permet de réaliser des gains en pourcentage encore plus élevés en période de récession.

Enfin, le tableau 5 donne une comparaison de l'importance des révisions en période de récession et en période de non-récession pour les deux méthodes les plus efficaces. Les résultats indiquent que la méthode X-11-ARMMI avec facteurs saisonniers courants, qui est la méthode de désaisonnalisation officielle de Statistique Canada, produit des EAM inférieures à celles produites par la méthode X-11 avec facteurs saisonniers courants. La plupart des ratios de la première colonne sont très près de l'unité; nous pouvons en conclure que l'utilisation des extrapolations ARMMI entraîne des révisions du même ordre en période de récession comme en période de non-récession. Lorsqu'on applique la méthode X-11 avec facteurs saisonniers courants, les révisions sont sensiblement plus élevées en période de récession qu'en période "normale" pour la plupart des séries. Cette divergence tient au fait que les variations rapides de l'ordre de grandeur de la série, provoquées par les nouvelles observations relatives aux années de récession, ne sont pas estimées avec autant de précision par les filtres aux extrémités. De fait, des mouvements progressifs et une partie de l'augmentation de l'ordre de grandeur de la série sont imputés à la composante saisonnière.

La seule série qui fait exception est celle du chômage chez les femmes de 15 à 24 ans; les deux méthodes entraînent dans ce cas des révisions plus faibles en période de difficultés économiques. Cela s'explique par le comportement particulier de cette série au cours de la période analysée, lequel est caractérisé par de fortes augmentations annuelles (environ 15% entre 1966 et 1973 et 8.5% entre 1973 et 1980) et une composante saisonnière additive, indépendante du cycle économique (c'est-à-dire que la variation de l'ordre de grandeur de la série exprimait plus un changement de comportement des jeunes femmes qu'un effet du cycle économique).

La série du chômage chez les hommes de 25 ans et plus constitue un autre cas particulier. En dépit d'extrapolations ARMMI, les années de récession présentent en effet des révisions beaucoup plus importantes que celles des années de non-récession, comme en fait foi le rapport de 1.42. Cet écart appréciable entre les révisions des deux périodes s'explique par les très fortes variations que subit la composition du mouvement saisonnier de cette série en période de récession. Sans extrapolations ARMMI, l'écart entre les révisions des deux périodes est encore plus grand (ratio de 1.59) puisque, outre les variations de la composition du mouvement saisonnier, les estimations saisonnières peu précises produites pour les années de récession contribuent à accroître cet écart.

3. DÉSAISONNALISATION COURANTE DES SÉRIES DE COMPOSITION ADDITIVE OU MULTIPLICATIVE EN PÉRIODES DE RÉCESSION ET DE NON-RÉCESSION

On soutient souvent que le modèle de décomposition additif convient mieux que le modèle de décomposition multiplicatif en période de récession, si l'on vise à minimiser les révisions. Deux grandes raisons sont invoquées à l'appui de cette affirmation: (1) dans un modèle additif, les composantes de la série chronologique sont supposées indépendantes; ainsi, l'effet saisonnier ne se trouve pas influencé par le niveau de la tendance-cycle, contrairement à ce qui se passe dans un modèle multiplicatif; (2) les filtres aux extrémités sont trop rigides pour permettre d'estimer avec précision un mouvement saisonnier rapide.

Tableau 4

Comparaison de l'EAM(N) produite par la méthode X-11-ARMMI avec facteurs saisonniers courants et des EAM(N) produites par trois autres méthodes pour la désaisonnalisation des séries de composition multiplicative sur l'emploi et le chômage en période de non-récession ($N = 60$)

Séries	X-11 facteurs saisonniers courants / X-11-ARMMI facteurs saisonniers courants (1)[a]	X-11-ARMMI facteurs saisonniers prévus / X-11-ARMMI facteurs saisonniers courants (2)[b]	X-11 facteurs saisonniers prévus / X-11-ARMMI facteurs saisonniers courants (3)[c]
Chômage			
Hommes, 25 ans et plus	1.26	1.62	1.99
Femmes, 25 ans et plus	1.31	1.59	1.93
Hommes, 15 à 24	1.35	1.61	2.01
Femmes, 15 à 24	1.49	1.34	1.65
Emploi			
Hommes, 25 ans et plus	1.25	1.50	1.62
Femmes, 25 ans et plus	1.17	1.43	1.48
Hommes, 15 à 24	1.18	1.49	1.77
Femmes, 15 à 24	1.14	1.58	1.86

[a] La colonne (1) est égale au quotient des colonnes (2) et (1) du tableau 3.
[b] La colonne (2) est égale au quotient des colonnes (3) et (1) du tableau 3.
[c] La colonne (3) est égale au quotient des colonnes (4) et (1) du tableau 3.

Tableau 5

Comparaison des EAM(N) des facteurs saisonniers courants des méthodes X-11-ARMMI et X-11 appliquées à des séries de composition multiplicative en période de récession et en période de non-récession

Séries	X-11-ARIMA facteurs courants période de récession ($N = 24$) / période de non-récession ($N = 60$) (1)[a]	X-11 facteurs courants période de récession ($N = 24$) / période de non-récession ($N = 60$) (2)[b]
Chômage		
Hommes, 25 ans et plus	1.42	1.59
Femmes, 25 ans et plus	1.05	1.22
Hommes, 15 à 24	1.09	1.35
Femmes, 15 à 24	0.67	0.60
Emploi		
Hommes, 25 ans et plus	1.00	1.20
Femmes, 25 ans et plus	1.00	1.07
Hommes, 15 à 24	1.05	1.27
Femmes, 15 à 24	1.16	1.54

[a] La colonne (1) est égale au quotient de la colonne (1) du tableau 1 par la colonne (1) du tableau 3.
[b] La colonne (2) est égale au quotient de la colonne (2) du tableau 1 par la colonne (2) du tableau 3.

Afin de vérifier ces affirmations, nous avons désaisonnalisé les huit séries sur la population active analysées précédemment suivant un modèle additif. Les résultats de cet exercice confirment ceux qui ont été obtenus au moyen d'un modèle multiplicatif. En effet, la méthode X-11-ARMMI avec facteurs saisonniers courants est celle qui entraîne les plus faibles révisions; viennent ensuite les méthodes X-11 avec facteurs saisonniers courants et X-11-ARMMI avec facteurs prévus une année à l'avance. La méthode X-11 avec facteurs prévus une année à l'avance est celle qui produit les estimations les moins précises. Il importe de souligner que, dans la désaisonnalisation de séries de modèle additif, les facteurs sont des *facteurs implicites* en ce sens qu'ils découlent du quotient de la série d'origine par la série désaisonnalisée.

Les tableaux 6 et 7 comparent l'importance relative des révisions produites par chacune des trois méthodes moins efficaces et de celles produites par la méthode X-11-ARMMI avec facteurs saisonniers courants pour les périodes de récession et de non-récession respectivement. Toutes les valeurs des tableaux sont supérieures à l'unité, ce qui indique qu'aucune des trois méthodes n'entraîne des révisions inférieures à celles produites par la méthode X-11-ARMMI avec facteurs saisonniers courants. Comme celle-ci s'avère la plus efficace tant dans la version additive que dans la version multiplicative, nous allons chercher à savoir, à l'aide de chaque série, lequel des deux modèles de décomposition produit les plus faibles révisions.

En ce qui a trait aux deux séries qui influent le plus sur le taux de chômage (c'est-à-dire, le chômage chez les hommes de 25 ans et plus et l'emploi chez les hommes de 25 ans et plus), les données du tableau 8 indiquent que l'utilisation d'un modèle multiplicatif est préférable en période de récession comme en période de non-récession. Ces données confirment pour la plupart les modèles de décomposition qu'a choisis Statistique Canada en se fondant sur les tests de modèles (Morry 1975; Higginson 1977). Seule la série de l'emploi chez les hommes de 15 à 24 ans semble faire exception à la règle puisque, dans ce cas, un modèle additif serait plus efficace. Mais comme les révisions sont déjà très faibles, un changement de modèle n'apporterait rien de plus. Les EAM, qui sont de 0.41 (période de récession) et de 0.39 (période de non-récession) avec un modèle multiplicatif, sont réduites à 0.33 et à 0.31 respectivement avec un modèle additif.

Enfin, les données du tableau 8 nous permettent de constater qu'un modèle multiplicatif conviendrait mieux qu'un modèle additif, en période de récession, pour la série du chômage chez les femmes de 25 ans et plus.

4. CONCLUSIONS

Les résultats présentés dans les sections 2 et 3 peuvent se résumer comme suit:

1) En période de récession comme en période de non-récession, la méthode X-11-ARMMI avec facteurs saisonniers courants entraîne les plus faibles révisions pour chaque série, peu importe si la série est de composition additive ou multiplicative.
2) Les ratios entre l'EAM produite par la méthode X-11-ARMMI (facteurs saisonniers courants) appliquée à un modèle additif et l'EAM produite par la même méthode appliquée à un modèle multiplicatif indiquent clairement que les deux séries qui influent le plus sur le taux de chômage (chômage chez les hommes de 25 ans et plus et emploi chez les hommes de 25 ans et plus) sont de composition multiplicative en période de récession comme en période de non-récession.
3) En période de récession, les méthodes X-11-ARMMI avec facteurs prévus une année à l'avance et X-11 avec facteurs saisonniers courants produisent des EAM identiques pour les séries de l'emploi et du chômage chez les hommes de 25 ans et plus. Pour ce qui a trait aux six autres séries toutefois, la méthode X-11 avec facteurs saisonniers courants s'avère la plus efficace après la X-11-ARMMI (facteurs saisonniers courants).

Tableau 6

Comparaison de l'EAM(N) produite par la méthode X-11-ARMMI avec facteurs saisonniers courants et des EAM(N) produites par trois autres méthodes pour la désaisonnalisation des séries de composition additive sur l'emploi et le chômage en période de récession ($N = 24$)

Séries	X-11 facteurs saisonniers courants	X-11-ARIMA facteurs implicites prévus	X-11 facteurs implicites prévus
	X-11-ARMMI facteurs saisonniers courants	X-11-ARMMI facteurs saisonniers courants	X-11-ARMMI facteurs saisonniers courants
Chômage			
Hommes, 25 ans et plus	1.18	1.29	1.38
Femmes, 25 ans et plus	1.16	1.49	1.75
Hommes, 15 à 24	1.21	1.48	1.70
Femmes, 15 à 24	1.33	1.74	1.84
Emploi			
Hommes, 25 ans et plus	1.44	1.69	2.08
Femmes, 25 ans et plus	1.26	1.33	1.65
Hommes, 15 à 24	1.02	1.05	1.34
Femmes, 15 à 24	1.50	1.50	2.05

Tableau 7

Comparaison de l'EAM(N) produites par la méthode X-11-ARMMI avec facteurs saisonniers courants et des EAM(N) produites par trois autres méthodes pour la désaisonnalisation des séries de composition additive sur l'emploi et le chômage en période de non-récession ($N = 60$)

Series	X-11 facteurs saisonniers courants	X-11-ARIMA facteurs saisonniers prévus	X-11 facteurs implicites prévus
	X-11-ARMMI facteurs saisonniers courants	X-11-ARMMI facteurs saisonniers courants	X-11-ARMMI facteurs saisonniers courants
Chômage			
Hommes, 25 ans et plus	1.31	1.65	1.88
Femmes, 25 ans et plus	1.20	1.59	1.71
Hommes, 15 à 24	1.22	1.57	1.89
Femmes, 15 à 24	1.05	1.20	1.26
Emploi			
Hommes, 25 ans et plus	1.16	1.24	1.54
Femmes, 25 ans et plus	1.10	1.27	1.30
Hommes, 15 à 24	1.22	1.31	1.55
Femmes, 15 à 24	1.41	1.68	2.16

Tableau 8

Comparaison de l'EAM(N) produite par la méthode X-11-ARMMI (facteurs saisonniers courants) appliquée à un modèle additif et de l'EAM(N) produite par la même méthode appliquée à un modèle multiplicatif en périodes de récession et de non-récession

Séries	(N = 24) Période de récession	(N = 60) Période de non-recession
	Modèle additif X-11-ARIMA facteurs saisonniers courants	Modèle additif X-11-ARIMA facteurs saisonniers courants
	Modèle multiplicatif X-11-ARMMI facteurs saisonniers courants	Modèle multiplicatif X-11-ARMMI facteurs saisonniers courants
Chômage		
Hommes, 25 ans et plus	1.25	1.15
Femmes, 25 ans et plus	1.14	0.88
Hommes, 15 à 24	1.23	1.05
Femmes, 15 à 24	0.93	0.85
Emploi		
Hommes, 25 ans et plus	1.25	1.25
Femmes, 25 ans et plus	1.00	1.00
Hommes, 15 à 24	0.80	0.80
Femmes, 15 à 24	1.14	1.17

4) La méthode X-11 avec facteurs saisonniers prévus une année à l'avance est celle qui produit les estimations les moins précises pour toutes les séries et dans toutes les situations étudiées.
5) En comparant les EAM produites par la méthode X-11-ARMMI (facteurs saisonniers courants) en période de récession et en période de non-récession, nous constatons qu'elles sont à peu près du même ordre, sauf en ce qui a trait à la série du chômage chez les hommes de 25 ans et plus, pour laquelle les révisions sont beaucoup plus fortes en période de récession. Ce cas d'exception s'explique par les brusques variations saisonnières auxquelles cette série est soumise à cause des changements profonds que subit sa composition. Les révisions sont plus importantes à cause principalement de ces modifications.

   Par ailleurs, la méthode X-11 avec facteurs saisonniers courants entraîne, pour la plupart des séries, des révisions très différentes selon qu'il s'agit d'une période de récession ou d'une période de non-récession. Cette divergence indique que les révisions tiennent surtout au fait que les filtres aux extrémités ne permettent pas d'estimer avec précision les ordres de grandeur, très volatiles, observés en période de récession.

   Il n'y a qu'une série pour laquelle les deux méthodes les plus efficaces entraînent des révisions beaucoup plus fortes en période de non-récession; c'est celle du chômage chez les femmes de 15 à 24 ans. Cela s'explique par le comportement particulier de cette série au cours de la période observée, lequel est caractérisé par de fortes augmentations annuelles (environ 15% entre 1966 et 1973 et 8.5% entre 1973 et 1980), qui masquent l'effet du cycle économique, et par une composante saisonnière indépendante de ce cycle.

Compte tenu des observations précédentes, nous pouvons affirmer que la méthode de désaisonnalisation officielle de Statistique Canada produira les meilleures estimations possibles en période de récession.

REMERCIEMENT

Les auteures tiennent à exprimer leur reconnaissance aux deux arbitres qui, par leurs précieux conseils, ont contribué à parfaire cette étude.

BIBLIOGRAPHIE

BOX, G.E.P., et JENKINS, G.M. (1970). *Times Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

DAGUM, E.B. (1978). *Comparison and Assessment of Seasonal Adjustment Methods for Labor Force Series*. Washington, D.C.: U.S. Government Printing Office.

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

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

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

DAGUM, E.B., et MORRY, M. (1982). L'estimation des variations saisonnières dans les indices des prix à la consommation. Actes du colloque sur *La mesure du niveau des prix*, N° 22-24 au catalogue, Ottawa, Canada: Statistique Canada.

HIGGINSON, J. (1977). Manuel d'utilisation du test de Bell Canada relatif à la sélection de schémas de composition des séries. Document de recherche n° 77-01-001, Groupe de la désaisonnalisation et des séries chronologiques, Statistique Canada.

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

KUIPER, J. (1978). A survey and comparative analysis of various methods of seasonal adjustment. *Seasonal Analysis of Economic Time Series*, (Ed. Arnold Zellner), Washington, D.C.: U.S. Government Printing Office, 59-76.

KUIPER, J. (1981). The treatment of extreme values in the X-11-ARIMA program. *Time Series Analysis and Forecasting*, (Eds. Anderson, O. et Perryman, M.R.), Amsterdam: North-Holland Publishing Co., 257-266.

McKENZIE, S. (1982). An evaluation of concurrent adjustment on Census Bureau time series. *Proceedings of the Business and Economics Section of the American Statistical Association*.

MORRY, M. (1975). A test for model selection. Document de recherche n° 75-12-016, Groupe de la désaisonnalisation et des séries chronologiques, Statistique Canada.

OTTO, M. (1985). Effects of forecasts on the revisions of seasonally adjusted values using the X-11 seasonal adjustment procedure. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association* (à paraître).

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

PIERCE, D., et McKENZIE S. (1985). On concurrent seasonal adjustment. Document technique, U.S. Bureau of the Census.

SHISKIN, J., YOUNG, A.H., et MUSGRAVE, J.C. (1967). The X-11 variant of census method II seasonal adjustment program. Document technique n° 15, Washington, D.C.: U.S. Bureau of Census.

WALLIS, K.F. (1974). Seasonal Adjustment and Relations Between Variables. *Journal of the American Statistical Association*, 69, 18-31.

WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: Linear filters for the X-11 Method. *Journal of the Royal Statistical Society*, série A, 145, 74-85.

YOUNG, A.H. (1968). Linear approximations to census and BLS seasonal adjustment methods. *Journal of the American Statistical Association*, 63, 445-457.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 163–179
Statistique Canada

Formes de relation entre le niveau total de chômage et le nombre de bénéficiaires de l'assurance-chômage au Canada

ESTELA BEE DAGUM, GUY HUOT, NAZIRA GAIT,
et NORMAND LANIEL[1]

RÉSUMÉ

La présente étude vise à vérifier à l'aide de séries chronologiques unidimensionnelles et multidimensionnelles s'il existe des relations temporelles entre le nombre de bénéficiaires de l'assurance-chômage, le niveau total de chômage, le nombre de personnes ayant perdu leur emploi et le nombre de personnes ayant quitté leur emploi au Canada. Les résultats indiquent que, pour la période 1975-1982, la série des bénéficiaires de l'assurance-chômage devance d'un mois la série du niveau total de chômage et de deux mois celle des personnes ayant quitté leur emploi. Par ailleurs, on constate une relation de rétroaction entre la série des bénéficiaires de l'assurance-chômage et celle des personnes ayant perdu leur emploi.

MOTS CLÉS: Personnes ayant perdu leur emploi; personnes ayant quitté leur emploi; ARMMI; modèles vectoriels ARMA; séries chronologiques multidimensionnelles.

1. INTRODUCTION

L'assurance-chômage (A-C) joue un rôle fondamental en permettant aux marchés du travail du pays de s'adapter aux variations du niveau de production et du niveau d'emploi causées par les fluctuations de la demande et des échanges commerciaux. Dans le cadre plus global d'une politique du marché du travail, l'A-C a pour principal objet d'assurer une protection financière convenable aux personnes qui sont temporairement en chômage pour leur permettre de vivre plus facilement cette période de transition. Comme l'A-C élimine la principale menace découlant du chômage, c'est-à-dire la perte de revenus, les chercheurs d'emploi ne se sentent plus obligés de céder aux pressions économiques en acceptant des emplois qui ne conviennent pas à leurs compétences ou à leurs aptitudes. En permettant une recherche d'emploi plus systématique ou plus variée, l'A-C favorise une réaffectation efficiente des ressources humaines. En outre, lorsqu'il y a des fermetures temporaires d'usines, l'A-C atteint son objectif en accordant une protection financière aux travailleurs touchés, de telle sorte que l'employeur peut rappeler à tout moment la même main-d'oeuvre expérimentée dont il bénéficiait auparavant. À la réouverture de l'usine, il n'a donc pas besoin de recruter et de former de nouveaux employés et évite, par le fait même, les dépenses inhérentes à ces activités. L'A-C permet en outre à l'employé de traverser sans trop d'ennuis la période de chômage et, partant, d'échapper aux difficultés financières.

Peu importe la situation, l'A-C doit être suffisamment flexible pour tenir compte de la conjoncture économique du moment, laquelle peut restreindre le nombre d'emplois disponibles et prolonger la période de chômage des chercheurs d'emploi. Dans le programme canadien d'assurance-chômage, cette flexibilité se traduit par une prolongation de la durée des prestations lorsque les taux de chômage dans les régions sont à la hausse.

L'écart entre la série du niveau total de chômage et celle des bénéficiaires de l'assurance-chômage tend à s'amenuiser en période de récession et à s'élargir en période de reprise. Lorsque la conjoncture économique se détériore et que les mises à pied surviennent, les personnes

[1] E.B. Dagum et G. Huot, Division des séries chronologiques recherche et analyse, Statistique Canada, N. Gait, Université de Sao Paulo, Brésil, cette dernière effectuait un séjour à Statistique Canada lorsque l'article a été écrit, et N. Laniel, Division des séries chronologiques recherche et analyse maintenant avec la Division des méthodes d'enquêtes-entreprises.

ayant perdu leur emploi représentent une plus forte proportion du nombre total de chômeurs. Comme la plupart des prestataires de l'assurance-chômage sont, de fait, des personnes qui ont perdu leur emploi, la proportion des bénéficiaires de l'assurance-chômage se trouve augmentée par rapport au niveau total de chômage.

La présente étude vise à vérifier s'il existe une relation temporelle entre le nombre de bénéficiaires de l'assurance-chômage (BAC) et le niveau total de chômage au Canada (NTC). L'analyse tient compte également des personnes ayant perdu leur emploi (JLo) et de celles ayant quitté leur emploi (JLe), deux groupes qui ont droit à des prestations et qui représentent la grande majorité des chômeurs au pays. L'existence d'une relation étroite entre ces variables pourrait servir à expliquer le comportement des marchés du travail. En outre, si une telle relation existe, elle peut nous amener à en découvrir d'autres du même genre, qui serviraient à estimer le niveau de chômage dans les petites régions où la taille de l'échantillon de l'enquête sur la population active est insuffisante. Dans la section 2, nous définissons chacune des quatre séries étudiées et analysons leurs principales caractéristiques à partir de leurs spectres respectifs. Dans la section 3, nous estimons, pour plusieurs décalages, les corrélations avec décalage des résidus des séries affectées d'un bruit blanc pour déterminer s'il existe des relations par paire et définir leur direction, s'il y a lieu. Les résidus sont calculés à l'aide de modèles ARMMI ajustés à chaque série. Dans la section 4, nous poursuivons les analyses des sections précédentes en définissant et en estimant deux modèles de séries chronologiques multidimensionnelles afin de comprendre la corrélation dynamique entre 1) les séries BAC et NTC et 2) les séries BAC, JLo et JLe. Enfin, la section 5 présente les principales conclusions de cette étude.

2. PRINCIPALES CARACTÉRISTIQUES DES SÉRIES ANALYSÉES

Pour comprendre le genre de relations qui existent entre la série BAC et la série NTC et ses principales composantes, JLo et JLe, nous commençons par définir ces séries et analyser leurs principales caractéristiques à l'aide de leurs spectres.

2.1 Niveau total de chômage (NTC)

La Division de l'enquête sur la population active (EPA) de Statistique Canada recueille des données mensuelles au moyen d'une enquête menée auprès de 56,000 ménages représentatifs répartis dans tout le pays. Malgré les améliorations qui ont été apportées depuis 1952, l'EPA a subi d'importantes modifications à partir de 1976.

Les estimations du nombre de personnes occupées, du nombre de chômeurs et du nombre d'inactifs portent sur la semaine de référence de l'enquête mensuelle, habituellement la semaine incluant le 15^{e} jour du mois. L'échantillon de l'enquête représente, à quelques exceptions près, toutes les personnes de 15 ans et plus qui résident au Canada.

La population active comprend les personnes qui avaient un emploi ou étaient en chômage pendant la semaine de référence. Les personnes occupées sont celles qui:

- ont fait un travail quelconque;
- avaient un emploi mais n'étaient pas au travail pour l'une des raisons suivantes: maladie ou invalidité, mauvais temps, conflit de travail, vacances ou obligations personnelles ou familiales.

Le groupe des chômeurs comprend les personnes qui:

- étaient sans travail mais avaient activement cherché du travail au cours des quatre dernières semaines et étaient prêtes à travailler;
- n'avaient pas activement cherché de travail au cours des quatre dernières semaines mais avaient été mises à pied depuis 26 semaines ou moins et étaient prêtes à travailler;
- n'avaient pas activement cherché de travail au cours des quatre dernières semaines mais devaient commencer à travailler à un nouvel emploi dans quatre semaines ou moins et étaient prêtes à travailler.

La série du niveau total de chômage regroupe les personnes ayant perdu leur emploi (JLo), celles ayant quitté leur emploi (JLe), les nouveaux venus sur le marché du travail, les personnes qui reviennent sur le marché du travail après un an ou moins et celles qui y reviennent après plus d'un an (Statistique Canada 1976). Les deux premiers groupes sont ceux qui comptent le plus pour notre étude puisque leurs membres sont admissibles à des prestations et représentent environ 70% de tous les chômeurs.

Comme il n'est pas possible d'obtenir de données sur l'augmentation du nombre de chômeurs pour la période précédant l'année 1975, toutes les séries analysées couvrent la période de janvier 1975 à décembre 1982 et comprennent, par conséquent, les données les plus récentes possibles.

La figure 1 illustre la série originale du niveau total de chômage, laquelle est caractérisée par un sommet dans les mois d'hiver et par un creux dans les mois d'été. La figure 2 décrit le spectre de cette série. On y observe une puissance élevée entre les fréquences 0.00 et 0.05, cette dernière équivalant à un cycle économique ($f = 0.05$ correspond à un cycle de 20 mois). De même, la puissance est relativement élevée autour de la fréquence saisonnière fondamentale (0.083) mais est moindre aux bandes harmoniques. Enfin, les fluctuations irrégulières comptent pour peu dans la variance totale par rapport aux deux autres composantes.

La figure 3 montre la série originale des personnes ayant perdu leur emploi et la Figure 4 décrit le spectre correspondant. Comme dans le cas de la série NTC, une forte puissance est observée aux fréquences correspondant à des cycles économiques mais, cette fois, la puissance saisonnière est concentrée dans la bande saisonnière fondamentale et est presque absente dans les bandes harmoniques. L'effet des fluctuations irrégulières est encore plus faible dans cette série.

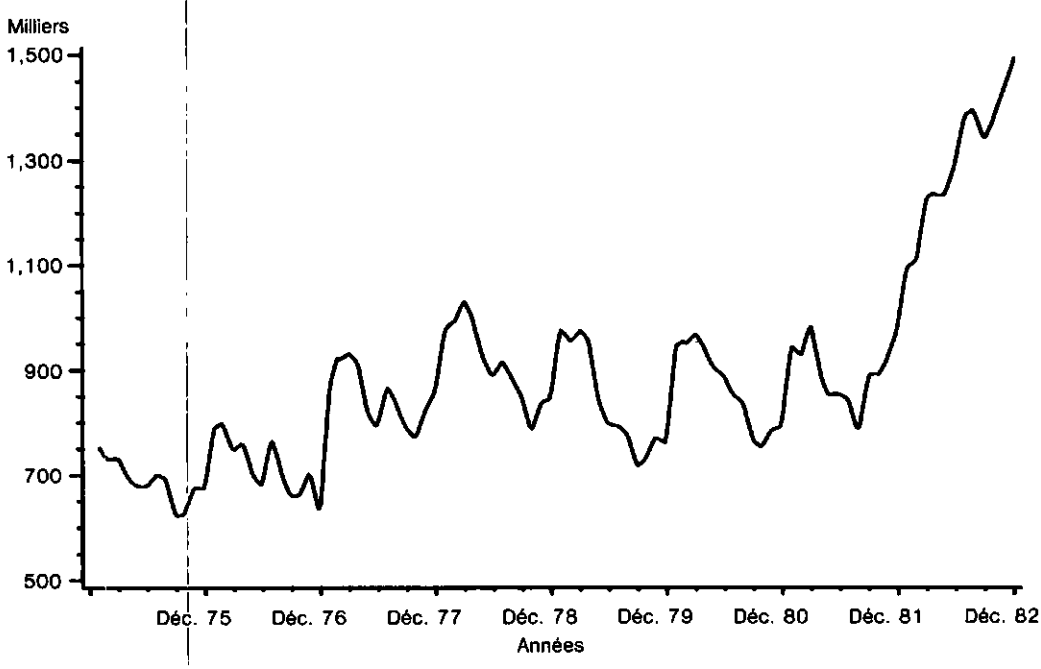


Figure 1. Série du niveau total de chômage

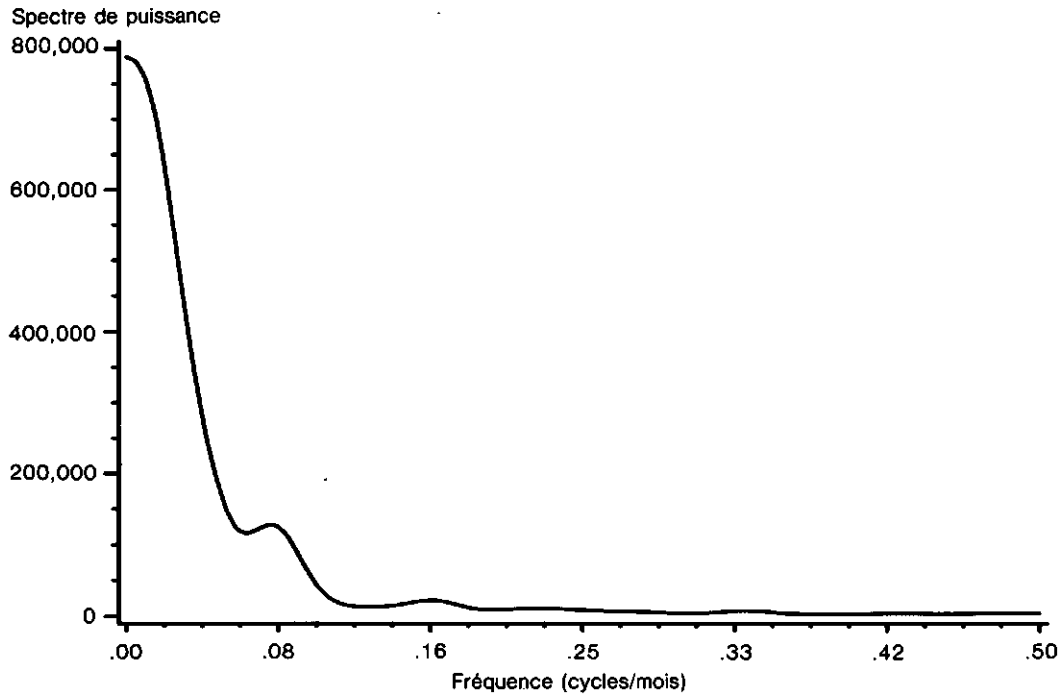


Figure 2. Spectre de la série du niveau total de chômage

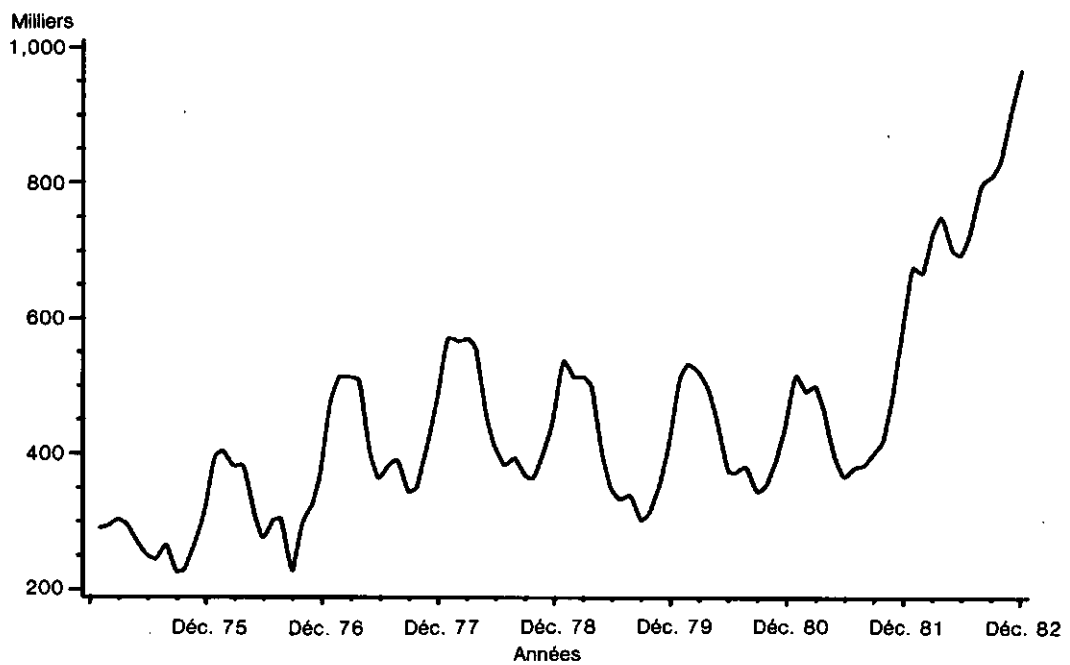


Figure 3. Série des personnes ayant perdu leur emploi

La figure 5 illustre la série des personnes ayant quitté leur emploi, laquelle est caractérisée par deux creux, un pour les mois d'hiver et l'autre pour les mois d'été. Le spectre de cette série est décrit à la Figure 6. Il indique que la puissance aux basses fréquences est surtout concentrée au niveau des fréquences correspondant au cycle économique, comme l'indique le sommet observé à la fréquence 0.022 (laquelle correspond à un cycle de 45 mois). En outre, les variations saisonnières sont concentrées autour de la première bande harmonique, ce qui confirme le fait que cette série présente deux creux saisonniers. Enfin, l'effet des fluctuations irrégulières sur la variance totale est plus notable que dans les deux séries précédentes.

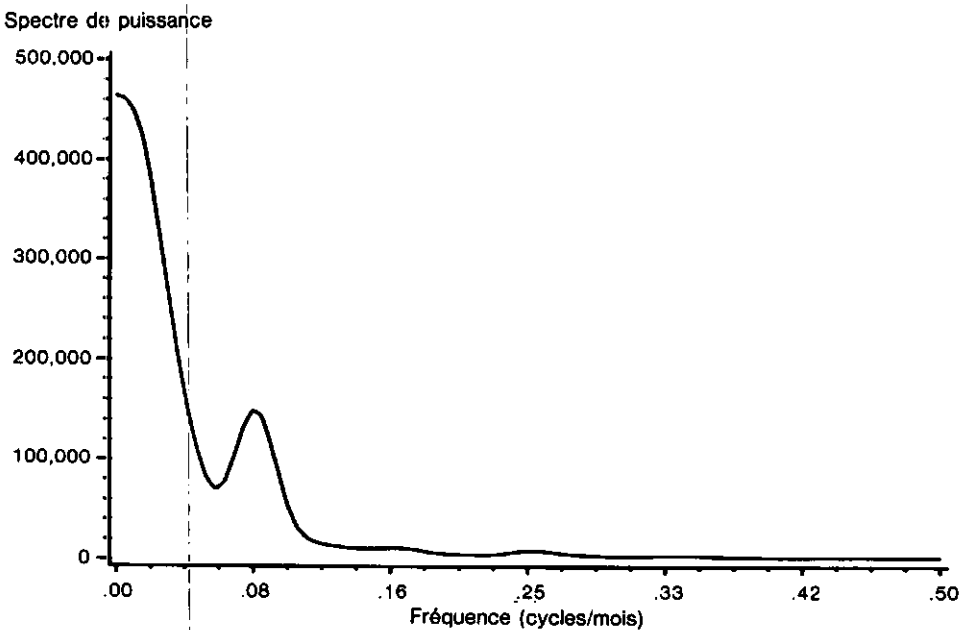


Figure 4. Spectre de la série des personnes ayant perdu leur emploi

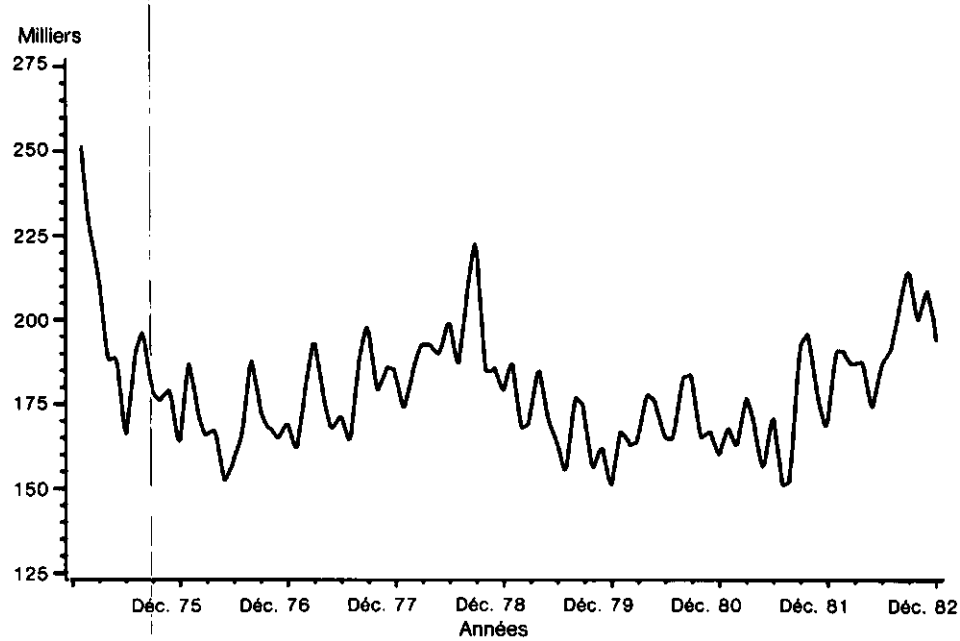


Figure 5. Série des personnes ayant quitté leur emploi

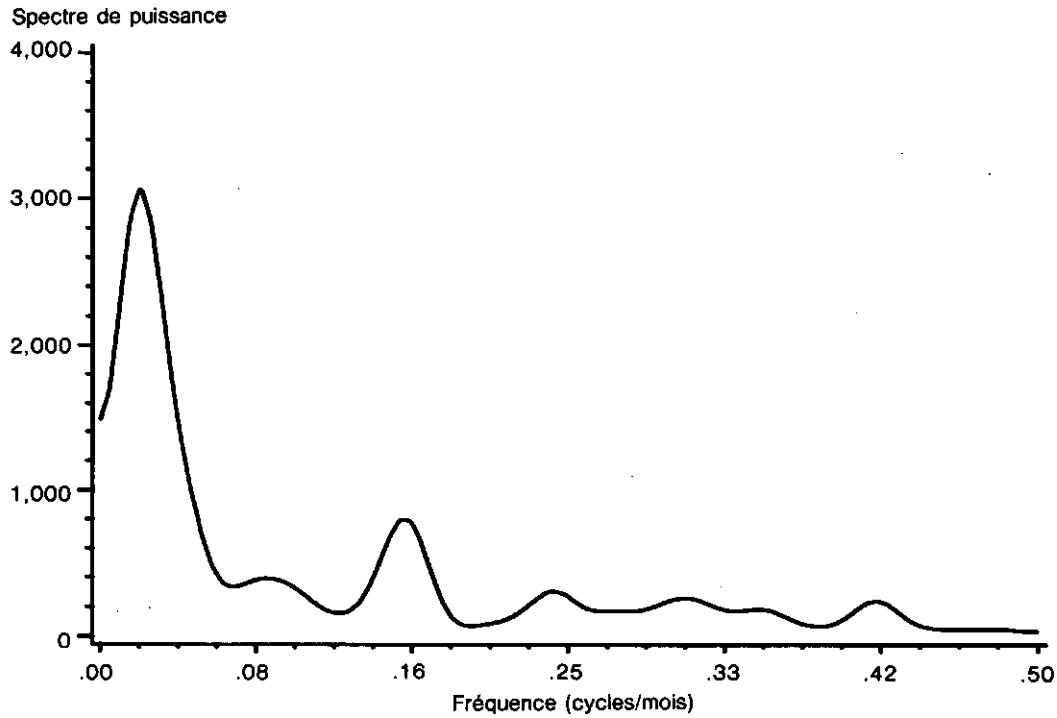


Figure 6. Spectre de la série des personnes ayant quitté leur emploi

2.2 Bénéficiaires de l'assurance-chômage (BAC)

Les données mensuelles concernant les bénéficiaires de l'assurance-chômage couvrent toutes les personnes qui reçoivent des prestations pour une semaine déterminée, notamment la semaine de référence de l'EPA. Il ne s'agit pas d'un échantillon puisque les données concernent l'ensemble des bénéficiaires. L'A-C vise pratiquement tous les membres de la population active qui reçoivent une rémunération ainsi que les membres des forces armées. Les principales exceptions sont :

- les personnes de 65 ans et plus;
- Les personnes qui travaillent moins de 15 heures par semaine;
- Les personnes qui gagnent moins de 20% de la rémunération hebdomadaire assurable maximum (en 1982, ce maximun était de $70).

Pour être admissible à des prestations, une personne doit être disponible pour travailler et être apte à occuper un emploi, capable de se trouver un emploi convenable et satisfaire aux conditions d'admissibilité. Auparavant, huit semaines de travail suffisaient à une personne pour qu'elle soit admissible à des prestations mais, depuis décembre 1977, ce nombre varie de 10 à 14 semaines selon le taux de chômage enregistré dans la région où le prestataire a son domicile. Les prestations commencent à être versées après une période de carence de deux semaines.

Les personnes admissibles à des prestations d'assurance-chômage peuvent gagner un revenu d'emploi n'excédant pas 25% du montant de leurs prestations et continuer de recevoir ces prestations. Cependant, ces personnes sont considérées comme occupées dans l'EPA. Si nous voulons déterminer la relation entre le nombre de bénéficiaires de l'assurance-chômage et le niveau total de chômage, il serait donc plus convenable d'utiliser la série des bénéficiaires de l'assurance-chômage qui ne reçoivent *aucune* rémunération. Ce sous-ensemble des

bénéficiaires de l'assurance-chômage représente une proportion relativement stable et importante du nombre total de chômeurs établi selon l'FPA. Il convient toutefois de souligner qu'à cause d'une différence de définitions, certaines personnes sont considérées comme des chômeurs dans l'EPA mais ne figurent pas dans les dossiers de l'A-C. Il s'agit, notamment, des nouveaux venus et des personnes qui reviennent sur le marché du travail, de toutes les personnes qui n'ont pas travaillé suffisamment longtemps pour avoir droit à des prestations et des personnes en chômage qui étaient auparavant des travailleurs autonomes. En revanche, certaines catégories de personnes qui sont admissibles aux prestations d'assurance-chômage ne sont pas considérées comme des chômeurs selon l'EPA. Mentionnons, par exemple, les pêcheurs indépendants, au cours de la saison morte, les femmes qui bénéficient d'un congé de maternité et les employés qui sont absents de leur poste pour cause de maladie ou d'invalidité.

La série des bénéficiaires de l'A-C (sans rémunération) est un indicateur sensible de la conjoncture du marché du travail. Elle trace un portrait fidèle de la main-d'o .e qui a récemment quitté le marché du travail et qui bénéficie de l'assurance-chômage.

Comme on peut le voir à la figure 7, la série originale des bénéficiaires de l'assurance-chômage connaît de fortes fluctuations saisonnières; son mouvement est caractérisé par des sommets durant les mois d'hiver, lorsque les mauvaises conditions atmosphériques limitent le travail à l'extérieur dans des secteurs comme la pêche, la construction et la coupe de bois d'oeuvre, ce qui provoque une forte hausse du nombre de demandes de prestations.

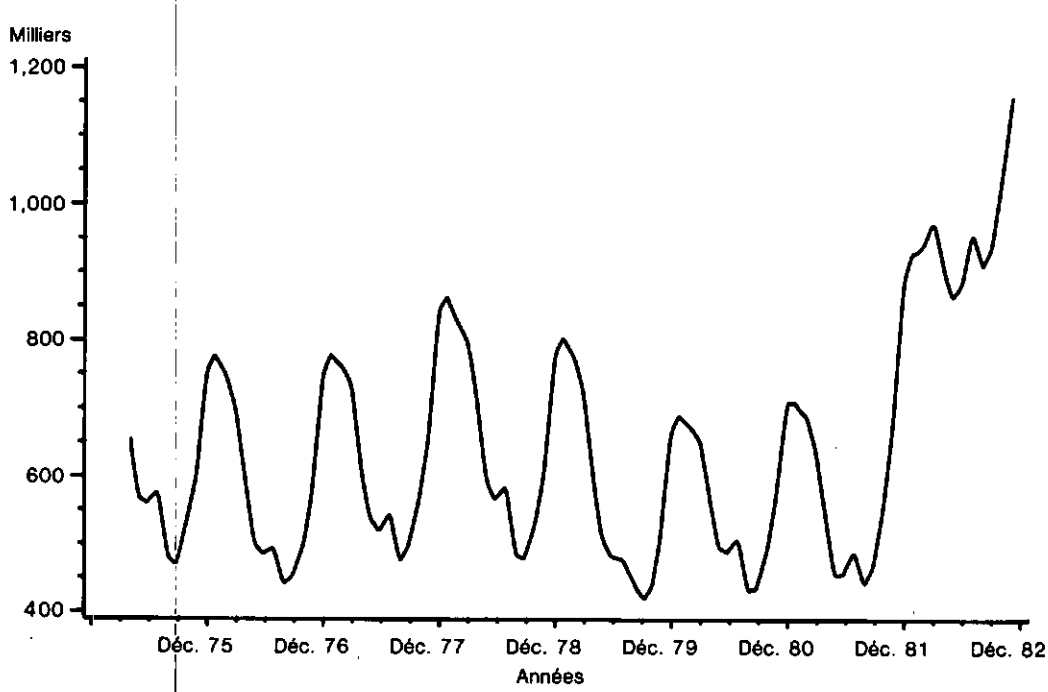


Figure 7. Série des bénéficiaires de l'assurance-chômage

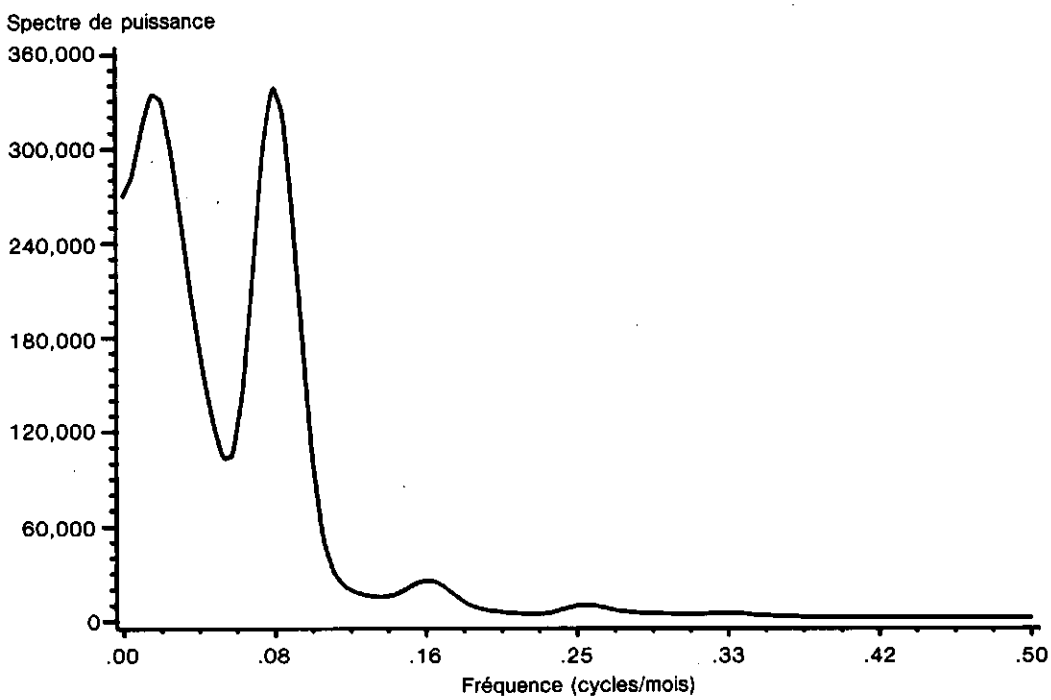


Figure 8. Spectre de la série des bénéficiaires de l'assurance-chômage

La figure 8 donne le spectre de la série des bénéficiaires de l'assurance-chômage. On y observe une très forte puissance à la fréquence 0.0167, qui correspond à un cycle de 60 mois, ainsi qu'aux fréquences associées à la bande saisonnière fondamentale. Les variations saisonnières comptent pour beaucoup plus dans la variance totale de la série qu'elles ne le faisaient pour la série NTC et ses deux grandes composantes. Enfin, les fluctuations irrégulières sont négligeables par rapport à la tendance-cycle et aux composantes saisonnières.

3. RELATIONS PAR PAIRE ENTRE LE NOMBRE DE BÉNÉFICIAIRES DE L'ASSURANCE-CHÔMAGE, LE NIVEAU TOTAL DE CHÔMAGE, LE NOMBRE DE PERSONNES AYANT PERDU LEUR EMPLOI ET LE NOMBRE DE PERSONNES AYANT QUITTÉ LEUR EMPLOI

Plusieurs études réalisées au Canada il y a déjà un certain nombre d'années (par exemple, Grubel et coll.. 1975; Green et Cousineau 1976; Jump et Rea 1975; et Siedule et coll. 1976) viennent confirmer l'affirmation générale selon laquelle la libéralisation du programme d'assurance-chômage en 1971 a eu tendance à accroître le niveau de chômage. Lazar (1978) montre que les modifications de 1971 ont eu pour effet d'accroître la durée du chômage et d'inciter un nombre proportionnellement plus élevé de personnes, surtout chez les jeunes et les femmes adultes, à quitter leur emploi. Ces études ont été faites avant les modifications de 1975 qui visaient à revaloriser les formes d'encouragement au travail. On espérait que les modifications apportées après 1975 supprimeraient les effets nuisibles du programme sur le niveau total de chômage.

Dans la présente section, nous nous livrons à une analyse préliminaire en tentant de découvrir des relations temporelles par paire entre le niveau total de chômage, le nombre de bénéficiaires de l'assurance-chômage, le nombre de personnes ayant perdu leur emploi et le nombre de celles ayant quitté leur emploi. L'existence de telles relations permettra de construire un modèle de séries chronologiques multidimensionnelles, qui servira à expliquer la dynamique des variables énumérées ci-dessus.

Les relations par paire entre les séries NTC, BAC, JLo et JLe sont déterminées au moyen des corrélations avec retard des résidus ou *innovations* des modèles ARMMI (Box et Jenkins, 1970) qui donnent une bonne approximation des données. Plusieurs auteurs dont Pierce et Haugh (1977) ont fait valoir avec raison que les corrélations avec décalage entre les bruits blancs obtenus au moyen de filtres différents sont telles qu'elles favorisent l'acceptation de l'hypothèse nulle de l'indépendance, alors que celle-ci n'existe pas. Pierce et Haugh (1977) proposent d'utiliser des modèles de régression dynamiques. Il faut toutefois, pour cela, déterminer quelle variable est la *cause* et quelle variable est l'*effet*. Pour l'instant, nous cherchons seulement à savoir s'il existe une relation temporelle dans chaque paire de variables analysée. Le tableau 1 donne les modèles ARMMI ajustés à chaque série, leurs paramètres estimés au moyen des moindres carrés inconditionnels, les résultats du test du *portmanteau* (Ljung et Box, 1978) et la variance résiduelle.

Les valeurs du paramètre statistique Q se trouvent dans l'intervalle d'acceptation de l'hypothèse nulle, selon laquelle les résidus présentent un caractère aléatoire dans chaque cas. Comme, toutefois, ce test est appliqué à un ensemble d'autocorrélations de résidus pour divers décalages, il pourrait ne pas révéler l'existence d'une forte autocorrélation pour un certain décalage k. Par conséquent, nous avons aussi vérifié s'il existait une autocorrélation de résidus pour chaque décalage. Pour tester la variance de l'autocorrélation, nous nous sommes servis d'une formule d'approximation plus précise que $1/N$ pour les petits échantillons, c'est-à-dire $(N - |k|)N^{-2}$ tell que la définit Haugh (1976).

Ayant obtenu des résultats satisfaisants avec ces modèles, nous avons calculé le coefficient de corrélation avec décalage $\hat{r}_{xy}(k)$ entre les séries analysées. On se sert du paramètre statistique S_M^* (Haugh 1976) pour tester l'indépendance des séries en supposant que les résidus sont distribués suivant une loi normale que $E[\hat{r}_{xy}(k)] = 0$ et Var $[\hat{r}_{xy}(k)] = (N - |k|)N^{-2}$ et que

$$S_M^* = N^2 \sum_{k=-M}^{M} (N - |k|)^{-1} \hat{r}_{xy}(k)^2$$

suit une distribution de *khi*-carré avec $2M + 1$ degrés de liberté. Afin de déterminer la direction des relations par paire, nous avons modifié le paramètre statistique S_M^*, qui n'a été calculé que pour des valeurs positives ou négatives de k, c'est-à-dire à l'exclusion de la valeur nulle.

Le tableau 2 donne les estimations de la corrélation avec décalage entre la série des bénéficiaires de l'assurance-chômage (BAC) et la série du niveau total de chômage (NTC) et ses deux principales composantes, les personnes ayant perdu leur emploi (JLo) et les personnes ayant quitté leur emploi (JLe). Certaines valeurs du tableau sont marquées d'un (a) ou d'un (b) suivant qu'elles sont significatives à un seuil de confiance de 5% ou de 1% respectivement. En ce qui concerne les séries BAC et JLo, nous avons calculé S_M^* pour les valeurs de k allant de ± 1 à ± 6 et de moins ± 1 à ± 2 pour déterminer s'il existait une relation unidirectionnelle dominante. Les résultats ont indiqué qu'il n'y avait aucune direction dominante entre chaque paire de variables mais qu'il existait une relation de rétroaction entre celles-ci.

Les résultats du tableau 2 peuvent être résumés comme suit:

1) Les données indiquent une relation unidirectionnelle entre les séries BAC et NTC de sorte que BAC devancerait NTC d'un mois.
2) Il y a rétroaction entre les séries BAC et JLo et cette rétroaction est caractérisée par une forte relation instantanée. Compte tenu du décalage entre les deux variables, la rétroaction semble être amorcée par la série JLo au décalage 2.

Tableau 1

Modèles ARMMI unidimensionnels

Séries	Modèles ARMMI	Q(24)	$\hat{\sigma}_a^2$
Bénéficiaires de l'assurance-chômage (BAC)	$(1 - 0.68B)\Delta\Delta^{12} \log_{10} BAC_t = (1 - 0.80B^{12})a_t$	11.55	0.000140
Niveau total de chômage (NTC)	$(1 - 0.25B^3)\Delta\Delta^{12} \log_{10} NTC_t = (1 - 0.84B^{12})a_t$	9.13	0.000395
Personnes ayant perdu leur emploi (JLo)	$(1 - 0.31B^3)\Delta\Delta^{12} \log_{10} JLo_t = (1 - 0.67B^{12})a_t$	15.78	0.000604
Personnes ayant quitté leur emploi (JLe)	$(1 - 0.37B^3)\Delta\Delta^{12} \log_{10} JLe_t = (1 - 0.40B - 0.25B^2)(1 - 0.87B^{12})a_t$	14.58	0.000627

Tableau 2

Corrélation avec décalage entre la série des bénéficiaires de l'assurance-chômage, et la série du niveau total de chômage et ses deux grandes composantes, personnes ayant perdu leur emploi et personnes ayant quitté leur emploi

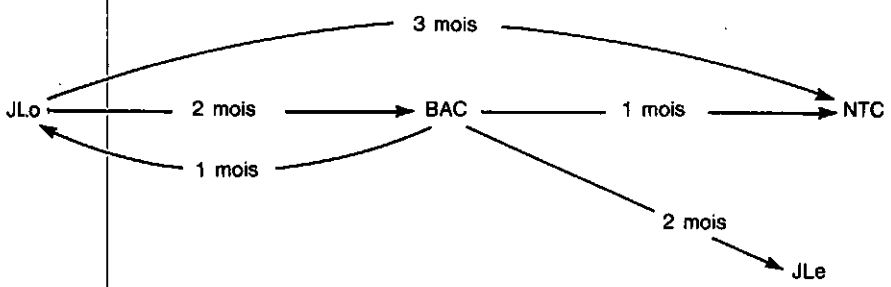
DÉCALAGES k	$BAC_{(t-k)} - NTC_t$ $\hat{r}(k)$	$BAC_{(t-k)} - JLo_t$ $\hat{r}(k)$	$BAC_{(t-k)} - JLe_t$ $\hat{r}(k)$	$JLo_{(t-k)} - NTC_t$ $\hat{r}(k)$
−6	−0.07	0.01	0.27[a]	0.15
−5	0.05	−0.04	−0.09	−0.04
−4	−0.09	0.03	−0.08	−0.01
−3	0.01	0.01	−0.11	−0.06
−2	0.14	0.28[a]	0.14	−0.01
−1	0.14	0.08	−0.01	0.21
0	0.16	0.32[b]	0.06	0.39[b]
1	0.22[a]	0.29[a]	0.04	0.14
2	0.12	0.12	0.26[a]	−0.16
3	−0.07	0.00	0.19	0.42[b]
4	0.12	−0.05	0.01	−0.05
5	−0.06	0.09	0.11	−0.04
6	0.13	0.00	0.08	0.05

[a] seuil de signification de 5%.
[b] seuil de signification de 1%.

3) On observe une relation unidirectionnelle entre les séries BAC et JLe de sorte que la première devancerait la seconde de 2 mois. On observe toutefois au décalage 6 l'effet d'une rétroaction tardive qui s'explique par le fait que la série JLe est caractérisée par deux sommets dans la même année, un en hiver et un en été, comme l'indique la figure 6.
4) Enfin, on observe une forte relation unidirectionnelle et instantanée entre les séries JLo et NTC de sorte que la première devancerait la seconde de 3 mois.

Ces observations nous ont permis de tracer le diagramme 1 qui servira à définir un modèle, plus complexe, de séries chronologiques multidimensionnelles, qui tient compte des relations partielles entre les variables.

Diagram 1



4. CONSTRUCTION D'UN MODÈLE DE SÉRIES CHRONOLOGIQUES MULTIDIMENSIONNELLES POUR LES BÉNÉFICIAIRES DE L'ASSURANCE-CHÔMAGE, LE NIVEAU TOTAL DE CHÔMAGE, LES PERSONNES AYANT PERDU LEUR EMPLOI ET LES PERSONNES AYANT QUITTÉ LEUR EMPLOI

Dans la section précédente, nous avons conclu qu'il existait des relations par paire, au sens où l'entendent Granger (1969) et Pierce et Haugh (1977), entre les quatre variables étudiées. Compte tenu de ces observations, nous voulons, dans la présente section, définir et estimer deux modèles de séries chronologiques multidimensionnelles à l'aide des méthodes élaborées par Tiao et Box (1981) et Tiao et Tsay (1983). Ces modèles serviront à expliquer la dynamique des variables en question.

Nous définissons comme suit un modèle vectoriel ARMA pour les séries saisonnières:

$$\underline{\phi}(B)\underline{\Phi}(B^s)\underline{Z}_t = \underline{\theta}(B)\underline{\Theta}(B^s)\underline{a}_t \tag{4.1}$$

où

$$\underline{\phi}(B) = \underline{I} - \underline{\phi}_1 B - \ldots - \underline{\phi}_p B^p \tag{4.2}$$

$$\underline{\Phi}(B^s) = \underline{I} - \underline{\Phi}_1 B^s - \ldots - \underline{\Phi}_P B^{sP} \tag{4.3}$$

$$\underline{\theta}(B) = \underline{I} - \underline{\theta}_1 B - \ldots - \underline{\theta}_q B_q \tag{4.4}$$

$$\underline{\Theta}(B^s) = \underline{I} - \underline{\Theta}_1 B^s - \ldots - \underline{\Theta}_Q B^{sQ} \tag{4.5}$$

sont les polynômes de matrice en B (l'opérateur de décalage défini par $B^m Z_t = Z_{t-m}$), les $\underline{\phi}$'s, $\underline{\Phi}$'s, $\underline{\theta}$'s et $\underline{\Theta}$s sont des matrices kxk, s est le mouvement périodique saisonnier, a_t est une suite de vecteurs aléatoires de changement $IID\ N(\underline{0}, \underline{\Sigma})$ et $\underline{Z}_t$ est un vecteur de séries chronologiques stationnaires.

Pour éviter qu'il y ait multicollinéarité entre les séries NTC et JLo, nous avons défini deux modèles vectoriels ARMA: un modèle $(1,2)(0,1)_{12}$, qui met en relation le nombre de bénéficiaires de l'assurance-chômage et le niveau total de chômage, et un modèle $(2,6)(0,1)_{12}$, qui met en relation les bénéficiaires de l'assurance-chômage, les personnes ayant perdu leur emploi et celles ayant quitté leur emploi. Ces modèles ont été définis et estimés par la méthode du maximum de vraisemblance du programme de Scientific Computing Associates (Liu et Hudak 1983). Les deux modèles sont ajustés respectivement aux données originales transformées:

$$\begin{pmatrix} bac_t \\ ntc_t \end{pmatrix} = (1 - B)(1 - B^{12}) \log_{10} \begin{pmatrix} BAC_t \\ NTC_t \end{pmatrix} \tag{4.6}$$

et

$$\begin{pmatrix} bac_t \\ jlo_t \\ jle_t \end{pmatrix} = (1 - B)(1 - B^{12}) \log_{10} \begin{pmatrix} BAC_t \\ JLo_t \\ JLe_t \end{pmatrix} \tag{4.7}$$

Le tableau 3 donne la valeur des paramètres du modèle vectoriel (1,2)(0,1) ainsi que l'erreur type des estimations, qui figure entre parenthèses. Les estimations de paramètres et la matrice des variances-covariances des résidus, qui sont présentées au tableau 3 ne peuvent être comparées à celles des modèles unidimensionnels (tableau 1) parce que, dans le premier cas, le modèle a été ajusté aux données transformées normalisées alors que, dans le deuxième cas, le modèle a été ajusté aux données non normalisées. Une analyse du tableau 4, où sont décrites les corrélations avec décalage des résidus, nous permet de croire que le modèle est approprié. On désigne par un signe positif (+) une estimation qui est supérieure au double de son erreur type et par un signe négatif (−) une estimation qui est inférieure au double de son erreur type et on désigne par un point une valeur non significative fondée sur le critère précédent.

Ainsi, le modèle vectoriel ARMA pour les séries BAC et NTC devient

$$bac_t = 0.669 bac_{t-1} + \hat{a}_{1t} - 0.794\, \hat{a}_{1(t-12)} \tag{4.8}$$

$$ntc_t = 0.475 bac_{t-1} - 0.347 ntc_{t-1} + \hat{a}_{2(t)} - 0.308 \hat{a}_{2(t-2)} - 0.705 \hat{a}_{2(t-12)} + 0.217 \hat{a}_{2(t-14)} \tag{4.9}$$

Tableau 3

Estimation de paramètres pour les séries BAC et NTC transformées

$\hat{\Phi}_1$		$\hat{\theta}_2$		$\hat{\Theta}_{12}$	
0.669	-	-	-	0.794	-
(0.089)				(0.090)	
0.475	−0.347	-	0.308	-	0.705
(0.098)	(0.115)		(0.116)		(0.086)

Σ		$\hat{\theta}_1$
0.429249	-	0
0.131532	0.544389	

Tableau 4
Matrices des corrélation avec décalage des résidus +, −, et ·

DÉCALAGES 1 À 6					
..	..	..	..	..	..
..	..	..	..	..	..
DÉCALAGES 7 À 12					
..	..	..	..	..	..
..	..	..	..	..	..
DÉCALAGES 13 À 18					
..	..	..	..	..	..
..	..	..	..	..	..
DÉCALAGES 19 À 24					
..	..	..	..	..	..
..	..	..	..	..	..

Les équations (4.8) et (4.9) indiquent que la série des bénéficiaires de l'assurance-chômage devance d'un mois la série du niveau total de chômage. En fait, lorsque nous analysons la relation entre les séries BAC et NTC, nous devons retenir qu'une augmentation du nombre de personnes ayant perdu leur emploi et, partant, du nombre de bénéficiaires de l'assurance-chômage peut amener d'autres membres de la famille à chercher un emploi afin de combler la perte de revenu. Ces chercheurs d'emploi, venant ou revenant au marché du travail, ne sont pas admissibles aux prestations d'assurance-chômage mais viennent grossir le nombre des chômeurs. Nous pouvons en outre concevoir une augmentation du niveau total de chômage sans un accroissement des transferts normaux de main-d'oeuvre entre les divers marchés puisqu'en période de récession, par exemple, on compte plus de bénéficiaires de l'assurance-chômage mais beaucoup moins d'emplois disponibles, ce qui accroît le niveau de chômage.

Les résultats de ce modèle concordent aussi avec les résultats préliminaires de l'analyse des corrélations avec décalage par paire, qui a été faite à la section précédente (voir diagramme 1). Toutefois, ce modèle-ci décrit mieux la dynamique de ces deux phénomènes. Nous constatons que le nombre de bénéficiaires de l'assurance-chômage est positivement relié au nombre de bénéficiaires du mois précédent. Le niveau total de chômage est, quant à lui, positivement relié au nombre de bénéficiaires du mois antérieur et négativement relié à son niveau du mois précédent. L'effet du mouvement saisonnier se reflète dans la portion de chaque équation réservée aux moyennes mobiles par un coefficient de changement aléatoire élevé au décalage 12.

Le tableau 5 contient les résultats de l'application du modèle vectoriel ARMA (2,6)(0,1) aux séries BAC, JLo et JLe transformées, telles qu'elles sont définies par le système d'équations (4.7).

Comme les éléments estimés des matrices de corrélations avec décalage des résidus du tableau 6 ne semblent décrire aucune tendance particulière, le modèle ci-dessus est jugé approprié.

La version finale du modèle vectoriel ARMA (2,6)(0,1) pour les trois variables est définie comme suit :

$$bac_t = 0.617bac_{t-1} + 0.268jlo_{t-2} + \hat{a}_{1(t)} + 0.221\hat{a}_{3(t-6)} - 0.831\hat{a}_{1(t-12)} - 0.176\hat{a}_{3(t-18)} \quad (4.10)$$

$$jlo_t = 0.577bac_{t-1} - 0.285jlo_{t-1} + \hat{a}_{2(t)} + 0.386\hat{a}_{3(t-6)} - 0.525\hat{a}_{2(t-12)} - 0.308\hat{a}_{3(t-18)} \quad (4.11)$$

$$jle_t = 0.303bac_{t-2} - 0.411jle_{t-1} - 0.403jle_{t-2} + \hat{a}_{3(t)} - 0.797\hat{a}_{3(t-12)}. \quad (4.12)$$

Tableau 5

Estimations de paramètres pour les séries BAC, JLo et JLe transformées

$\hat{\underline{\Phi}}_1$

0.617 (0.086)	–	–
0.577 (0.099)	−0.285 (0.096)	–
–	–	−0.411 (0.088)

$\hat{\underline{\Phi}}_2$

–	0.268 (0.080)	–
–	–	–
0.303 (0.083)	–	−0.403 (0.084)

$\hat{\underline{\theta}}_6$

–	–	−0.221 (0.087)
–	–	−0.386 (0.108)
–	–	–

$\hat{\underline{\Theta}}_{12}$

0.831 (0.094)	–	–
–	0.525 (0.096)	–
–	–	0.797 (0.077)

$\underline{\Sigma}$

0.339	–	–
0.117	0.483	–
0.014	0.153	0.428

$\hat{\underline{\theta}}_t, t = 1,2,\ldots,5$

$\underline{0}$

Tableau 6

Matrices des corrélations de décalage +, −, et ·

DÉCALAGES 1 À 6

… … … … … …
… … … … … …
… … … … … …

DÉCALAGES 7 À 12

·· + · … … … … …
… … … … … …
… … … … … …

DÉCALAGES 13 À 18

·· _ … … … _ ·· …
… … … … … …
… … … … … …

DÉCALAGES 19 À 24

… … … … … …
… … … … … …
… … … … … …

Les équations (4.10) et (4.11) indiquent l'existence d'une relation de rétroaction entre le nombre de personnes ayant perdu leur emploi et le nombre de bénéficiaires de l'assurance-chômage, cette relation étant comparable à celle qui a été mise à jour dans la section 3. La série JLo devance de deux mois la série BAC (équation 4.10) et le retard d'un mois qu'elle accuse par rapport à la série BAC (équation 4.11) influe sensiblement sur sa valeur courante. En outre, les valeurs des deux variables endogènes BAC et JLo sont influencées par leurs valeurs du mois précédent; positivement pour la série BAC et négativement pour la série JLo. Les coefficients de a_{t-6} et de a_{t-12} dénotent la relation que crée le mouvement saisonnier entre les deux séries. Le fait que la série JLe soit caractérisée par deux sommets accentués au cours d'une même année (un en hiver et un en été — voir figure 6) nous oblige à introduire un coefficient de moyenne mobile au décalage 6.

Ces observations ne viennent pas contredire la théorie économique. On a prétendu à bon droit que la causalité ne pouvait être uniquement déterminée par des résultats empiriques et qu'elle devait être aussi appuyée par la théorie économique (voir, par exemple Zellner 1979). Nous pouvons facilement admettre que, dans le contexte d'une récession économique, une augmentation du nombre de travailleurs licenciés se traduira par un nombre accru de bénéficiaires de l'assurance-chômage. En retour, un accroissement du nombre de ces bénéficiaires engendrera d'autres licenciements puisque, dans un tel contexte, la plupart des entreprises procèdent à des mises à pied temporaires pour pouvoir ensuite rappeler leurs employés lorsqu'une reprise s'amorce.

L'équation (4.12) soulève une question intéressante en indiquant que la série des bénéficiaires de l'assurance-chômage devance de deux mois la série des personnes ayant quitté leur emploi. Nous ne voyons pas à prime abord pourquoi il devrait en être ainsi. Nous pourrions trouver des explications plausibles de ce phénomène en analysant la dynamique à court terme des marchés du travail au Canada, mais une analyse approfondie de la question exigerait des données longitudinales. Nous pouvons toutefois avancer l'hypothèse, parmi tant d'autres, selon laquelle une augmentation du nombre de personnes ayant perdu leur emploi et, partant, du nombre de bénéficiares de l'assurance-chômage incite d'autres membres de la famille à chercher un emploi pour combler la perte de revenu. Ces chercheurs d'emploi sont les nouveaux venus et les personnes qui reviennent sur le marché du travail après une certaine absence. Ces deux catégories de personnes peuvent très difficilement se trouver un emploi en période de récession, lorsque le nombre de licenciements est à la hausse. Ces groupes sont surtout composés de jeunes et de femmes de plus de 25 ans, qui sont prêts dans un premier temps à accepter n'importe quel emploi pourvu que cet emploi procure un revenu additionnel à la famille. Ces personnes pourraient travailler le temps qu'il faut pour qu'elles deviennent admissibles aux prestations d'assurance-chômage. Une fois admissibles à ces prestations, elles quitteraient leur emploi afin d'en chercher un autre plus conforme à leurs goûts et à leurs aptitudes.

5. CONCLUSIONS

Le principal objet de cette étude a été de déterminer s'il existait des relations temporelles entre la série des bénéficiaires de l'assurance-chômage (BAC) et les séries du niveau total de chômage (NTC), du nombre de personnes ayant perdu leur emploi (JLo) et du nombre de personnes ayant quitté leur emploi (JLe), en construisant des modèles dynamiques de séries chronologiques multidimensionnelles.

Nous avons commencé par faire une analyse préliminaire en tentant de découvrir des relations temporelles par paire entre les séries NTC, BAC, JLo et JLe, selon le modèle défini par Granger (1969) et Pierce et Haugh (1977). Les résultats de notre analyse ont révélé l'existence de relations entre les quatre variables étudiées. Nous avons ensuite défini et estimé

deux modèles de séries chronologiques multidimensionnelles suivant les méthodes élaborées par Tiao et Box (1981) et Tiao et Tsay (1983). Les résultats des modèles vectoriels ARMA concordent avec les résultats préliminaires des modèles unidimensionnels ARMMI (corrélations avec décalage par paire des résidus).

Le premier modèle vectoriel ARMA montre que la série BAC devance d'un mois la série NTC. De plus, le modèle indique une relation directe entre les termes de la série BAC et une relation inverse entre ceux de la série NTC.

Le second modèle vectoriel ARMA indique que la série JLo devance de deux mois la série BAC et qu'il y a une rétroaction d'un mois de BAC sur JLo. On observe, en outre, une relation directe entre les termes de la série BAC mais une relation inverse entre ceux de la série JLo. Le second modèle ARMA indique aussi que la série BAC devance la série JLe de deux mois.

Ces résultats empiriques, fondés sur des données couvrant la période 1975-1982, ne sont pas contraires à la théorie économique. Ils sont, en outre, conformes aux résultats d'études antérieures réalisées au Canada et fondées sur des données portant sur une période antérieure à 1975. Les résultats de ces études confirmaient l'affirmation générale selon laquelle l'assouplissement du programme d'assurance-chômage avait eu pour effet d'inciter un plus grand nombre de personnes, surtout des jeunes et des femmes adultes, à quitter leur emploi et avait contribué à accroître le niveau de chômage. Il semble donc que les révisions apportées au programme d'assurance-chômage en 1977 n'aient rien changé à la situation.

Il aurait été très intéressant d'analyser les effets de la forte récession du début des années 1980 sur les séries étudiées mais, compte tenu de la longueur de ces séries, la suppression de cette période de récession aurait réduit les séries à un point tel qu'elles n'auraient convenu à aucun modèle statistique.

BIBLIOGRAPHIE

BOX, G.E.P., et JENKINS, G.M. (1970). *Time Series Analysis Forecasting and Control.* San Francisco: Holden Day.

GRANGER, C.W.J. (1969): Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 424-438.

GREEN, C., et COUSINEAU, J.M. (1976). *Chômage et programmes d'assurance-chômage.* Conseil économique du Canada, Ottawa.

GRUBEL, M.G., MAKI, D., et SAX, S. (1975). Real and insurance induced unemployment in Canada. *Revue canadienne d'économique*, VIII, 174-191.

HAUGH, L.D. (1976). Checking the independence of two covariance stationary time series: A univariate residual cross-correlation approach. *Journal of American Statistical Association*, 71, 378-385.

JUMP, G.V., et REA, S.A. (1975). *The Impact of the 1971 Unemployment Insurance Act on Work Incentives and the Aggregate Labour Market.* Institute for Policy Analysis, Université de Toronto.

LAZAR, F. (1978). The impact of the 1971 unemployment insurance revisions on unemployment rates: Another look. *Revue canadienne d'économique*, août, 559-570.

LIU, L.M., HUDAK G.B. (1983). *Univariate - Multivariate Time Series and General Statistical Analysis.* Scientific Computing Associates, Illinois.

LJUNG, G.M., et BOX, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-307.

PIERCE, D.A., et HAUGH, L.D. (1977). Causality in temporal systems. *Journal of Econometrics*, 5, 265-293.

SIEDULE, T., SKOULAS, N., et NEWTON, K. (1976). *La population active et les politiques économiques - une analyse économétrique*, Conseil économique du Canada, Ottawa.

STATISTIQUE CANADA (1976). *Méthodologie de l'enquête sur la population active du Canada*. N° de Catalogue 71-716.

TIAO, G.C., et BOX, G.E.P. (1981). Modeling multiple time series with applications. *Journal of American Statistical Association*, 76, 802-816.

TIAO, G.C., et TSAY, R.S. (1983). Multiple time series modeling and extended sample cross correlations. *Journal of Business and Economic Statistics*, 1, 43-56.

ZELLNER, A. (1979). Causality and econometrics. *Carnegie-Rochester Conference Series on Public Policy*, no. 10 (Karl Brunner et Allan H. Meltzer eds.), supp. du Journal of Monetary Economics, Amsterdam: North-Holland Publishing Company.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 181–191
Statistique Canada

Principes fondamentaux pour le développement des questionnaires

LARRY SWAIN[1]

RÉSUMÉ

Trente principes fondamentaux sont présentés à l'égard du développement des questionnaires comprenant le contenu, la formulation, la présentation et l'évaluation des questionnaires. L'importance du questionnaire comme composant intégral d'une enquête est soulignée comme aussi la considération de son rapport avec d'autres aspects du plan de l'enquête.

MOTS CLÉS: enquête; questionnaire; méthodologie.

1. INTRODUCTION

La plupart des enquêtes reposent sur un questionnaire qui doit être rempli par le répondant ou un représentant officiel de l'organisme qui mène l'enquête (sur place ou par téléphone). Comme le questionnaire est l'instrument qui traduit les objectifs d'une enquête en variables mesurables, la réalisation de ces objectifs exige un questionnaire efficace. En outre, le questionnaire peut aider à organiser, à normaliser et à contrôler la collecte des données de telle sorte que les informations nécessaires puissent être recueillies d'une manière satisfaisante. Pour bien élaborer un questionnaire, il faut appliquer certains principes de base et de la logique aux besoins particuliers de chaque enquête.

Bien que trente principes individuels de l'élaboration d'un questionnaire soient présentés ici, ils ne doivent pas être considérés comme indépendants les uns des autres ou du contexte dans lequel une enquête se déroule. Le questionnaire est une partie intégrante de toute enquête — on ne saurait trop insister là-dessus. Puisque le questionnaire *ne peut pas* être isolé des divers autres aspects d'une enquête, le lecteur doit, *pendant* l'élaboration du questionnaire, examiner les rapports entre cet instrument et les objectifs de l'enquête, la population visée, la collecte, le codage, la saisie, la vérification et l'imputation des données, le caractère confidentiel des informations fournies et les essais préliminaires.

Comme cet exposé n'a pas été conçu comme un traité complet sur l'élaboration des enquêtes ou des questionnaires, certains lecteurs, selon leur point de vue sur les différents aspects d'une enquête, remarqueront peut-être des lacunes dans les principes proposés ou préféreront exclure certains principes parce qu'ils semblent relever davantage d'une autre étape que la confection du questionnaire.

Les principes de base de l'élaboration d'un questionnaire énoncés portent sur le contenu, la formulation, la présentation et les essais préliminaires. Le questionnaire influe beaucoup sur la possibilité pour une enquête d'atteindre ses objectifs. Contrairement aux autres grandes étapes d'une enquête, par exemple la construction du plan de sondage ou les opérations de traitement des données, le questionnaire touche directement le répondant. Il est donc essentiel que le contenu, la formulation et la présentation du questionnaire permettent d'obtenir du répondant des renseignements fiables, valables et correspondant aux informations recherchées. L'auteur reconnait que bien que certains des principes semblent évidents quand ils

[1] Larry Swain, anciennement de la Division des méthodes de recensement et d'enquête-ménages, Statistique Canada; actuellement à la Division de la planification des ressources humaines, Commission de la Fonction publique, Ottawa, Canada K1A 0M7.

sont énoncés, souvent ils peuvent ne pas l'être en pratique. De même, certains des principes peuvent être mesurés, alors que d'autres ne peuvent pas l'être.

Dans ce qui suit, le terme *questionnaire* représente toujours les diverses formules utilisées pour recueillir des informations. Dans la documentation et en pratique, on distingue souvent les documents suivants:

a) questionnaires remplis par un répondant;
b) questionnaires remplis par un interviewer;
c) formules administratives remplies par un répondant ou un représentant officiel de l'organisme qui mène une enquête;
d) formules utilisées pour enregistrer des observations ou des mesures (remplies par un représentant officiel de l'organisme qui mène une enquête);
e) formules utilisées pour transcrire des informations figurant dans des dossiers administratifs (remplies par un représentant officiel de l'organisme qui mène une enquête).

Pour simplifier les choses, on emploie ici le terme *questionnaire* de manière à englober tous ces sens. En outre, on entend par *question* chaque question ou demande d'information, y compris le choix de réponses ou l'espace prévu pour une réponse.

Le terme *enquête* est utilisé dans un sens général pour toute activité de collecte de données comme les enquêtes par sondage, les recensements et le prélèvement de données administratives.

2. CONTENU

1. Toutes les questions d'un questionnaire doivent porter directement sur l'objectif et les fonctions de l'enquête.

Il est raisonnable d'exiger que la collecte des données soit organisée de manière à réduire au minimum le fardeau des répondants. Un moyen de satisfaire à cette condition est d'exclure les questions qui n'ont qu'un rapport vague avec les objectifs et les fonctions de l'enquête.

Les demandes de renseignements inutiles augmentent aussi la longueur d'un questionnaire et peuvent provoquer une attitude de méfiance chez le répondant, facteurs susceptibles d'accroître les taux de non-réponse (source possible d'erreurs systématiques), de baisser la qualité des données en fatiguant les interviewers ou les répondants ou en diminuant leur concentration et de coûter plus d'argent et de temps au client qui a commandé une enquête et aux répondants.

Du point de vue du concepteur d'un questionnaire, le fait même de préciser le lien entre chaque question et les objectifs et les fonctions de l'enquête l'aide à s'assurer que ces objectifs et fonctions sont bien définies et que le questionnaire y correspond effectivement.

2. Si un questionnaire demande des renseignements qui sont utiles pour l'enquête mais risquent de ne pas le sembler pour les répondants, il faut alors expliquer aux répondants la raison pour laquelle ces informations sont nécessaires.

Des variables de classification comme l'âge, le sexe, l'état matrimonial, la taille d'un organisme, l'effectif d'un employeur et des variables telles que le nom, l'adresse et le numéro de téléphone (renseignements utilisés pour des suivis ou des vérifications) sont des exemples possibles de données dont l'enquêteur doit songer à expliquer la nécessité (du moins à un niveau général) aux répondants.

3. Seules des questions auxquelles il est possible de répondre facilement et avec assez de fiabilité doivent être incluses.

Quand les répondants doivent chercher dans leur mémoire, les événements visés doivent être assez récents ou connus des répondants; si les informations voulues se trouvent dans des dossiers des répondants, l'effort (mesuré en temps et en argent) nécessaire pour obtenir ces renseignements ne doit pas dépasser les avantages qui découleront de cette recherche.

À cause du risque d'ambiguïtés dans les définitions, d'accroissement du fardeau des répondants et d'erreurs de traitement, il peut être préférable de ne pas demander aux répondants de traiter les données qui conduisent à un certain résultat. On peut parfois faciliter la collecte des données et éviter des erreurs en demandant aux répondants de fournir des informations dont ils disposent et laisser l'organisme qui mène l'enquête faire le traitement de ces renseignements par la suite.

4. Il ne faut pas poser de questions auxquelles il n'est pas évident que les répondants peuvent répondre.

Les questions ne doivent pas présupposer que le répondant possède des connaissances sur un sujet précis ou exerce une activité particulière. Des questions filtres permettent d'omettre les questions qui ne s'appliquent pas à un répondant à cause de ses caractéristiques, de sa situation ou de ses opinions.

Si les répondants trouvent que beaucoup de questions ne s'adressent pas à eux, ils risquent de penser que le questionnaire leur a été donné par erreur, ce qui peut aggraver la non-réponse ou nuire aux relations avec les répondants.

Une question filtre indique clairement si un répondant doit répondre à une question ou à une série de questions. Cette propriété est utile pour le traitement et l'analyse des données d'une enquête. Si aucune réponse n'est fournie à une question, il peut être difficile de distinguer entre les cas de non-réponse (refus ou oubli accidentel) et les cas où une question n'est pas applicable (pour une réponse numérique, on risque de conclure que la question ne s'appliquait pas au répondant, alors que la réponse est en fait zéro). Une question filtre aide à résoudre ce problème en montrant quels répondants devaient répondre à une question.

On doit toutefois éviter de rédiger des directives compliquées sur les questions à omettre, en particulier pour les questionnaires que les répondants remplissent eux-mêmes. Il faut aussi réduire au minimum le nombre de questions filtres.

Pour les réponses numériques, on peut éviter de poser une question filtre en incluant la catégorie *aucun* ou *aucune*.

3. FORMULATION

1. La formulation d'une question doit convenir au répondant.

Si un répondant ne comprend pas une question, il donnera probablement une réponse inexacte ou ne répondra pas. Il faut que les mots, le style et la structure des phrases soient faciles à comprendre et conviennent aux personnes qui doivent fournir des renseignements.

Il faut éviter les abréviations à moins que les répondants ne les comprennent.

2. Lorsque la demande est suffisante, un questionnaire doit être traduit en d'autres langues.

Il est important de vérifier si la traduction reproduit bien le sens du texte de départ.

3. Le concepteur d'un questionnaire doit choisir les types de question qui conviennent le plus aux renseignements demandés et permettent de réduire au minimum l'erreur de réponse et le fardeau des répondants.

Les principaux types de question sont les questions ouvertes à réponse libre, les questions fermées à réponses fixes et les questions à réponse structurée. Les questions fermées comprennent une liste des réponses possibles. Les questions à réponse structurée semblent ouvertes parce qu'aucun choix de réponses n'est fourni, mais elles sont implicitement fermées parce que le répondant est normalement obligé de se limiter à un chiffre, un jour de la semaine, une province, etc.

En général, le fardeau du répondant et/ou de l'interviewer est moins grand quand les questions sont fermées parce que les répondants n'ont pas à chercher leurs propres mots pour formuler des réponses et que ces réponses n'ont pas à être transcrites textuellement.

4. Lorsqu'un répondant doit choisir entre deux réponses bien définies ou plus, une question fermée ou à réponse structurée doit être utilisée.

Si les réponses à une question sont trop nombreuses pour être toutes énumérées, il est recommandé d'utiliser la catégorie *autre* pour englober des réponses peu fréquentes, de poser une question à réponse structurée ou à réponse libre ou de regrouper certaines réponses pour réduire le nombre de catégories. On peut même utiliser une question à réponse structurée si les réponses et leurs codes numériques figurent dans un livret d'instructions accompagnant le questionnaire. Dans les enquêtes sur les entreprises, l'agriculture et les institutions, des questions à réponse structurée sont souvent employées quand les réponses sont numériques. Le choix et le nombre de réponses à une question fermée dépend de la complexité de la notion mesurée, de la manière dont les données seront utilisées et des informations préalables dont le concepteur d'un questionnaire dispose.

5. Si le choix de réponses à une question n'est pas bien défini, une question ouverte doit être utilisée.

Les questions ouvertes sont souvent utilisées dans les recherches préliminaires ou exploratoires entreprises en vue de formuler des hypothèses précises et de structurer les questions pour les versions ultérieures d'un questionnaire. Les questions ouvertes offrent également un moyen d'obtenir des renseignements supplémentaires ou des précisions, de vérifier les réponses à d'autres questions, d'interpréter les données, de modifier le rythme du questionnaire ou d'aborder un nouveau sujet.

6. Si la facilité, la rapidité et le coût du traitement des données pour la saisie sont des facteurs importants, des questions fermées doivent être utilisées.

Les réponses à des questions ouvertes doivent être codées, opération qui peut être coûteuse, prendre beaucoup de temps et introduire des erreurs d'interprétation et de codage.

Un autre problème est que les questions ouvertes ne fournissent aucun schéma de référence et que les répondants choisissent alors différents schémas pour leurs réponses. Comme les répondants peuvent avoir différents schémas de référence et fournir différentes quantités d'information, il est difficile d'interpréter, de coder et d'analyser les réponses. En revanche, une question fermée contient un schéma de référence précis qui permet d'éviter le problème susmentionné mais peut susciter une réponse artificiellement. Ce problème est particulièrement grave quand le répondant connaît peu un sujet donné, l'ignore ou n'a pas d'opinion marquée. Le concepteur du questionnaire doit donc considérer les schémas de référence possibles des répondants avant de décider de la forme d'une question.

Le choix du type de question et de la formulation impose des restrictions sur la manière dont le renseignement demandé peut être utilisé ou analysé et les hypothèses précises qui peuvent être vérifiées. Il s'ensuit que la définition des objectifs, des utilisations, des hypothèses à vérifier et des analyses voulues doit se faire avant que ne soit fixée la forme définitive d'une question. Le fait que ces éléments soient établis à l'avance n'exclut pas la possibilité que les données recueillies puissent servir à d'autres analyses et besoins à l'intérieur des limites imposées par la nature des questions.

Outre les facteurs mentionnés plus haut, l'expérience du concepteur d'un questionnaire influera sur le choix des formes de questions qui conviennent à différentes situations.

7. Il ne doit pas y avoir de chevauchement dans les réponses aux questions fermées et ces réponses doivent être exhaustives (c'est-à-dire qu'elles sont mutuellement exclusives et complètes).

Les réponses à une question particulière doivent être distinctes et inclure toutes les possibilités.

Le fait que les réponses à une question fermée doivent être exclusives ne signifie nullement qu'aucune question ne peut avoir plus d'une réponse. S'il y a plus d'une réponse, une note telle que cocher *autant de réponses que nécessaire* doit être incluse avec la question.

Si seulement une réponse doit être fournie à une question, une note telle que *cocher une case seulement* doit être incluse dans la question (sauf dans les cas les plus évidents, par exemple quand la réponse est *oui* ou *non*). Si plus d'une réponse peut être bonne mais que le concepteur veut que le répondant choisisse seulement une catégorie afin de limiter les réponses, une note telle que *cocher la réponse la plus appropriée* doit figurer dans la question.

8. Les unités de mesure de chaque réponse doivent être précisées.

Il faut soit préciser l'unité de mesure de la réponse (kilogrammes, tonnes, pourcentage, heures par semaine, etc.) à l'intérieur de la question, soit demander au répondant de l'indiquer lui-même. Autrement, il peut y avoir de la confusion au sujet de l'unité dans laquelle une réponse est exprimée.

9. Des notions et des définitions normalisées doivent être utilisées.

Pour faciliter la comparaison des données d'une enquête à celles d'autres sources d'information (publications, enquêtes) et rendre les données aussi utiles que possible (y compris pour les analyses secondaires), il est important d'employer des définitions normalisées (généralement comprises et utilisées) quand elles existent et qu'elles sont claires, appropriées et à jour. Statistique Canada publie des classifications types pour les professions, les activités économiques, les marchandises et les unités géographiques, de même qu'un répertoire de concepts sociaux. Par ailleurs, les notions et les catégories établies pour le recensement sont souvent utilisées comme normes.

10. La formulation des questions doit être précise, complète, cohérente, brève, simple et claire.

Les notions et les termes que les répondants ne connaissent pas ou peuvent mal comprendre doivent être expliqués, définis ou évités. Pour que les résultats soient interprétés de manière cohérente, le schéma de référence approprié (par exemple la période de référence, l'endroit visé ou le type de dépense étudié) doit être fourni. Si la cohérence des données risque d'être imparfaite (par exemple si différentes questions portent sur différentes périodes), tous les changements de schéma de référence doivent être soulignés dans le questionnaire.

Dans les contextes où plusieurs mots peuvent être utilisés indifféremment, il faut choisir un mot et l'employer partout dans le questionnaire. Si un synonyme d'un mot qui figure ailleurs le remplace dans un autre passage, les répondants risquent d'y accorder un sens différent.

11. Les questions doubles doivent être évitées.

Une question double demande au répondant de fournir une seule réponse à ce qui est en réalité deux questions. La réponse ne permet pas de savoir à laquelle des deux questions le répondant pensait ou si ce dernier pensait aux deux à la fois. Les deux renseignements recherchés doivent être demandés séparément, sauf dans les cas particuliers où il faut nécessairement les demander en même temps pour transmettre le message voulu. On doit alors indiquer clairement aux répondants que les deux aspects de la question doivent être envisagés ensemble.

12. Les questions tendancieuses doivent être évitées.

Une question tendancieuse est formulée ou présentée de manière à porter le répondant à choisir une réponse ou série de réponses en particulier.

Certaines questions peuvent être considérées comme tendencieuses si elles renferment des choix que les répondants peuvent trouver socialement inacceptables si on ne les assure pas que leur réponse ne suscitera aucun jugement de valeur.

Dans les enquêtes sur les attitudes, deux principes de base ont été établis pour réduire (mais pas nécessairement éliminer) le biais de réponse. Le choix des réponses doit être équilibré, de sorte qu'il y a autant de réponses positives que de réponses négatives pour éviter de faire pencher le répondant pour un type de réponse. Deuxièmement, s'il y a une série de questions qui ont le même choix de réponses, cette série doit soit contenir un mélange de phrases positives et négatives, soit être divisée, soit être présentée dans un ordre varié afin de réduire la possibilité que les répondants puissent répondre de la même manière pour toute la série (même si cela peut n'être pas approprié) sans penser très sérieusement à chaque réponse.

4. PRÉSENTATION

1. Chaque questionnaire ou ensemble de formules d'enquête doit contenir un texte qui explique et présente l'enquête.
2. Le texte d'introduction doit préciser le titre de l'enquête, le nom des organismes qui l'ont commandée et les objectifs visés.
3. Il faut songer à inclure un avis assurant les répondants quelles données fournies sont confidentielles.
4. Le questionnaire doit indiquer le nom (s'il y a lieu) et le numéro de téléphone ou l'adresse postale d'une personne avec qui les répondants peuvent entrer en contact dans l'organisme qui a commandé l'enquête s'ils ont besoin de renseignements supplémentaires.

En général, le texte de présentation peut être une lettre ou une brochure envoyée au répondant; ce texte peut également être rédigé à l'avance et lu par un interviewer; ou il peut être imprimé sur le questionnaire. Ce texte doit fournir des renseignements généraux essentiels pour identifier l'organisme qui mène l'enquête, justifier l'enquête et informer les répondants de leurs droits (s'il y a lieu).

5. Les répondants doivent être identifiés d'une manière convenable sur le questionnaire.

Pour l'estimation, le contrôle sur le terrain, l'enchaînement avec d'autres enregistrements ou les suivis auprès des non-répondants, les répondants doivent être identifiés (par un code numérique ou autrement) sur les questionnaires.

6. Les questions et les pages du questionnaire doivent être numérotées.

Pour faciliter la lecture du questionnaire pour les interviewers et les répondants et simplifier les opérations de codage et les instructions, les questions et les pages doivent être numérotées consécutivement (à l'aide de lettres ou de chiffres) dans l'ensemble du questionnaire. Si des questions sont imprimées sur les deux côtés d'une feuille, une indication (comme *suite au verso*) doit apparaître au bas du recto pour assurer qu'une réponse est fournie aux questions qui figurent sur le deuxième côté.

7. Les caractères d'impression doivent être faciles à lire pour le répondant moyen.

On doit penser à la personne qui répondra au questionnaire quand on décide de la grandeur des caractères d'impression (par exemple, de petits caractères pourraient poser des problèmes pour les personnes dont la vue n'est pas bonne) et qu'on choisit les couleurs et le degré de contraste entre le papier et les caractères. Il est normalement préférable d'utiliser des caractères différents (de grandeur ou de forme différente) pour les questions et les instructions, de sorte qu'on puisse facilement distinguer ces parties.

8. Des instructions concernant la manière de répondre aux questions doivent être imprimées sur le questionnaire ou l'accompagner.

Pour aider les répondants, les interviewers ou d'autres représentants à bien remplir le questionnaire, des instructions brèves mais claires doivent être imprimées sur le questionnaire ou un autre document (par exemple un manuel d'interviewer ou un livret d'instructions). Toutefois, les questions doivent être aussi claires que possible pour éviter d'avoir à donner des instructions compliquées.

Dans le cas des questionnaires lus par un lecteur optique de caractères, des instructions claires doivent être fournies pour assurer qu'ils sont bien remplis.

Les instructions indiquant aux répondants ou aux interviewers les questions à omettre après les questions filtres doivent être assez simples et faciles à suivre. Il peut être approprié d'utiliser des flèches et des indications. On doit éviter des séries complexes de directives sur les questions à omettre, surtout dans les questionnaires remplis par les répondants eux-mêmes.

9. Les instructions pour le renvoi du questionnaire doivent figurer sur le questionnaire.

Si un questionnaire doit être retourné par la poste, le nom et l'adresse de la personne (ou de l'organisme) à laquelle il faut l'envoyer doivent être indiqués sur le questionnaire. Les lettres de présentation et les enveloppes de retour peuvent facilement se perdre ou être séparées du questionnaire.

La date limite avant laquelle les répondants doivent retourner les questionnaires remplis doit également être précisée.

Si un questionnaire doit être recueilli par un représentant sur le terrain, on doit prévoir un espace sur le questionnaire pour le nom, le numéro de téléphone ou l'adresse postale où il est possible de joindre le représentant et indiquer la date et l'heure approximative à laquelle ce dernier viendra.

10. Les chiffres et les codes utilisés pour la saisie des données doivent figurer sur le questionnaire (quand la saisie doit se faire directement à partir du questionnaire).

Dans certains cas, les données peuvent être saisies plus rapidement et avec moins d'erreurs directement à partir du questionnaire. Il faut alors que les chiffres et les codes puissent être lus facilement par les personnes chargées de la saisie des données, mais ils ne doivent pas nuire à la concentration du répondant, de l'interviewer ou d'un autre représentant qui remplit le questionnaire.

Quand les données doivent être codées avant la saisie, les cases pour le codage peuvent être imprimées sur le questionnaire ou sur une feuille séparée. Si les cases de codage figurent sur le questionnaire, elles doivent être clairement distinguées des cases pour les réponses, peut-être par une indication comme *réservé au bureau* ou par une couleur appropriée.

Le codage et la saisie des données sont souvent considérés comme des étapes liées à la structure du questionnaire. Il est essentiel d'y penser pendant l'élaboration du questionnaire pour assurer l'efficacité de leur déroulement.

11. La disposition des espaces pour les réponses doit être uniforme dans tout le questionnaire et ces espaces doivent être assez grands pour que le questionnaire soit lisible et offre assez de place pour les réponses.

L'uniformité des espaces pour les réponses facilite le travail du répondant, de l'interviewer ou du représentant qui remplit le questionnaire et aide à réduire les erreurs dues à l'omission accidentielle d'une question, au choix d'une réponse inexacte ou à la transposition de réponses.

Il peut être utile d'employer des formes d'espace différentes pour les questions fermées et celles dont la réponse est un chiffre. Des cercles sont parfois utilisés pour les questions fermées, des cases carrées pour les réponses numériques.

L'espacement doit être étendu pour plusieurs raisons: pour rendre le questionnaire facile à remplir, lui donner une bonne apparence, le rendre plus lisible et fournir assez de place au répondant, à l'interviewer ou au représentant pour inscrire la réponse à chaque question.

12. Les questions doivent se suivre dans un ordre logique pour en faciliter la lecture et établir le schéma de référence nécessaire.

L'ordre des questions doit sembler logique au répondant (dont la logique peut être différente de celle du concepteur du questionnaire) et les questions qui sont liées entre elles doivent se trouver ensemble. Une méthode qui est parfois recommandée consiste à commencer par les questions les plus générales avant de poser les questions les plus détaillées. L'ordre des questions doit autant que possible correspondre à l'ordre dans lequel les répondants fourniront des renseignements. Le concepteur d'un questionnaire doit se rappeler qu'une question peut susciter une réponse visant également une autre question qui (si l'ordre des questions est logique) n'est pas très loin.

La transition entre les différentes sections du questionnaire ne doit pas être brusque. Un en-tête ou une explication doit figurer au début de chaque section. Si un questionnaire est utilisé pour transcrire des renseignements figurant sur d'autres documents, un ordre logique serait celui des documents dans lesquels les informations sont puisées.

Dans les enquêtes sur les attitudes, le concepteur d'un questionnaire doit éviter de conditionner au départ les répondants en fonction d'un schéma de référence qui peut fausser les réponses aux questions posées par la suite. Par exemple, les questions mesurant le degré de connaissance d'un sujet doivent précéder toutes les autres questions qui abordent ce sujet. Les questions délicates doivent être regroupées avec d'autres questions connexes pour justifier leur inclusion autant que possible et atténuer quelque peu l'embarras que le répondant peut éprouver.

13. La version finale du questionnaire ne doit contenir aucune erreur typographique ou grammaticale.

La présence d'erreurs dans le questionnaire peut avoir un effet nuisible sur la qualité des données à cause du risque que le questionnaire ne soit pas pris au sérieux ou soit mal compris par les personnes qui le remplissent. Par ailleurs, les erreurs peuvent se répercuter négativement sur la manière dont le public perçoit l'organisme qui mène l'enquête.

5. ESSAIS

1. Les questionnaires qui sont utilisés pour la première fois ou ont été modifiés en profondeur doivent être mis à l'essai avant d'être employés pour la collecte de données.

Même si tous les principes décrits dans les élaborations précédentes ont été respectés et que le concepteur du questionnaire a été très consciencieux, il n'existe aucune garantie que le questionnaire proposé réponde entièrement aux objectifs de l'enquête. Il se produit presque toujours des problèmes imprévus lors de l'utilisation d'un questionnaire. Il est donc essentiel de faire un essai préliminaire du questionnaire pour toute enquête nouvelle ou très modifiée, afin de déterminer si le questionnaire proposé peut répondre aux objectifs de la recherche.

Le concepteur du questionnaire peut, par exemple, étudier: la formulation, l'agencement et la disposition des questions pour voir si les répondants et les interviewers comprennent les questions et leur enchaînement; la nécessité d'inclure des questions particulières; le choix des genres de questions; l'emploi de questions spéciales comme celles où l'enquêté doit classer un choix de réponses selon le rang ou appliquer une échelle d'évaluation; la structure et la définition des catégories de réponses; la fréquence de la catégorie "autre" dans les réponses; la facilité de remplir le questionnaire; le temps requis pour remplir les diverses sections du questionnaire; la traduction du questionnaire; la possibilité d'un biais dans les questions; la nature des différences ethniques, régionales ou linguistiques; le fardeau que le questionnaire impose au répondant; l'efficacité avec laquelle le questionnaire mesure les notions visées; les lettres ou autres moyens utilisés pour présenter l'enquête et la qualité de la méthode de collecte des données.

Il faut effectuer un essai préliminaire auprès d'au moins un petit échantillon de répondants (généralement une vingtaine ou une trentaine) de la population cible. Il est préférable de choisir ces répondants dans les divers sous-groupes de la population susceptibles de présenter des différences ou des problèmes. Les sous-populations pour un essai peuvent être définies en fonction de variables comme la région géographique, le niveau d'instruction, l'âge, le sexe, la langue, la taille de l'entreprise et la catégorie d'activité économique. Selon les buts particuliers de l'essai préliminaire, la sélection des répondants peut reposer sur un échantillon aléatoire ou non aléatoire mais, dans la plupart des cas, la méthode de sélection est non aléatoire. Il est également possible d'organiser un groupe de travail pour discuter du questionnaire dans le cadre de l'essai préliminaire.

La méthode de collecte utilisée pour l'essai préliminaire doit être identique à celle prévue pour l'enquête principale. Mais une interview sur place est recommandée pour au moins une partie des enquêtés de l'essai préliminaire afin que l'interviewer puisse noter les réactions, verbales ou autres, des répondants ainsi que leurs suggestions et impressions. À la fin de chaque interview de l'essai préliminaire, l'interviewer peut discuter des difficultés éprouvées par le répondant, de l'interprétation des questions et des catégories de réponse et ainsi de suite. Le concepteur du questionnaire peut ensuite organiser une réunion avec les interviewers qui ont participé à l'essai préliminaire pour discuter de ces difficultés et faire le bilan de l'expérience. Dans certains cas, il est préférable de faire appel à des interviewers qualifiés et expérimentés afin de tirer profit au maximum de l'essai préliminaire.

L'essai préliminaire est une étape souvent négligée. Il révèle presque toujours des améliorations qui peuvent être apportées ou permet du moins au concepteur du questionnaire de s'assurer que l'utilisation du questionnaire dans l'enquête principale, qui coûte beaucoup plus qu'un essai préliminaire, sera probablement assez efficace. Bien entendu, rien ne garantit jamais que tous les problèmes seront résolus, mais la plupart des grandes difficultés devraient l'être. Un essai préliminaire n'est pas nécessairement coûteux, ne prend pas forcément beaucoup de temps et est recommandé pour les questionnaires nouveaux ou modifiés.

REMERCIEMENTS

L'auteur remercie l'arbitre et le conseil de rédaction pour leurs observations utiles, ainsi que les nombreuses personnes qui ont examiné une version antérieure de ce texte qui faisait partie d'une ébauche partielle d'un document intitulé "Survey Design Standards and Guidelines".

BIBLIOGRAPHIE

ANDERSON, J.F., et BERDIE, D.R. (1974). *Questionnaires: Design and Use.* Metuchen, N.J.: The Scarecrow Press, Inc.

BERTHIER, N., et F. (1971). *Le sondage d'opinion.* Paris: Bordas.

BON, F. (1974). *Les sondages - peuvent-ils se tromper?* France: Calmann-Lévy.

CARSON, E. (1974). Questionnaire Design, Some Principles and Related Topics. Document non publié, Statistique Canada.

CORBIN, R., SWAIN, L., et WILHELM, E. (1977). Exposé pour un atelier sur la conception des questionnaires. Document non publié, Statistique Canada.

CORBIN, R., SWAIN, L., et WILHELM, E. (1977). Outline for a Workshop on Basic Questionnaire Design. Document non publié, Statistique Canada.

GHIGLIONE, R., et MATALON, B. (1978). *Les enquêtes sociologiques: théories et pratique.* Paris: Colin.

JAVEAU, C. (1974). *L'enquête par questionnaire. Manuel à l'usage du praticien,* troisième édition. Bruxelles: Institut de sociologie de l'Université libre de Bruxelles.

MOSER, C.A., et KALTON, G.J. (1972). *Survey Methods in Social Investigation,* deuxième édition. New York: Basic Books.

OPPENHEIM, A.N. (1966). *Questionnaire Design and Attitude Measurement.* Londres: Heinemann.

PAYNE, S.L. (1951). *The Art of Asking Questions.* Princeton: Princeton University Press.

STATISTIQUE CANADA (1979). Basic Questionnaire Design, deuxième édition. Document non publié, Statistique Canada.

STATISTIQUE CANADA (1981). Conception des questionnaires, Manuel d'atelier, troisième édition. Document non publié, Statistique Canada.

WARWICK, D.P., et LININGER, C.A. (1975). *The Sample Survey: Theory and Practice.* New York: McGraw-Hill.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 193–202
Statistique Canada

Vue d'ensemble des avantages et inconvénients des fichiers de données administratives choisis[1]

RAVI B.P. VERMA et PIERRE PARENT[2]

RÉSUMÉ

Dans ce texte, on évalue les possibilités de tirer des données de migration de douze fichiers de données administratives, dans l'éventualité où l'universalité des données d'allocations familiales, qui sont utilisées présentement,serait remise en question par une loi fédérale.

Il est montré qu'aucun fichier ne présente des avantages et des inconvénients exactement comparables à ceux de l'allocation familiale. Il est fortement recommandé de poursuivre le développement des fichiers d'assurance-santé, et, dans une moindre mesure du fichier de la sécurité de la vieillesse.

MOTS CLÉS: Fichiers administratifs; migration; évaluation qualitative.

1. INTRODUCTION

Au Canada, les fichiers des allocations familiales et de l'impôt sur le revenu sont utilisés de diverses façons pour produire des estimations de la migration et de la population pour les différentes régions géographiques (voir les publications n^{os} 91-001, 91-210, 91-211 et 91-212 au catalogue). Les données provenant des fichiers des allocations familiales sont publiées deux à trois mois après la date de référence, tandis que les données tirées des fichiers de l'impôt sur le revenu sont, elles, publiées douze à quinze mois après la date de référence. Cependant, les données de l'impôt sur le revenu produisent des estimations des flux migratoires par division de recensement et également selon l'âge et le sexe.

En termes de précision des estimations de la population, les deux types de fichiers donnent de bons résultats et sont comparables (voir Norris et Standish 1983; Norris 1983; Verma et. al. 1984; Verma et Basavarajappa 1985). Une des caractéristiques particulières des fichiers des allocations familiales et des fichiers de l'impôt sur le revenu réside dans le fait qu'ils ont un caractère national. Autre caractéristique, les dossiers contiennent des adresses où figure le code postal, de sorte qu'il est possible d'obtenir les renseignements sur les migrations locales. Toutefois, depuis quelques années, il semble y avoir possibilité que les allocations familiales cessent d'être universelles par suite de mesures législatives gouvernementales. Par exemple, les secteurs couverts pourraient être limités aux secteurs de la population des tranches de revenu inférieures et intermédiaires. Si ce fichier cessait d'être universel, son utilité comme source de données sur les migrations pourrait être sérieusement réduite. Par conséquent, notre capacité d'établir des estimations de la population pourrait être affaiblie, ce qui pourrait, par voie de conséquence, influer sur d'autres programmes comme le programme ayant trait au partage des revenus, qui suppose la répartition annuelle de 20 milliards de dollars entre les provinces.

Pour cette raison, il faut explorer d'autres sources de données. Nous tentons ici d'évaluer les avantages et les inconvénients de certains des fichiers de données administratives choisis pour l'estimation des migrations et de la population dans les provinces et territoires, les divisions de recensement, les régions métropolitaines de recensement ainsi que dans d'autres régions du Canada.

---

[1] Version abrégée du document présenté à la réunion du Comité fédéral-provincial sur la démographie tenues les 28 et 29 novembre 1985 à Ottawa.

[2] Ravi B.P. Verma et Pierre Parent, Division de la démographie, Direction de la statistique démographique et du recensement, Statistique Canada, 4^{ième} étage, Immeuble Jean Talon,Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6.

Les douze fichiers administratifs sont évalués du point de vue qualitatif comme source de données pouvant remplacer les fichiers des allocations familiales. Ces fichiers ont été divisés en trois groupes principaux de la façon suivante en fonction de leurs avantages et de leurs inconvénients.

Fichiers présentant le plus de potentiel pour l'estimation des flux migratoires

i) Fichiers de l'assurance-maladie
ii) Fichier de la sécurité de la vieillesse

Fichiers présentant le plus de potentiel comme indicateurs symptomatiques des variations de la population et de la migration nette

iii) Abonnés des services d'hydro-électricité
iv) Abonnés du téléphone
v) Inscriptions scolaires

Autres fichiers, présentant un potentiel limité ou incertain pour l'estimation des flux migratoires ou des migrations nettes

vi) Permis de conduire
vii) Permis de bâtir
viii) Bénéficiaires de l'assurance-chômage
ix) Enquête sur la population active
x) Listes électorales
xi) Ventes au détail
xii) Statistiques du camionnage

1.1 Critères d'évaluation des fichiers de données administratives

L'utilité des diverses sources de données administratives pour l'estimation des migrations interprovinciales et intraprovinciales est évaluée en fonction de dix critères: univers, couverture de l'univers, méthode de détermination de l'information sur les migrations, types de migration, caractéristiques des dossiers, date ou période de référence (et possibilité d'avoir des données mensuelles), délai, possibilité d'obtenir des séries chronologiques, cohérence et informatisation (Almond 1982).

La nouvelle source de données à utiliser présenterait un grand potentiel si elle contenait les caractéristiques des fichiers des allocations familiales, décrites au tableau 1. Les critères les plus importants sont les suivants: couverture, actualité des données, cohérence, possibilité d'obtenir des données mensuelles ou trimestrielles, niveau de désagrégation utilisant le code postal ou d'autres systèmes de géocodage. Le fichier ou l'ensemble de fichiers qui pourrait satisfaire ces critères aurait probablement les qualités voulues pour remplacer les fichiers des allocations familiales.

2. FICHIERS PRÉSENTANT UN GRAND POTENTIEL POUR L'ESTIMATION DES FLUX MIGRATOIRES

Les fichiers de l'assurance-maladie et de la sécurité de la vieillesse sont ceux qui ont le plus de potentiel pour l'estimation des flux migratoires entre les provinces, territoires et divisions de recensement. Les avantages et les inconvénients de chacun de ces fichiers sont décrits ci-après.

2.1 Fichier de l'assurance-maladie

L'assurance-maladie est du ressort des provinces. Chaque province tient donc à jour un dossier des personnes admissibles au programme. Tous les résidents d'une province (y compris

Tableau 1

Description des fichiers de données administratives actuellement utilisés pour générer les données sur les migrations au Canada

Critères	Rapport sur les statistiques mensuelles concernant les allocations familiales	Fichers M0024 des allocations familiales
Univers	Enfants pour lesquels une allocation familiale est versée	Enfants ayant droit aux allocations familiales (par opposition aux enfants pour lesquels une allocation familiale est versée)
Couverture de l'univers	25% de la population totale en 1984. Presques 100% de tous les enfants âgés entre 0 et 17 ans.	Analogue à la couverture des statistiques mensuelles des allocations familiales
Méthode de détermination de la migration	Compilation des avis de changement d'adresse	Compilation des avis de changement d'adresse
Types de migration	Migration interprovinciale, selon la province d'origine et de destination	Analogue aux types de migration couverts par les statistiques mensuelles des allocations familiales, en plus de la migration internationale
Caractéristiques	Origine-destination. Âge: total des enfants de 0 à 17 ans seulement. Taille de la famille: désigne le nombre d'enfants dans la famille	Origine-destination. Âge: année et mois de naissance. Langue (A ou F). Type de compte (parent régulier, parent adoptif, personne de l'étranger ou agence)
Date ou période de référence	Mois: désigne le volume de renseignements traités durant cette période.	Mois: désigne le mois où la migration s'est réellement faite.
Délais	Les données traitées un mois donnée peuvent être obtenues à la fin de ce mois et concernent les migrations d'environ deux mois aupravant	Données diffusées deux fois par année. Renferme des informations sur la migration des six derniers mois. Les statistiques peuvent être obtenues trois mois environ après la fin de la version semi-annuelle.
Possibilité d'avoir des séries chronologiques	À partir de janvier 1974 jusqu'à aujourd'hui pour les données concernant les migrations d'enfants. De 1947 à 1973, seules des données sur les migrations de familles peuvent être obtenues	De décembre 1977 à aujourd'hui
Cohérence	DANS LE TEMPS: modifications apportées en 1974. Léger problèmes depuis 1980. ENTRE LES PROVINCES: bonne	DANS LE TEMPS: généralement bonne. Léger problèmes depuis 1982. ENTRE LES PROVINCES: problèmes posés par l'Ontario depuis 1982. Légers problèmes posés par T.-N. et la N.-É. en 1983
Degré d'informatisation	Dans les bureaux provinciaux, oui. Mais les données sont envoyées au bureau central de Santé et Bien-Être social Canada sous forme d'imprimés	Oui, assez poussé

Tableau 1 (suite)

Critères	Programme F55	Fichier de Revenu Canada
Univers	Enfants ayant droit aux allocations familiales	Souscripteurs de déclaration d'impôt (doivent avoir rempli une déclaration d'impôt deux années consécutives)
Couverture de l'univers	Analogue à la couverture des statistiques memsuelles des allocations familiales	Les souscripteurs dont les déclarations d'impôt s'apparient deux années consécutives représentent environ 75% de la population âgée de 18 ans et plus
Méthode de détermination de la migration	Indicateur symptomatique (estimations des variations de la population totale à partir des variations de la population couverte)	Comparaison de l'adresse de retour figurant sur les déclarations d'impôt appariées. Des corrections sont apportées aux déclarations d'impôt non appariées
Types de migrations	Migrations nettes indirectes	Intraprovinciale, Interprovinciale et Internationale
Caractéristiques	Nombre d'enfants par région géographique. Âge	Origine-destination, grands groupes d'âges, sexe
Date ou période de référence	Deux fois par année, le 1er juin et le 1er décembre (désigne le nombre d'enfants ayant droit aux allocations familiales à ces dates)	Année: désigne la période entre la production d'une déclaration d'impôt deux années consécutives, c'est-à-dire à peu près la période d'avril d'une année à mars de l'année suivante. Les données relatives à cette période servent comme données de la période de juin à mai
Délais	Les statistiques peuvent être obtenues environ trois mois après la date de référence	Des données préliminaires peuvent être obtenues six à huit mois après la fin de la période de référence, alors que les données finales peuvent être obtenues dix à douze mois après la fin de la période de référence
Possibilité d'avoir des séries chronologiques	De décembre 1977 jusqu'à aujourd'hui, pour l'information concernant les enfants ayant droit aux allocations familiales. Il est possible d'avoir des statistique qui remonte à 1974 dans le cas des enfants pour lesquels des allocations familiales ont été versées	De 1966-1967 à aujourd'hui
Cohérence	Généralement bonne	Des modifications aux lois fiscales ont entraîné des changements au niveau de la couverture et du nombre de déclaration d'impôt appariées dans le temps et selon les provinces
Degré d'informatisation	Oui, assez poussé .	Oui, assez poussé

les immigrants nouvellement arrivés et les étudiants étrangers) sont couverts par l'assurance-maladie de la province,à l'exception du personnel de la GRC et des Forces armées et des détenus des prisons fédérales qui sont, eux, couverts par l'assurance-maladie du gouvernement fédéral. Tous ceux qui établissent leur résidence dans une province donnée doivent remplir une demande en bonne et due forme, à partir de laquelle des données peuvent être compilées sur les immigrants internes par province d'origine et sur les immigrants provenant de l'étranger. La couverture presque complète de l'univers, la possibilité d'avoir des données mensuelles, les délais minimums et les données habituellement désagrégées selon l'âge, le sexe et la composition de la famille des migrants sont les principaux avantages de ces fichiers. Les migrants interprovinciaux devraient également être fortement incités à faire une demande pour être couverts par ce programme. Par conséquent, les données sur les migrations provenant de ces fichiers devraient être fiables.

Ces fichiers présentent toutefois certains inconvénients. Leur principale lacune réside dans le fait que ni l'Ontario ni le Québec ne sont en mesure de fournir des données sur les migrations. Des changements s'annoncent dans le cas du Québec mais, pour l'Ontario, il n'y a rien à attendre. À moins de pouvoir générer une source spéciale pour l'Ontario, cela diminuerait le potentiel de ce fichier. Ce fichier risque également de poser des problèmes de cohérence étant donné que chaque province gère son fichier de façon indépendante.

Au niveau infraprovincial, il serait également possible de générer des données sur les migrations étant donné que les bureaux des services de santé des provinces devraient être informés de tout changement d'adresse. Dans les faits, cependant, ce n'est pas toujours le cas.

On pourrait également utiliser les fichiers des services de santé dans les estimations par régression, en particulier dans les provinces qui vérifient périodiquement les adresses pour épurer leur fichier et compter uniquement la population voulue.

2.2 Dossiers de la sécurité de la vieillesse

Santé et Bien-être social Canada est responsable de l'administration du fichier de la sécurité de la vieillesse. Les résidents canadiens âgés de 65 ans et plus qui totalisent suffisamment d'années de résidence au pays sont admissibles. Ils représentent environ 10% de la population totale. Presque toutes les personnes admissibles sont couvertes par le programme. En outre, l'incitation financière à signaler un changement d'adresse est très forte. Un autre avantage du fichier, c'est l'actualité de ses données. Les renseignements sur les personnes qui ont déménagé un mois donné seront compilés et reçus à Statistique Canada deux ou trois mois plus tard. Enfin, la sécurité de la vieillesse, étant un programme fédéral, fournit des données comparables d'une province à l'autre; même si les renseignements sont compilés par les bureaux provinciaux régionaux, la procédure est la même partout au Canada.

La principale lacune de ce fichier, du point de vue l'estimation des migrations, réside dans le fait qu'il concerne une faible partie de la population (variant de 7.3% en Alberta à 12.2%à l'Île-du-Prince-Édouard) et, qui plus est, les personnes âgées ont un comportement migratoire assez différent de celui du reste de la population. Contrairement à la migration chez les enfants, qui peut évidemment être liée à la migration chez les adultes et ensuite faire l'objet d'une projection en vue d'une estimation de la migration totale, le fichier de la sécurité de la vieillesse ne permet pas d'élaborer de méthode analogue et aussi efficace pour estimer la migration totale. Même si le fichier de la sécurité de la vieillesse n'a pu être utilisé comme source principale de données pour produire des estimations de la migration, il doit être considéré comme une estimation très intéressante de la migration chez les personnes âgées.

3. FICHIERS PRÉSENTANT LE PLUS DE POTENTIEL COMME INDICATEURS SYMPTOMATIQUES DES VARIATIONS DE LA POPULATION ET DES MIGRATIONS NETTES

Les données tirées de certains fichiers administratifs pourraient servir à générer des estimations de la population totale. Par exemple, les effectifs scolaires, les abonnés des services

d'hydro-électricité ou les abonnés résidentiels du téléphone pourraient être utilisés dans les techniques de régression comme indicateurs symptomatiques (voir McRae 1985, pour une application des données sur les abonnés des services d'hydro-électricité à l'estimation de la population). Cette méthode et les sources correspondantes sont généralement utilisées pour produire des estimations de la population dans les petites régions, mais si aucune autre méthode ne permet d'obtenir des estimations valables au niveau provincial, ces sources seront sérieusement prises en considération.

3.1 Abonnés des services d'hydro-électricité

Les compagnies d'électricité conservent des dossiers sur leurs clients. Ces dossiers renseignent sur le type de compte (résidentiel, commercial, exploitation agricole, etc.) et l'adresse et le code postal de l'abonné. Dans certaines provinces, il y a un seul fichier pour la province, mais, dans d'autres, deux compagnies (au Manitoba et à Terre-Neuve) ou même plus (en Colombie-Britannique et en Ontario) fournissent l'électricité à la population de la province. Par contre, dans la plupart des provinces, des statistiques peuvent être produites pour le territoire entier, à tout moment et rapidement, mais il y a quand même un petit nombre de provinces pour lesquelles il peut être difficile d'obtenir des données. Les principales faiblesses des fichiers sont de deux types. Outre le problème mentionné plus haut, il est possible qu'il y ait de légères incohérences à cause de la différence qui existe dans les définitions provinciales des ménages résidentiels (puisqu'ils satisfont aux critères administratifs) et même à l'intérieur d'une même province lorsque plus d'une compagnie dessert le territoire. Néanmoins, les abonnés des services d'hydro-électricité pourraient être une très bonne source de données pour l'estimation de la population. En fait, cette source de données a été utilisée en Colombie-Britannique, où des estimations de la population ont été produites pour les municipalités et les districts scolaires, et les résultats ont été bons. Cette méthode pourrait également être utilisée et éventuellement élargie pour produire, au besoin, des estimations au niveau provincial.

3.2 Compagnies de téléphone

Au Canada, la majorité des services téléphoniques sont assurés par 14 principales compagnies de téléphone. Il est possible d'obtenir de l'information sur les abonnés des lignes résidentielles (adresse et code postal). La situation est sensiblement la même que celle des fichiers des abonnés à l'Hydro. Les données peuvent en général être obtenues assez rapidement et la couverture de l'univers est assez élevée. Encore ici, il est possible que plus d'une compagnie desserve un territoire donné et, aussi, qu'une même compagnie desserve plus d'une province. Malgré qu'aucune estimation fondée sur les fichiers des compagnies de téléphone n'ait encore été tentée par Statistique Canada, on estime que ces fichiers ont le potentiel pour produire de bons résultats.

3.3 Inscriptions scolaires

Chaque gouvernement provincial tient à jour un fichier informatisé sur les étudiants inscrits dans son réseau d'écoles primaires et secondaires; ce fichier contient des renseignements sur les adresses des écoles avec le code postal et le nombre d'étudiants, selon l'âge et le niveau d'études. L'information sur le nombre d'étudiants se rapporte au 30 septembre de chaque année et peut être obtenue de quatre à dix mois après la date de référence, les délais variant selon les provinces. La couverture de l'univers des étudiants est également très bonne.

Ce fichier présente certains inconvénients. Par exemple, son caractère annuel empêche qu'il soit utilisé pour produire des estimations trimestrielles. Une date de référence convenant

mal (le 30 septembre au lieu du 1er juin) et des délais pouvant atteindre dix mois sont d'autres handicaps. Au niveau infraprovincial, enfin, on constate souvent qu'un certain nombre d'étudiants résident dans une région administrative donnée, mais vont à l'école dans une autre. Cela aussi risque d'influer sur la qualité des estimations. Il faut remarquer ici que les données sur les effectifs scolaires ont déjà été utilisées par Statistique Canada (et aussi par le U.S. Bureau of the Census, à l'aide d'une méthode des composantes mise au point par ce bureau; voir U.S. Bureau of the Census 1973, Chap. 23, p. 51); les écarts obtenus à l'aide de cette dernière méthode ont été beaucoup plus grands que les écarts obtenus à l'aide d'autres méthodes. Cependant, si aucun autre fichier ne pouvait produire d'estimations adéquates de la population, les estimations obtenues par régression au moyen de ce fichier pourraient être acceptables, au moins au niveau provincial.

4. FICHIERS PRÉSENTANT UN POTENTIEL LIMITÉ

4.1 Permis de conduire

Toutes les provinces ont un fichier des personnes âgées de 15 (ou de 16 ou de 17) ans et plus qui ont un permis les autorisant à conduire un véhicule automobile. À l'aide de ces fichiers provinciaux, les migrations pourraient être estimées de deux façons: 1) par compilation des changements d'adresse sur les permis de conduire pour estimer les flux migratoires et 2) en utilisant les fichiers comme indicateurs symptomatiques des variations de la population établies d'après les variations enregistrées dans le nombre des personnes ayant un permis de conduire dans une région donnée. À l'heure actuelle, l'Ontario utilise les permis de conduire pour estimer les migrations intraprovinciales, mais très peu d'autres provinces pourraient fournir de renseignements sur les flux migratoires, en particulier au niveau infraprovincial. La fourniture de tels renseignements aurait exigé trop de travaux et nécessité des consultations avec le ministre de la province. En dépit du fait que la loi oblige les conducteurs à indiquer leur changement d'adresse, tous ne le font pas et il n'y a pas de statistiques suffisamment détaillées.

Le fichier des permis de conduire pourrait également être utilisé dans les techniques de régression. La possibilité d'obtenir des données à n'importe quelle date précise et rapidement est un élément positif. La couverture de l'univers et les problèmes de cohérence pourraient toutefois réduire la qualité des données. Par exemple, 83% des adultes en Saskatchewan ont un permis de conduire, comparativement à 73% au Manitoba (85% des hommes et 62% des femmes de cette province ont un permis de conduire). En outre, la proportion de ceux qui ont un permis de conduire chez les pauvres, les immigrants relativement récents, les Indiens et les résidents des collectivités éloignées du Nord est inférieure aux moyennes (Stock 1981, p. 44). Pour faire une estimation, il est souvent préférable d'avoir une couverture à 100% d'une petite partie de la population (par exemple les enfants) plutôt qu'une couverture à 80 ou 85% d'une grande partie de la population (par exemple les adultes), surtout si la couverture est sélective sur le plan de la migration. Même si cela ne fait pas nécessairement que le fichier ne convient pas pour l'estimation, cela en affecte le potentiel.

4.2 Permis de bâtir

Statistique Canada recueille des données sur les nouveaux permis de bâtir au niveau des villes et des régions rurales au Canada. En moyenne, les taux de couverture diffèrent selon qu'il s'agit de villes (98.5%) ou de régions rurales (62.5%). Il est possible d'obtenir des données mensuelles sur les permis de bâtir au niveau des divisions de recensement. Ces statistiques pourraient aussi être utilisées comme indicateurs symptomatiques des variations de la population. Un des inconvénients des données sur les permis de bâtir réside toutefois dans

le fait que la date à laquelle ces données s'appliquent est la date d'émission du permis. Pour cette raison, on ne peut être certain que la maison a effectivement été construite, ni qu'elle a été occupée. Un autre inconvénient du fichier, c'est que le nombre de permis émis n'est pas nécessairement directement lié aux variations de la population, surtout dans le cas d'une population décroissante.

Par conséquent, l'utilisation des données sur les permis de construire semble également présenter un potentiel limité pour estimer la population des diverses régions géographiques.

4.3 Commission d'assurance-chômage

La Commission d'assurance-chômage tient à jour une liste des bénéficiaires du programme. Un échantillon couvrant 10% de ce fichier a été prélevé pour produire des statistiques et pourrait produire des renseignements sur les migrations. Toutefois, ce fichier pourrait difficilement être utilisé pour estimer les migrations au Canada, et ce pour deux raisons. Premièrement, un échantillon couvrant 10% de l'univers des chômeurs correspond à moins de 1% de la population totale. Il est impossible de générer des statistiques sur les flux migratoires entre les provinces à partir d'une sous-population aussi petite. Deuxièmement, le fait que cet échantillon ne soit pas représentatif (les jeunes adultes représentant une bonne partie des chômeurs) invite à la prudence relativement aux données sur les migrations tirées de ce fichier.

La Commission tient également à jour un fichier des salariés qui versent des cotisations d'assurance-chômage. Aucune analyse approfondie de ce fichier n'a toutefois encore été faite.

4.4 Enquête sur la population active

En 1982, Statistique Canada a réalisé une enquête-échantillon auprès de 56,000 ménages au Canada. On a posé aux enquêtés de la population civile hors institutions âgée de 15 ans et plus qui faisait partie de l'échantillon une question concernant leurs antécédents migratoires des cinq ou six dernières années. D'autres renseignements utiles peuvent également être obtenus. Toutefois, la très petite taille de l'échantillon (½% de la population totale) et le fait que l'enquête n'a été menée qu'une seule fois éliminent l'Enquête sur la population active comme source de données à des fins d'estimation des migrations.

4.5 Listes électorales

Il est en général possible d'obtenir des données sur les électeurs au Canada. Les listes électorales fédérales et provinciales pourraient facilement être obtenues, tandis qu'il faudrait plus de travail pour obtenir les listes électorales municipales. Ces listes contiennent des renseignements sur le nombre de citoyens canadiens âgés de 18 ans et plus (les immigrants reçus sont inclus au niveau municipal seulement). Elles couvrent en moyenne 90 à 95% de la population cible. La principale lacune de cette source de données réside dans le fait que les statistiques ne peuvent être obtenues à des intervalles réguliers. Les listes électorales fédérales et provinciales sont dressées pour des élections qui ont lieu à peu près tous les quatre ans environ à des dates qui ne conviennent pas à des fins d'estimation. Il semble donc inutile d'envisager d'utiliser les listes électorales.

4.6 Ventes au détail

Des données sur les ventes au détail sont recueillies par Statistique Canada sous forme de chiffre d'affaires auprès de tous les grands magasins et d'un échantillon d'entreprises plus petites. Ces données sont recueillies mensuellement et peuvent être obtenues trois mois après la date de référence. L'utilité de cette série de données semble toutefois limitée sur le plan de l'estimation de la population et de la migration. Cela pourrait être attribuable au fait que les ventes au détail sont très fortement tributaires des fluctuations économiques qui peuvent ne pas toujours traduire fidèlement les variations dans la taille de la population.

4.7 Stastistiques du camionnage (entreprises de déménagement)

Il est possible d'obtenir les statistiques tirées d'un échantillon de cinq entreprises de déménagement parmi les principales au Canada. On pourrait évaluer les flux migratoires interprovinciaux en pondérant le nombre de déménagements déclarés entre deux provinces ou territoires. L'utilité des statistiques du camionnage est toutefois sérieusement diminuée par le fait qu'elles datent toujours d'au moins deux ans.

5. CONCLUSIONS

Dans le présent rapport, nous avons présenté une vue d'ensemble des avantages et inconvénients de douze fichiers de données administratives dans le but de faire des recommandations concernant le choix d'une source de données autre que les fichiers des allocations familiales. Nous avons constaté qu'aucun fichier n'avait des avantages et des inconvénients strictement comparables à ceux des fichiers des allocations familiales. Toutefois, advenant le cas où le programme des allocations familiales cesserait d'être universel, nous pourrions faire les recommandations suivantes:

- continuer d'explorer la possibilité d'utiliser le fichier de l'assurance-maladie des provinces et les dossiers de la sécurité de la vieillesse du gouvernement fédéral pour produire des estimations de la population totale et de la migration sur une base trimestrielle;
- examiner la qualité des estimations annuelles de la population des provinces et des territoires produites par la Méthode des composantes II à l'aide des estimations de la migration générées à partir des fichiers provinciaux des données sur les effectifs scolaires;
- vérifier la qualité des statistiques obtenues en utilisant les fichiers provinciaux de données administratives (fichiers de l'assurance-maladie, abonnés à l'Hydro, compagnies de téléphone et permis de conduire) comme indicateurs symptomatiques des variations de la population et de la migration nette résiduelle pour les régions infraprovinciales (divisions de recensement et régions métropolitaines de recensement au Canada).

BIBLIOGRAPHIE

ALMOND, M.M. (1982). An Inventory of Sources of Canadian Migration Data. Document de travail, Division de la démographie, Statistique Canada.

McRAE, D.G. (1985). Utilisation des comptes de l'hydro dans le modèle d'estimation par régression de la population en Colombie-Britannique. Document présenté au Comité fédéral-provincial de la démographie, Ottawa, Canada.

NORRIS, D.A., et STANDISH, L.D. (1983). Un Rapport technique sur la production d'estimations des migrations à partir de dossiers fiscaux. Rapport technique, Division de l'exploitation des données administratives, Statistique Canada.

NORRIS, D.A. (1983). Nouvelles sources de données sur la migration à l'échelle des petites régions. *Review of Public Data Use*, 11-25.

STATISTIQUE CANADA (Trimestriel). *Estimations trimestrielles de la population du Canada, des provinces et des territoires*. N° 91-001 au catalogue, Ottawa ministre des Approvisionnements et Services Canada.

STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population suivant l'état matrimonial, l'âge, le sexe et composantes de l'accroissement, Canada, provinces et territoires*. N° 91-210 au catalogue, Ottawa, ministre des Approvisionnements et Services Canada.

STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population des divisions et régions métropolitaines de recensement (méthode de régression)*. N° 91-211 au catalogue, Ottawa, ministre des Approvisionnements et Services Canada.

STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population des divisions et régions métropolitaines de recensement (méthode des composantes)*. Ottawa, N° 91-212 au catalogue, Ottawa, ministre des Approvisionnements et Services Canada.

STOCK, R. (1981). *Migration Estimates from Current Administrative Files: Data Sources and Methodologies*. Canadian Plains Research Center, Université de Régina.

U.S. BUREAU OF THE CENSUS (1973). *The Methods and Materials of Demography*. Washington, D.C.: U.S. Government Printing Office.

VERMA, R.B.P., BASAVARAJAPPA, K.G., et BENDER, R.K. (1984). The Regression Estimates of Population for Sub-provincial Areas in Canada. *Techniques d'enquête*, 9, 219-240.

VERMA, R.B.P., et BASAVARAJAPPA, K.G. (1985). Recent Developments in the Estimation of Population for Small Areas in Canada by Regression. Document présenté au Symposium international sur les statistiques régionales, Ottawa, Canada.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 203–209
Statistique Canada

Utilisation de fichiers de données administratives pour la production d'estimations de la migration: une étude de cas du fichier des permis de conduire en Ontario[1]

RAGHUBAR D. SHARMA et CHEUK WONG[2]

RÉSUMÉ

Au Canada, les démographes provinciaux et fédéraux ont entrepris d'utiliser divers ensembles de données administratives pour estimer les flux migratoires. Le présent document traite de la production d'estimations des migrations intraprovinciales à l'aide de données sur les permis de conduire en Ontario. Une évaluation de ces estimations de la migration est faite à l'aide d'une comparaison des estimations produites à partir des données de l'impôt sur le revenu de Statistique Canada. Les deux fichiers donnent des estimations tout aussi bonnes et complémentaires des migrations intraprovinciales.

MOTS CLÉS: Fichiers administratifs; estimations de population; méthode des composantes; petites régions; erreur en fin de période; migration intraprovinciale.

1. INTRODUCTION

La migration est une composante importante des projections et des estimations de la population. Étant donné qu'on ne tient aucun fichier concernant le mouvement de la population au Canada, les démographes du gouvernement fédéral et ceux des administrations provinciales ont entrepris d'utiliser divers ensembles de données administratives pour estimer les flux migratoires. Statistique Canada utilise les données sur le revenu (Norris et Standish 1983), la Colombie-Britannique utilise les listes d'abonnés des services d'hydro-électricité (McRae 1985) et l'Alberta utilise les dossiers sur les bénéficiaires des soins de santé (Bureau de la statistique de l'Alberta 1985). Depuis 1979, l'Ontario utilise les changements d'adresse des détenteurs de permis de conduire pour estimer les migrations intraprovinciales. Outre leur fiabilité, les données sur les permis de conduire ont cet autre avantage important d'être diffusées rapidement. En effet, le délai entre la date de réception des données et la date de référence n'est que de 4 à 5 semaines comparativement à plus d'une année et demie dans le cas des données sur le revenu. Dans le présent document, nous évaluons les estimations des migrations intraprovinciales produites à partir des données sur les permis de conduire en Ontario. Aux États-Unis, l'État de Californie utilise aussi les changements d'adresse des détenteurs de permis de conduire pour estimer les mouvements migratoires à l'intérieur de l'État (Hoag 1984).

2. FICHIER DES DONNÉES SUR LES PERMIS DE CONDUIRE

L'information relative aux changements d'adresse des détenteurs de permis de conduire est fournie par le ministère des Transports et des Communications de l'Ontario. Un conducteur doit aviser le Ministère de sa nouvelle adresse dans les 90 jours suivant la date de son changement d'adresse. L'information est ventilée par région de code postal. Ces régions peuvent ensuite être converties en régions infraprovinciales, comme des comtés et des municipalités. Comme le tableau 1 l'indique, les données peuvent être obtenues pour les sept dernières années. Depuis 1979, les données peuvent également être produites par trimestre.

---

[1] Version abrégée d'un document présenté aux réunions du Comité fédéral-provincial de la démographie, les 28 et 29 novembre, 1985, Statistique Canada, Ottawa.

[2] Raghubar D. Sharma et Cheuk Wong, Direction des politiques sectorielles et régionales, ministère du Trésor et de l'Économie de l'Ontario, Queen's Park, Toronto (Ontario), M7A 1Y9.

Plus d'un million de changements d'adresse sont enregistrés chaque année. La plupart de ces déménagements se produisent à l'intérieur même des divisions de recensement (c.-à-d. un comté ou une municipalité régionale). Cependant, les migrations nettes inter-comtés s'élevent en moyenne à 22,000 seulement par année. On constate, d'après le tableau 1, qu'un tiers environ des enregistrements n'indiquent pas le code postal de l'origine et/ou de la destination du migrant.

En Ontario, une personne peut obtenir un permis de conduire dès l'âge de 16 ans. Aussi plus de 75% de la population visée détient un permis de conduire. Par ailleurs, la proportion de personnes âgées et de femmes qui possèdent un permis de conduire est moins élevé que celle des autres catégories de détenteurs (tableau 2).

Tableau 1

Nombre total de migrants et nombre de migrants dont l'origine et/ou la destination n'ont pas été précisées, Ontario, 1975-1985

Année	Nombre de migrations dont l'origine et la destination sont connues (inter-comté et infracomté)	Nombre de migrations dont l'origine et/ou la destination n'ont pas été précisées	Total	Pourcentage de migration non précisées
1979 (année civile)	881,000	0	881,000	0
1979/80	586,000	301,000	887,000	34
1980/81	566,000	306,000	872,000	35
1981/82	617,000	270,000	887,000	30
1982/83	648,000	259,000	907,000	29
1983/84	822,000	320,000	1,142,000	28
1984/85	831,000	330,000	1,161,000	28

Source: Ministère des Transports et des Communications de l'Ontario.

Tableau 2

Pourcentage de la population détenant un permis de conduire 1981

	% de la population détenant un permis de conduire, 1981		
Âge	Hommes	Femmes	Total
16-19 ans	63.9	36.5	49.1
20-24 ans	92.7	73.0	85.7
25-34 ans	98.6	81.6	90.0
35-44 ans	99.7	79.9	90.0
45-54 ans	96.7	67.8	82.4
55-64 ans	93.2	56.7	74.2
65 ans et plus	73.1	27.4	46.4
Total	90.6	62.2	75.8

Source: Ministère des Transports et des Communications de l'Ontario.

3. CONVERSION DU NOMBRE DE CONDUCTEURS EN NOMBRE DE MIGRANTS

Pour connaître le nombre de migrants, un facteur de rajustement est appliqué au nombre de conducteurs. Ce facteur est calculé de la façon suivante:

$$FA = \frac{\text{Total des mouvements connus et des mouvements inconnus}}{\text{Nombre de mouvements connus}}$$

$$FB = \frac{\text{Population totale}}{\text{Population détenant un permis de conduire}}$$

$$F = FA \times FB.$$

FA tient compte des origines et/ou des destinations non déclarées tandis que FB tient compte des personnes qui ne détiennent pas de permis de conduire. Le facteur de rajustement *FA* suppose que les comportements migratoires de ceux qui n'ont pas de permis de conduire ne diffèrent pas des comportements migratoires des détenteurs d'un permis de conduire, tout comme le facteur *FB* suppose que les comportements migratoires des migrants dont les mouvements n'ont pas été précisés ne diffèrent pas des comportements migratoires des migrants dont les mouvements ont été précisés.

4. ESTIMATION DES MIGRATIONS INTRAPROVINCIALES: FICHIER DES PERMIS DE CONDUIRE ET FICHIER DE L'IMPÔT SUR LE REVENU

Statistique Canada utilise les changements d'adresse indiqués par les contribuables dans leur déclaration annuelle d'impôt. Le nombre d'enfants est estimé à partir du nombre de personnes à charge déclaré par le contribuable. Comme dans le cas des données sur les permis de conduire, il faut introduire un facteur de rajustement dans les données sur le revenu pour tenir compte des cas de codes postaux non précisés et des cas des personnes qui ne produisent pas de déclaration d'impôt. En outre, certains contribuables inscrivent une adresse postale autre que celle de leur domicile dans leur déclaration d'impôt.

Il a fallu vérifier la qualité des estimations des migrations produites à partir du fichier de l'impôt sur le revenu et celle des estimations des migrations produites à partir du fichier des données sur les permis de conduire. Nous avons utilisé trois mesures pour vérifier l'exactitude relative des estimations des migrations produites à partir de ces deux fichiers; ce sont:

A. les erreurs en fin de période intercensitaire
B. la comparaison des taux de croissance
C. l'indice de dissemblance.

Idéalement, les erreurs en fin de période intercensitaire et les taux de croissance devraient être calculés à partir des estimations de la population d'une année de recensement donnée à l'année de recensement suivante. Comme on dispose de données fiables sur les changements d'adresse des détenteurs de permis de conduire seulement depuis 1979 en Ontario, les estimations intercensitaires de la population faites en 1979 et les chiffres estimatifs de la population en 1981 ont servi de base aux calculs. Deux séries d'estimations de la population ont été calculées pour l'estimation des migrations intraprovinciales: une première à l'aide du fichier des adresses des détenteurs de permis de conduire et une deuxième à l'aide du fichier de l'impôt sur le revenu. Toutes les autres composantes, c'est-à-dire les naissances, les décès, les migrations interprovinciales et les migrations internationales sont demeurées les mêmes pour les deux séries d'estimations de la population.

4.1 Erreurs en fin de période intercensitaire

Nous avons comparé les deux séries d'estimations de la population des divisions de recensement (l'une produite à partir des données sur les permis de conduire et l'autre, à partir des données sur l'impôt sur le revenu, dans les deux cas pour les migrations intraprovinciales) avec les chiffres de population du recensement de 1981. La différence exprimée en pourcentage entre la population estimative et le chiffre de population du recensement est appelée "erreur en fin de période intercensitaire". Pour 23 des 49 divisions de recensement, l'erreur est plus petite lorsque les données sur les permis de conduire sont utilisées et, pour 26 DR, l'erreur est plus petite lorsque les données sur l'impôt sur le revenu sont utilisées pour l'estimation des migrations intraprovinciales.

4.2 Conurbation de Toronto

La conurbation de Toronto, qui englobe six municipalités régionales (tableau 3), est un cas assez intéressant. Les données sur les permis de conduire produisent une erreur en fin de période intercensitaire plus petite pour la conurbation dans son ensemble et sous-estiment la population des régions situées à l'extérieur de la région métropolitaine de Toronto.

Avec les données du fichier de l'impôt sur le revenu, on obtient une erreur en fin de période intercensitaire plus petite pour chacune des divisions de recensement, alors que dans le cas de la conurbation dans son ensemble, l'erreur en fin de période intercensitaire est plus grande que celle qui est calculée à partir des données du fichier des permis de conduire (tableau 3). Pour cette raison, nous avons utilisé les données sur les permis de conduire pour estimer les migrations intraprovinciales des résidents de la conurbation de Toronto dans son ensemble, alors que nous avons réparti les migrants selon chaque division de recensement de la conurbation d'après les données sur le revenu.

4.3 Taux de croissance de la population

Nous avons calculé le taux de variation entre 1979 et 1981 des estimations de la population produites à partir du fichier de l'impôt sur le revenu et celui des estimations produites à partir des données sur les permis de conduire. Les taux de croissance ainsi obtenus ont ensuite été comparés aux taux de croissance indiqués par les données du recensement de 1981.

Les taux de croissance obtenus à partir des deux fichiers ne diffèrent pas beaucoup des taux de croissance indiqués par les données du recensement. Les divisions de recensement où la population ne varie pas dans le même sens que le taux de croissance indiqué par les données du recensement sont au nombre de 3, lorsque les données du fichier des permis

Tableau 3

Erreurs en fin de période intercensitaire, conurbation de Toronto

Division de recensement	Erreurs en fin de période intercensitaire	
	Fichier impôt sur le revenu	Fichier permis de conduire
M.R. de Durham	−0.21	−0.41
M.R. de Halton	0.45	−0.51
M.R. de Hamilton-Wentworth	0.10	−0.11
M.R. de Peel	−0.63	−1.94
M.R. de Toronto	0.01	1.26
M.R. de York	−0.11	−5.70
Conurbation de Toronto, Total	−0.05	−0.01

de conduire sont utilisées, et au nombre de 10, lorsque les données du fichier de l'impôt sur le revenu sont utilisées (tableau 4). Il s'agit d'un cas où les données sur les permis de conduire ont semblé produire des estimations plus fiables que les estimations produites à partir des données du fichier de l'impôt sur le revenu.

4.4 Indice de dissemblance

L'indice de dissemblance a été calculé pour la migration d'entrée et la migration de sortie séparément, étant donné que le sens de la migration nette indiqué par les deux séries d'estimations n'était pas le même dans certains comtés. La valeur de l'indice de dissemblance peut varier de 0 à 100. Cet indice mesure la moitié de la somme des écarts absolus entre les deux répartitions en pourcentage correspondantes et équivaut à la somme des écarts positifs ou à la somme des écarts négatifs (Shryock et Siegel, 19781). La formule générale est:

$$ID = \tfrac{1}{2} \sum |r_2 - r_1|$$

où r_2 et r_1 sont les pourcentages correspondants dans les deux répartitions.

Tableau 4

Division de recensement où la variation de population est différente de celle indiquée par les données du recensement

Division de recensement	Variation en % d'après	
	les données sur l'impôt sur le revenu	les données du recensement
Bruce	0.11	−1.77
Grey	−0.04	0.17
Hastings	0.12	−1.03
Leeds et Grenville	0.21	−0.32
Niagara	0.15	−0.46
Northumberland	0.31	−0.82
Oxford	0.17	−0.09
Parry Sound	1.20	−0.51
Stormond/Dundas/Glengarry	0.44	−0.17
Sudbury T.D.	0.41	−2.28
Province	1.60	1.46
	Variation en % d'après	
	les données sur les permis de conduire	les données du recensement
Leeds et Grenville	0.02	−0.32
Parry Sound	0.81	−0.51
Thunder Bay	0.42	−0.20
Province	1.60	1.46

Tableau 5
Indice de dissemblance

	Indice de dissemblance	
Année	Migration d'entrée	Migration de sortie
1979-80	5.61	3.50
1980-81	5.54	3.82
1981-82	5.26	4.82
1982-83	4.41	4.87

Les valeurs peu élevées de l'indice indiquent que les deux fichiers (des permis de conduire et de l'impôt sur le revenu) produisent des estimations assez semblables des migrations intraprovinciales dans le cas des divisions de recensement en Ontario. Au cours des quatre dernières années, toutefois, le degré de dissemblance augmente pour ce qui a trait à la migration de sortie et diminue pour ce qui a trait à la migration d'entrée.

5. CONCLUSION ET SOMMAIRE

Cet exposé présente une comparaison des estimations des migrations intraprovinciales produites à partir du fichier des permis de conduire et de celles qui ont été produites à partir du fichier de l'impôt sur le revenu. Les deux fichiers donnent des mesures raisonnablement bonnes des migrations intraprovinciales pour les divisions de recensement en Ontario.

Tandis que les données du fichier des permis de conduire semblent avoir produit de meilleures estimations de la direction des migrations intraprovinciales, l'utilisation des données du fichier de l'impôt sur le revenu a produit des erreurs en fin de période intercensitaire plus petites pour un nombre légèrement plus grand de comtés et a de plus donné des résultats un peu meilleurs pour certaines grandes régions (par exemple, la répartition des migrants dans la conurbation de Toronto/Hamilton). Étant donné leurs avantages respectifs, la meilleure méthode semble être de combiner l'utilisation des deux sources de données.

Il faut également noter que l'évaluation a porté sur trois années seulement, c'est-à-dire de 1979 à 1981. Il est impossible d'évaluer de façon plus précise la qualité des estimations produites à partir de ces deux fichiers tant que les données du recensement de 1986 ne seront pas connues.

Pour améliorer davantage la qualité et les applications des estimations produites à partir des données sur les permis de conduire, nous proposons d'orienter les recherches dans les deux domaines suivants:

- Vérification du facteur de rajustement FA en tenant compte des chiffres réels des migrations dont l'origine et/ou la destination sont inconnues au moyen du codage manuel des adresses.
- Extension de l'utilisation du fichier des données sur les permis de conduire comme source s'ajoutant aux données des allocations familiales et aux données sur le revenu, pour l'estimation des migrations interprovinciales.

Le fichier des permis de conduire tend à surestimer le nombre de migrants de la région métropolitaine de Toronto et à sous-estimer celui des migrants des régions environnantes. Il semble que ce soit l'inverse qui se produit avec le fichier de l'impôt sur le revenu. Pour l'ensemble de la conurbation de Toronto, le fichier des permis de conduire produit de meilleures estimations des migrations intraprovinciales que le fichier de l'impôt sur le revenu qui, en revanche, donne une meilleure répartition des migrations intraprovinciales dans les comtés de cette région.

BIBLIOGRAPHIE

BUREAU DE LA STATISTIQUE DE L'ALBERTA (1985). Utilisation des dossiers de l'assurance-maladie de l'Alberta pour la production d'estimations régionales de la population. Document présenté aux réunions du Comité fédéral-provincial de la démographie, Ottawa.

HOAG, ELIZABETH (1984). Estimating Annual Migration for California Counties using Driver's Licence Address Change. Document présenté aux réunions de la Population Association of America, Minneapolis, Minnesota.

McRAE, DONALD G. (1985). Utilisation des comptes de l'Hydro dans le modèle d'estimation par régression de la population de la Colombie-Britannique. Document présenté aux réunions du Comité fédéral-provincial de la démographie, Ottawa.

NORRIS, D., et STANDISH, L. (1983). *Rapport technique sur la production d'estimations à partir de dossiers fiscaux*. Division de l'exploitation des données administratives, Statistique Canada.

SHRYOCK, HENRY, et SIEGEL, JACOB S. (1971). *The Methods and Materials od Demography*. Washington: U.S. Bureau of Census.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 211–221
Statistique Canada

Utilisation des dossiers de l'assurance-maladie de l'Alberta pour la production d'estimations régionales de la population[1]

F. AHMAD, R. CHOW, O. DEVRIES,
A. HASHMI, et M. MARCOGLIESE[2]

RÉSUMÉ

Ce document traite de l'utilisation des fichiers administratifs du régime d'assurance-maladie de l'Alberta et de la statistique de l'état civil pour la production d'estimations de la population. Les données obtenues et comparées avec les données du recensement montrent que les fichiers d'assurance-maladie peuvent servir à la production d'estimations précises à l'échelle de la province et au niveau de petites régions telles que les divisions de recensement et les municipalités.

MOTS CLÉS: Fichiers administratifs; méthode des composantes; petites régions; migration nette résiduelle.

1. HISTORIQUE

Du milieu à la fin des années 70, la province de l'Alberta a connu une croissance économique rapide, en raison surtout de l'activité du secteur du pétrole et du gaz, qui a entraîné un accroissement marqué de la population. Les divers ordres de gouvernement ont exigé des données à jour sur l'importance et la répartition régionale de cette croissance démographique, afin de pouvoir offrir à la population des biens et services adéquats. Face à ce besoin, on a jugé que les recensements menés tous les cinq ans par le gouvernement fédéral n'étaient ni assez fréquents, ni suffisamment à jour puisque les chiffres du recensement sont publiés de douze à dix-huit mois après l'année de référence. En conséquence, des organismes provinciaux, et tout particulièrement le Bureau de la statistique de l'Alberta (désigné ci-après le Bureau), ont entrepris des recherches en vue de trouver de nouvelles sources de données, disponibles rapidement, sur la population.

Après avoir examiné un certain nombre de possibilités, le Bureau a procédé à l'évaluation des données administratives sur l'assurance-maladie tirées des fichiers de l'Alberta Health Care Insurance Plan (AHCIP) en vue de la production de statistiques démographiques. Le présent document souligne les travaux entrepris par le Bureau pour faire en sorte que les données tirées des dossiers de l'AHCIP puissent servir à des estimations de population pour des petites régions.

2. PRODUCTION DE CHIFFRES SUR LES BÉNÉFICIAIRES DE L'ASSURANCE-MALADIE À PARTIR DES DOSSIERS DE L'AHCIP

La présente partie décrit brièvement la nature des dossiers de l'AHCIP et évalue les totalisations produites à partir de l'information qui y est contenue.

2.1 Élaboration des données sur les bénéficiaires de l'assurance-maladie

Chaque trimestre, le Bureau reçoit sur bande informatisée un certain nombre d'enregistrements tirés du système d'inscription et de facturation de l'AHCIP. Le fichier

[1] Version condensée du document présenté à la renconte du Comité fédéral-provincial de la démographie tenue les 26 et 27 novembre 1985 à Ottawa.

[2] F. Ahmad, R. Chow, O. DeVries, A. Hashmi et M. Marcogliese, Bureau de la statistique de l'Alberta, Sir Frederik W. Haultain Building, 9811-109th Street, Edmonton, Alberta, Canada, T5K 0C8.

sur bande constitue une liste partielle puisqu'on y a biffé tous les noms et autres identifiants de façon à préserver le caractère confidentiel des renseignements sur les individus. Les renseignements types contenus dans le fichier sont l'adresse, le code postal, les dates d'inscription et d'annulation, l'âge et le sexe. (On peut obtenir sur demande une description détaillée du cliché d'enregistrement).

L'unité de déclaration du fichier de l'AHCIP est l'inscription. Chaque inscription peut comprendre jusqu'à vingt-cinq individus: un requérant (habituellement la personne qui paie les primes) et jusqu'à concurrence de vingt-quatre personnes à charge. Il y a actuellement environ 1.7 million de dossiers d'inscription actifs qui représentent environ 2.6 millions d'individus. De plus, le fichier est chronologique et englobe toutes les personnes protégées en vertu de l'AHCIP depuis son entrée en vigueur en 1969.

Le traitement du fichier se fait en quatre étapes:

a) Vérification - notation ou correction des erreurs en fonction de critères de contrôle précis.
b) Épuration - sélection des individus actifs à partir du fichier des données corrigées à l'étape d).
c) Codification - appariement des codes postaux du fichier épuré et du fichier de conversion des codes postaux du Bureau (FCCP) et attribution de codes de référence géographiques aux enregistrements de l'AHCIP.
d) Agrégation - totalisation des individus des deux sexes par année d'âge et pour chaque code postal, à partir du fichier codifié. Cette opération permet de réduire le nombre des enregistrements/individus d'environ 2.6 millions à moins de 120,000 et donc de diminuer sensiblement les coûts de traitement informatique subséquents.

Le fichier agrégatif sert à la production de chiffres de la population selon l'âge et le sexe pour toute région géographique pouvant être définie à l'aide des 60,000 codes du FCCP de l'Alberta.

2.2 Évaluation des résultats

Pour évaluer les chiffres sur les bénéficiaires de l'assurance-maladie, on les a comparés aux chiffres du recensement de la population du Canada de 1976 et de 1981. Comme les données du fichier de l'AHCIP de 1981 ont été jugées plus précises que celles du fichier de 1976, on s'est davantage fondé sur les comparaisons avec les chiffres du recensement de 1981. Les données des recensements municipaux ont été utilisées comme deuxième base de comparaison, bien qu'elles ne soient pas jugées aussi fiables dans l'ensemble que les chiffres du recensement du Canada. Toutefois, les données des recensements municipaux ont permis de mieux saisir l'importance des variations ainsi que les distributions relatives selon l'âge et le sexe et l'évolution des tendances (croissance ou déclin). On a en outre établi des comparaisons avec les estimations intercensitaires de la population produites par le Bureau et par Statistique Canada.

Principales constatations

a) À l'échelle provinciale, les chiffres basés sur les dossiers de l'AHCIP surestiment d'environ 3.5% à 4.5% aussi bien les chiffres du recensement du Canada que les chiffres des recensements municipaux. Les distributions selon l'âge et le sexe sont plus précises et les coefficients de corrélation témoignent de l'uniformité des tendances (surestimation/sous-estimation) dans le temps.
b) Au niveau des divisions de recensement (DR), on enregistre des écarts allant de −2.6% à 9.7% entre les chiffres de l'AHCIP et ceux du recensement du Canada (voir tableau 1). La variation est sensiblement la même dans le cas des comparaisons avec les estimations intercensitaires de la population. Comme pour les données provinciales, les distributions selon l'âge et le sexe et l'uniformité des tendances se sont révélées d'une fiabilité élevée.

Tableau 1
Comparaisons des chiffres selon l'AHCIP et des chiffres du recensement du Canada, divisions de recensement de l'Alberta

Division de recensement	Année							
	1976				1981			
	Chiffres du recensement	Chiffres selon l'AHCIP	Écart en pourcentage	Écart en chiffres réels	Chiffres du recensement	Chiffres selon l'AHCIP	Écart en pourcentage	Écart en chiffres réels
1	46,990	45,789	−2.56	−1,201	55,375	55,748	0.67	373
2	96,995	97,229	0.24	234	110,477	111,567	0.99	1,090
3	32,898	33,884	3.00	986	35,652	36,463	2.27	811
4	12,130	12,101	−0.24	−29	12,119	12,038	−0.67	−81
5	35,424	35,656	0.65	232	38,382	38,457	0.20	75
6	524,554	538,432	2.65	13,878	668,682	699,999	4.68	31,317
7	37,866	38,235	0.97	369	40,071	40,359	0.72	288
8	95,384	95,063	−0.34	−321	123,642	124,666	0.83	1,024
9	19,903	21,832	9.69	1,929	21,670	23,338	7.70	1,668
10	67,171	67,168	0.00	−3	78,417	78,532	0.15	115
11	632,909	646,799	2.19	13,890	762,041	796,884	4.57	34,843
12	63,129	62,011	−1.77	−1,118	84,221	86,183	2.33	1,962
13	46,305	47,258	2.06	953	53,701	54,282	1.08	581
14	19,386	21,039	8.53	1,653	24,635	25,991	5.50	1,356
15	106,993	111,678	4.38	4,685	128,639	134,451	4.52	5,812
Non connue[a]		48,462				19,279		
Alberta	1,838,037	1,922,636	4.60	84,599	2,237,724	2,338,237	4.49	100,513

[a] Cette rubrique englobe les totalisations obtenues pour tous les cas pour lesquels il n'y avait pas d'identification d'adresse.
Source: Recensements de Statistique Canada de 1976 et 1981; Alberta Health Care Insurance Plan

c) Au niveau des subdivisions de recensement unifiées (SRU), l'écart entre les données sur les bénéficiaires de l'assurance-maladie et les chiffres du recensement ±10% dans cinquante des soixante et onze SRU. L'écart le plus important est de −56.5% (district municipal 135).

Pour la plupart des régions problèmes, on a constaté que le comté, district municipal ou district d'amélioration locale était situé à proximité d'un grand centre urbain. Aucune anomalie précise n'a été relevée lors de la vérification des distributions selon l'âge et le sexe, quoique les relations n'étaient pas aussi étroites que dans le cas des comparaisons au niveau de la province et des divisions de recensement.

d) Au niveau des subdivisions de recensement (SDR), les chiffres préliminaires révélaient des écarts variant entre −100% et +955% entre les chiffres selon l'AHCIP et les chiffres du recensement de 1981. On a donc décidé d'utiliser à ce niveau d'agrégation les vingt-neuf régions les plus importantes comptant plus de 5,000 habitants. Les six SDR les plus grandes (Edmonton, Calgary, Lethbridge, Medicine Hat, Red Deer et St. Albert) affichaient un surdénombrement variant de 3% à 9%. Pour huit autres SDR, l'écart se situait à ±20% et dans seize autre cas, la variation était légèrement supérieure à ±20%. Une fois de plus, aucune anomalie précise n'a été décelée dans la distribution selon l'âge et le sexe, quoique les écarts étaient plus importants qu'à des niveaux d'agrégation plus hauts. A preuve, vingt-sept des vingt-neuf SDR observées affichaient un niveau élevé d'uniformité des tendances.

À mesure que la taille de la région géographique diminue, les chiffres selon l'AHCIP deviennent moins fiables; les distributions selon l'âge et le sexe, bien que moins précises, demeurent adéquates et l'uniformité des tendances (évolution des chiffres dans le temps) continue à présenter un degré élevé de corrélation, à quelques exceptions près. Les limites que présentent les chiffres basés sur les dossiers de l'AHCIP en tant qu'indicateurs de la population peuvent essentiellement être attribuées à l'une des deux grandes causes suivantes: a) les inexactitudes imputables aux procédures administratives de l'AHCIP; b) l'utilisation des codes postaux.

a) *Procédures administratives de l'AHCIP*

1) Comme il s'agit d'un régime d'assurance, la préoccupation première est d'offrir le service à l'ensemble de la population. En conséquence, pour assurer l'universalité du régime, on consacre davantage d'efforts à l'entrée des gens dans le système qu'au contrôle des sorties. C'est pourquoi le nombre des gens inscrits dépasse la population réelle de la province.
2) On utilise les adresses postales plutôt que les adresses résidentielles, ce qui complique l'attribution des codes géographiques. On note des écarts importants dans les régions comprenant une population rurale importante qui vit à proximité d'un centre urbain et ramasse son courrier dans le centre urbain. Le cas échéant, il y a surdénombrement dans la plupart des régions urbaines et sous-dénombrement dans les régions rurales.
3) Des données incomplètes et inexactes, particulièrement celles touchant aux codes postaux, rendent difficile la production de statistiques régionales à cause du sous-dénombrement.
4) Les délais de déclaration et d'enregistrement des données influent sur les totalisations. Dans la plupart des cas, il faut compter de trois à six mois avant qu'un individu n'entre dans le système (naissance, immigration), mais il faut encore plus de temps pour retirer un dossier actif du système (décès, émigration). Toutefois, ces délais sont difficiles à calculer et varient considérablement selon les circonstances.

b) *Codes postaux*

1) Le code postal délimite la zone du service de livraison (l'endroit où une personne ramasse son courrier) qui n'est pas nécessairement le lieu de résidence. Ce facteur limite la précision de la répartition géographique des enregistrements de l'AHCIP. Comme on l'a vu, cela entraîne un problème de démarcation entre les régions urbaines et rurales.
2) Le code postal à six chiffres ne suffit pas toujours pour déterminer la zone du service de livraison. Pour une adresse plus exacte, il est parfois nécessaire de préciser également la route rurale, le service de banlieue ou la case postale.
3) Les codes postaux n'ont pu donner des niveaux d'agrégation adéquats, surtout dans les régions rurales. À titre d'exemple, il y a environ 363 subdivisions de recensement en Alberta, mais le FCCP du Bureau peut en extraire seulement 324.

Les problèmes énoncés ci-dessus ont empêché la diffusion des chiffres basés sur les dossiers de l'AHCIP en tant qu'approximations de la population réelle. Il est vrai qu'on a obtenu un degré élevé de précision dans certaines régions, mais dans d'autres, les chiffres fondés sur les dossiers de l'AHCIP étaient peu représentatifs. Cependant, compte tenu des relations étroites entre les distributions selon l'âge et le sexe des bénéficiaires de l'assurance-maladie et les distributions fondées sur les chiffres du recensement du Canada et compte tenu également de l'uniformité des tendances dans le temps, les chiffres basés sur les dossiers de l'AHCIP ont été inclus dans la méthode qu'utilise le Bureau pour produire des estimations de la population (méthode expliquée dans la partie qui suit).

3. UTILISATION DES CHIFFRES SELON L'AHCIP POUR LA PRODUCTION D'ESTIMATIONS DE LA POPULATION DES PETITES RÉGIONS

Depuis près de dix ans, le Bureau produit des estimations intercensitaires de la population de l'Alberta et de ses divisions de recensement. Au cours de cette période, diverses méthodes ainsi que diverses sources de données ont été analysées et quelques-unes utilisées pour améliorer la qualité des estimations. À ce jour, on a obtenu des résultats très satisfaisants avec la méthode des composantes, en utilisant comme données d'entrée les chiffres sur les bénéficiaires de l'assurance-maladie. Ces données ont servi à calculer la structure par âge et par sexe de la population de l'Alberta, au niveau de la province et des divisions de recensement, et à produire des estimations de la population à ces mêmes niveaux. Elles ont également été utilisées récemment dans des tests visant à déterminer leur utilité dans la production d'estimations de la population au niveau des subdivisions de recensement.

3.1 Méthode d'estimation

La méthode d'estimation utilisée par le Bureau pour produire des estimations infraprovinciales de la population comprend deux volets: l'estimation de la population migrante et la production des chiffres de population.

a) Estimation de la population migrante à partir des chiffres sur les bénéficiaires de l'assurance-maladie

Le bureau tire ses données de trois fichiers administratifs: les chiffres fondés sur les dossiers de l'AHCIP; les données sur les naissances fournies par la statistique de l'état civil de l'Alberta; et les données sur les décès également fournies par la statistique de l'état civil de l'Alberta. Les données des trois sources servent au calcul de la migration nette. Fondamentalement, pour toute petite région donnée, on mesure l'accroissement du nombre des bénéficiaires de l'assurance-maladie en faisant la différence des totaux enregistrés entre le moment t et le moment t-1. De ce résidu, on retranche l'accroissement naturel de la population de la région (les naissances moins les décès) pour calculer les entrées (ou les sorties) nettes d'individus, c'est-à-dire la migration nette. L'équation mathématique prend la forme:

$$HMIG = [(HC_t - HC_{t-1}) - (B - D)]$$

où:

$HMIG$ = estimation de la migration nette des bénéficiaires de l'assurance-maladie entre le moment t et le moment $t-1$

HC_t = nombre de bénéficiaires de l'assurance-maladie au moment t

HC_{t-1} = nombre des bénéficiaires de l'assurance-maladie au moment $t-1$

B = naissances totales au cours de l'intervalle t à $t-1$

D = décès totaux au cours de l'intervalle t à $t-1$.

Toutefois, les problèmes de surdénombrement et de sous-dénombrement énoncés dans la partie 2 s'appliquent également à cette méthode d'estimation de la population migrante des bénéficiaires de l'assurance-maladie. Aussi, bien que des estimations infraprovinciales puissent être obtenues avec l'application d'une telle méthode, une fois ramenées à un niveau d'agrégation provincial, elles sont moins fiables que les estimations que le Bureau produit

actuellement en se basant sur les mouvements interprovinciaux des bénéficiaires d'allocations familiales. (Comme Statistique Canada utilise également les fichiers des allocations familiales, les estimations du Bureau sont généralement très proches des estimations produites par l'agence fédérale.)

Pour améliorer les estimations infraprovinciales de la migration et assurer leur compatibilité avec les estimations provinciales, on a introduit un facteur de correction dans les estimations. Il s'agit d'établir le rapport entre les chiffres de migration nette des bénéficiaires de l'assurance-maladie pour une région donnée et les chiffres de migration nette des bénéficiaires de l'assurance-maladie pour la province, et de multiplier ce quotient par la migration nette de la province calculée à partir des estimations trimestrielles de la population produites par le Bureau. L'équation mathématique s'écrit:

$$AMIG_i = \frac{HMIG_i}{HMIG_a} \times PMIG$$

où:

$AMIG_i$ = migration nette corigée dans la région i

$HMIG_i$ = chiffres de migration nette des bénéficiaires de l'assurance-maladie pour la région i

$HMIG_a$ = chiffres de migration nette des bénéficiaires de l'assurance-maladie pour l'Alberta

$PMIG$ = migration provinciale nette estimée à partir des estimations trimestrielles de la population.

Ces estimations corrigées de la migration (AMIG) sont ensuite utilisées pour la production d'estimations de la population.

b) Estimations régionales de la population des petites régions

L'estimation corrigée de la migration nette (AMIG) calculée pour chaque région est introduite dans une équation comprenant les composantes de la croissance démographique (naissances, décès et migration):

$$P_{i_t} = P_{i_{t-1}} + (B_i - D_i) + AMIG_i$$

Où:

P_{i_t} = population estimée dans la région i au moment t

$P_{i_{t-1}}$ = population dans la région i au moment t-1

3.2 Évaluation des estimations pour les petites régions

En utilisant la méthode décrite ci-dessus, le Bureau a produit des estimations de la population pour les quinze divisions de recensement de l'Alberta et vingt-huit municipalités comptant plus de 5,000 habitants. Les résultats sont jusqu'ici très encourageants.

Le tableau 2 présente les résultats d'une comparaison, au niveau des divisions de recensement, entre les chiffres du recensement de 1981 et les estimations relatives à la même année produites, à l'aide de la méthode, à partir des chiffres de la population dénombrée au recensement de 1976. Pour treize des quinze divisions, les estimations se situaient à $+2.0\%$ des chiffres du recensement de 1981. Seules les deux plus petites DR (9 et 14) affichaient pour cette période de cinq ans un écart plus grand que 2.0%. Les écarts absolus moyens (ou écarts annuels moyens) ne dépassaient pas 0.5% dans l'ensemble des divisions du recensement.

Les estimations de la population de vingt-huit municipalités ont été comparées aux chiffres du recensement de 1981 ainsi qu'aux données tirées des recensements municipaux menés entre 1982 et 1984 (voir tableaux 3 et 4). Les comparaisons avec les chiffres du recensement fédéral ont révélé que les estimations de dix-neuf des vingt-huit municipalités présentaient un écart absolu moyen de $\pm 1.0\%$. Seulement six municipalités affichaient un écart annuel supérieur à 2.0%. Les comparaisons avec les recensements municipaux menés entre 1982 et 1984 ont donné vingt-deux cas d'écarts en deçà de $\pm 1.0\%$, quatorze cas d'écarts variant entre $\pm 1.0\%$ et $\pm 3.0\%$ et des écarts supérieurs à $\pm 3.0\%$ dans neuf cas.

Les résultats de l'estimation sont dans l'ensemble satisfaisants et encourageants. La production des chiffres à partir des dossiers d'inscription de l'AHCIP ainsi que la méthode des composantes utilisée ont amélioré la précision des estimations de la population et ont ouvert de nouvelles avenues pour la production d'estimations sur les petites régions géographiques définies par les utilisateurs. Le Bureau continuera de rechercher des façons d'améliorer les totalisations fondées sur les dossiers de l'AHCIP (certaines concernent l'adoption de nouvelles procédures administratives pour l'AHCIP). La méthode d'estimation de la population fera elle-même l'objet de perfectionnements à mesure que le Bureau aura accès à de nouvelles données ou nouvelles techniques.

3.3 Résumé des avantages et désavantages de l'utilisation des dossiers de l'AHCIP

L'utilisation des chiffres sur les bénéficiaires de l'assurance-maladie pour la production d'estimations régionales de la population comporte un certain nombre d'avantages et de désavantages.

Avantages

a) Les données sur les enregistrements de l'AHCIP s'appliquent à l'ensemble des habitants de l'Alberta.
b) Les délais d'inscription semblent aléatoires et n'influent pas sur les distributions ou les tendances observées.
c) Les données sont diffusées rapidement et fréquemment.
d) Le fichier contient des renseignements socio-économiques sur les bénéficiaires (notamment l'âge, le sexe et l'état civil), ce qui permet la production d'estimations détaillées sur la population.

Désavantages

a) Le lieu de résidence déterminé en fonction du code postal peut entraîner certaines inexactitudes.
b) Les personnes inscrites aux dossiers de l'AHCIP peuvent quitter le système, notamment à la suite d'un décès ou d'un départ vers une autre province, sans que les bureaux de l'AHMC en aient été avisés, ce qui entraîne un surdénombrement.
c) Les procédures administratives peuvent être la cause d'écarts/d'inexactitudes dans les comptes des bénéficiaires de l'assurance-maladie de l'Alberta.

Tableau 2

Comparaisons des chiffres du recensement du Canada et des estimations de la population du B.S.A., divisions de recensement de l'Alberta

Division de recensement	Recensement de 1976	Estimations du bureau[a]			
		Accroissement naturel[b] de 1976-1981	Migration nette 1976-1981	Croissance 1976-1981	Population 1981
1	47,000	2,730	6,080	8,810	55,810
2	96,980	6,120	7,190	13,310	110,290
3	32,870	2,310	100	2,410	35,280
4	12,140	490	−520	−30	12,110
5	35,460	1,820	790	2,610	38,070
6	524,570	33,860	107,540	141,400	665,970
7	37,820	2,010	−10	2,000	39,820
8	95,400	6,140	20,860	27,000	122,400
9	19,850	1,040	200	1,240	21,090
10	67,230	1,650	8,550	10,200	77,430
11	632,830	43,880	90,880	134,760	767,590
12	63,130	6,470	16,130	22,600	85,730
13	46,300	2,040	4,320	6,360	52,660
14	19,450	2,200	2,430	4,630	24,080
15	107,010	10,260	10,040	20,300	127,310
Alberta	1,838,040	123,020	274,580	397,600	2,235,640

Division de recensement	Recensement 1981	Écart		Écart absolu moyen
		Nombre	%	
1	55,360	450	0.81	0.16
2	110,470	−180	−0.16	0.03
3	35,640	−360	−1.01	0.20
4	12,120	−10	−0.08	0.02
5	38,430	−360	−0.94	0.19
6	668,680	−2,710	−0.41	0.08
7	40,030	−210	−0.52	0.10
8	123,690	−1,290	−1.04	0.21
9	21,630	−540	−2.50	0.50
10	78,390	−960	−1.22	0.24
11	762,080	5,510	0.72	0.14
12	84,220	1,510	1.79	0.36
13	53,690	−1,030	−1.92	0.38
14	24,650	−570	−2.31	0.46
15	128,640	−1,330	−1.03	0.21
Alberta	2,237,720	−2,080	−0.09	0.02

[a] Données expérimentales.
[b] Accroissement naturel désigne le nombre des naissances moins le nombre de décès.
Note: Les chiffres ayant été arrondis, leur somme peut ne pas correspondre aux totaux.
Source: Statistique Canada, recensements de 1976 et 1981; estimations du Bureau de la statistique de l'Alberta.

Tableau 3

Comparaisons des chiffres du recensement du Canada et des estimations de la population du B.S.A., certaines municipalité de l'Alberta

Municipalité	Recensement de 1976	Estimations du bureau[a]			Recensement de 1981	Écart %	Écart absolu moyen
		Accroissement naturel[b] 1976-1981	Migration nette 1976-1981	Population 1981			
Airdrie	1,410	580	5,090	7,070	8,410	−15.9	3.2
Brooks	6,340	730	2,370	9,440	9,420	0.2	0.0
Calgary	469,920	30,310	93,760	593,990	592,740	0.2	0.0
Camrose	10,100	150	2,570	12,830	12,570	2.1	0.4
Crowsnest Pass	5,250	40	−410	4,880	7,310	−33.2	6.6
Drayton Valley	4,300	530	1,760	6,590	5,040	30.8	6.2
Drumheller	6,150	20	220	6,390	6,510	−1.8	0.4
Edmonton	461,360	27,900	51,240	540,510	532,250	1.6	0.3
Edson	4,040	510	2,490	7,040	5,840	20.5	4.1
Fort McMurray	15,420	2,900	14,140	32,460	31,000	4.7	0.9
Fort Saskatchewan	8,300	800	2,660	11,760	12,170	−3.4	0.7
Grande Prairie	17,630	1,970	6,300	25,900	24,260	6.8	1.4
Hinton	6,730	760	−820	6,670	8,340	−20.0	4.0
Innisfail	2,900	230	1,930	5,060	5,250	−3.6	0.7
Lacombe	3,890	150	1,210	5,240	5,590	−6.3	1.3
Leduc	8,580	920	3,430	12,930	12,470	3.7	0.7
Lethbridge	46,750	2,070	4,400	53,220	54,070	−1.6	0.3
Medicine Hat	32,810	1,770	6,010	40,590	40,380	0.5	0.1
Peace River	4,840	580	970	6,390	5,910	8.1	1.6
Ponoka	4,640	−10	530	5,160	5,220	−1.1	0.2
Red Deer	32,180	2,300	11,790	46,270	46,390	−0.3	0.1
Spruce Grove	6,910	1,110	4,710	12,730	10,330	23.2	4.6
St. Albert	24,130	2,360	6,670	33,160	32,000	3.6	0.7
Stettler	4,180	500	580	5,270	5,140	2.5	0.5
Taber	5,300	320	410	6,020	5,990	0.5	0.1
Vegreville	4,160	80	860	5,090	5,250	−3.0	0.6
Wetaskiwin	6,750	300	2,440	9,490	9,600	−1.1	0.2
Whitehorse	3,880	600	1,150	5,630	5,590	0.7	0.1
Alberta	1,838,040	123,020	274,580	2,235,630	2,237,720	−0.1	0.0

[a] Données expérimentales.
[b] Accroissement naturel désigne le nombre des naissances moins le nombre de décès.
Note: Les chiffres ayant été arrondis, leur somme peut ne pas correspondre aux totaux.
Source: Statistique Canada, recensements de 1976 et 1981; estimations du Bureau de la statistique de l'Alberta.

Tableau 4

Comparaisons des chiffres des recensements municipaux de l'Alberta et des estimations du B.S.A., certaines municipalités

Municipalité	1982			1983			1984		
	Estimations du Bureau[a]	Recensement municipal	Écart %	Estitions du Bureau[a]	Recensement municipal	Écart %	Estitions du Bureau[a]	Recensement municipal	Écart %
Airdrie	9,450	9,980	−5.3	9,830	10,430	−5.8	10,080	--	--
Brooks	9,640	--	--	9,790	--	--	9,510	--	--
Calgary	614,930	623,130	−1.3	622,510	620,690	0.3	615,140	619,810	−0.8
Camrose	12,880	12,810	0.6	12,970	--	--	13,070	12,750	2.5
Crowsnest Pass	7,490	7,580	−1.1	7,530	--	--	7,350	--	--
Drayton Valley	5,120	4,870	5.2	5,200	--	--	5,310	4,920	7.9
Drumheller	6,660	--	--	6,700	6,670	0.4	6,620	--	--
Edmonton[b]	550,930	551,310	−0.1	557,400	560,090	−0.5	551,140	--	--
Edson[b]	6,110	6,290	−2.9	6,220	--	--	6,080	7,110	−14.5
Fort McMurray	32,930	33,580	−1.9	33,600	34,490	−2.6	35,150	35,350	−0.6
Fort Saskatchewan	12,530	12,460	0.6	12,650	12,470	1.4	12,620	--	--
Grande Prairie	24,650	--	--	24,910	24,080	3.5	25,370	24,410	3.9
Hinton	8,820	8,820	0.0	8,980	8,830	1.8	8,950	8,900	0.6
Innisfail	5,420	5,440	−0.4	5,460	--	--	5,440	5,440	0.0
Lacombe	5,810	5,720	1.5	5,850	5,850	5,950	−1.8	5,850	--
Leduc	12,880	--	--	13,010	--	--	13,290	--	--
Lethbridge[b]	55,440	56,500	−1.9	55,900	58,090	−3.8	57,500	--	--
Medicine Hat[b]	41,070	--	--	41,440	42,270	0.7	41,540	--	--
Peace River	6,080	--	--	6,150	--	--	6,250	--	--
Ponoka	5,310	--	--	5,310	--	--	5,280	--	--
Red Deer	48,450	48,560	−0.2	49,230	50,260	−2.0	50,860	51,070	−0.4
Spruce Grove	11,080	10,780	2.7	11,410	11,310	0.9	11,550	11,570	−0.1
St. Albert	33,170	32,980	0.6	33,740	35,030	−3.7	34,840	35,530	−1.9
Stettler	5,180	--	--	5,220	--	--	5,300	--	--
Taber	6,140	--	--	6,210	--	--	6,360	6,380	−0.4
Vegreville	5,280	5,250	0.6	5,290	--	--	5,390	--	--
Wetaskiwin	9,880	9,900	−0.2	9,990	10,020	−0.3	10,080	--	--
Whitecourt	5,710	--	--	5,840	--	--	5,710	--	--

[a] Données expérimentales.
[b] Une fusion a eu lieu entre 1982 et 1984.
Note: "--" indique qu'il n'y a pas eu de recensement municipal.
Source: Municipal Affairs d'Alberta, recensements municipaux 1982-1984; estimations du Bureau de la statistique de l'Alberta.

4. CONCLUSION

Notre expérience de l'exploitation des dossiers de l'assurance-maladie s'est révélée très positive. Les principaux avantages sont l'utilisation des chiffres dans un modèle des composantes appliqué à la production d'estimations infraprovinciales de la population ainsi que l'obtention d'excellents ratios de distribution selon l'âge et le sexe et des tendances uniformes. Les coûts liés au développement du système d'enregistrement des données démographiques ne sont pas jugés excessifs compte tenu de ces avantages. Le Bureau est à la disposition de tout organisme provincial qui envisage de tirer ainsi profit des fichiers de l'assurance-maladie et qui aimerait traiter de la question plus en détail et obtenir des renseignements supplémentaires comme les clichés d'enregistrement et les coûts de traitement des systèmes.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 223–227
Statistique Canada

Utilisation des comptes de l'Hydro dans le modèle d'estimation par régression de la population en Colombie-Britannique[1]

DONALD G. McRAE[2]

RÉSUMÉ

La précision des estimations régionales de la population qui sont établies à l'aide d'un modèle de régression repose foncièrement sur la capacité des indicateurs sélectionnés de traduire avec exactitude les changements démographiques. Par conséquent, il est important de connaître préalablement les caractéristiques des données administratives pouvant servir d'indicateurs de la population dans un modèle de régression. Ce document présente sommairement les avantages et les inconvénients de l'utilisation des comptes résidentiels de l'Hydro dans le modèle d'estimation par régression de la population de la Colombie-Britannique.

MOTS CLÉS: Estimations régionales de la population; méthode de régression; méthode de corrélation des différences; indicateurs de population; comptes de l'Hydro; prestataires d'allocations familiales.

1. INTRODUCTION

Le Bureau central de la statistique produit des estimations postcensitaires de la population pour diverses unités géographiques de la Colombie-Britannique (par ex. les municipalités, les districts sanitaires locaux, les divisions de recensement, les régions de la GRC, et autres). Les estimations courantes de la population sont produites pour ces zones infraprovinciales à l'aide d'une méthode de régression, la méthode de corrélation des différences (MCD). Une description détaillée de cette méthode a été présentée dans des documents antérieurs (Bureau central de la statistique, 1982, McRae 1985). Les données utilisées comme indicateurs de la population sont le nombre de prestataires d'allocations familiales (F), et(ou) le nombre de comptes résidentiels de l'Hydro (H). Les caractéristiques de cette deuxième source de données, les comptes résidentiels de l'Hydro, tels qu'ils sont utilisés dans le modèle de la C.-B., sont examinées dans le reste de cet exposé.

2. SOURCES DE DONNÉES

Les données des comptes résidentiels de l'Hydro à l'intérieur de la province proviennent de neuf administrations et sociétés différentes :

Organisation	% de l'ensemble des comptes de l'Hydro (1985)
(1) British Columbia Hydro	90.8
(2) West Kootenay Power and Light	4.7
(3) Princeton Light and Power Co.	0.2
(4) Ville de Kelowna	0.8
(5) Ville de Penticton	0.8
(6) District de Summerland	0.3
(7) Ville de New Westminster	1.7
(8) Ville de Grand Forks	0.1
(9) Ville de Nelson	0.6

[1] Exposé présenté à la rencontre du Comité fédéral-provincial de la démographie, les 28 et 29 novembre 1985.

[2] D.G. McRae, Bureau central de la statistique, ministère du développement de l'industrie et de la petite entreprise de la Colombie-Britannique, 2e étage, 1405, Douglas Street, Victoria, C.-B., Canada, V8W 3C1.
Les opinions exprimées dans ce document sont celles de l'auteur et ne représentent pas nécessairement le point de vue du gouvernement de la Colombie-Britannique.

Les principaux fournisseurs d'électricité sont la B.C. Hydro et la West Kootenay Power and Light. Les autres administrations achètent l'électricité aux deux grands fournisseurs et la revendent à leurs propres clients (normalement les résidents de la municipalité).

3. PRÉSENTATION DES DONNÉES

Sur les neuf sources de données des comptes résidentiels de l'Hydro, seulement celles qui proviennent de la B.C. Hydro sont ordinolingues, c'est-à-dire lisibles par machine. Les huit autres administrations fournissent les données totalisées par municipalité (urbaine), et un total pour les clients ruraux (non municipaux) s'il y a lieu. La date de référence pour toutes les données est la date de facturation du 31 mai. Dans la plupart des cas, les données peuvent être obtenues dans les 2 à 3 semaines qui suivent la date de facturation.

Les données fournies par la B.C. Hydro se présentent sous deux formes. La première indique le nombre de compteurs résidentiels en date du 31 mai par code de district capital. Un code de district capital (il en existe environ 248 dans la province) est une unité administrative utilisée par la B.C. Hydro, qui correspond à une municipalité là où des municipalités existent. Aux termes d'une entente, les deux principaux fournisseurs d'électricité dans la province paient à chaque municipalité un certain pourcentage des recettes annuelles perçues auprès des clients dans cette municipalité, ce qui tient lieu d'impôt foncier. Par conséquent, ces sociétés, comme la B.C. Hydro, établissent leur système de comptabilité en fonction des clients résidant dans les municipalités visées. En outre, la B.C. Hydro cherche à maintenir une correspondance étroite entre les limites des districts capitaux et celles des districts scolaires.

La deuxième forme de présentation indique pour chacun des compteurs résidentiels, qui sont au nombre d'un million et plus, le code postal de l'adresse de facturation. Ce deuxième fichier de données permet de convertir facilement les comptes de l'Hydro en unités géographiques autres que les municipalités et les districts scolaires, au moyen du code postal.

4. AVANTAGES DES DONNÉES DE L'HYDRO DANS UN MODÈLE DE RÉGRESSION

Des tests empiriques portant sur les deux sources de données différentes, les comptes de l'Hydro (H) et les comptes d'allocations familiales (F), ont été effectués par la production d'estimations de la population pour 1981, avec chacune prise séparément et avec les deux ensemble. Les coefficients de régression utilisés étaient fondés sur les périodes 1971/1976 et 1976/1981, et l'année de base était 1976. Les résultats ont été comparés aux données du recensement de 1981, et les erreurs absolues moyennes en pourcentage (EAMP) ont été calculées (voir tableaux 1 et 2).

Comme on peut voir au tableau 1, les estimations de la population fondées sur les données de l'Hydro produisent, en moyenne, des pourcentages d'erreur inférieurs à ceux des estimations fondées sur les comptes d'allocations familiales. Si on examine de plus près le tableau 1, on constate que l'amélioration de la précision des estimations touche en presque totalité les estimations relatives aux régions d'une population inférieure à 4,000 habitants. Cette observation est confirmée par le tableau 2 où l'on peut voir que, en termes statistiques, il existe une différence importante sur le plan de la précision entre les estimations fondées sur les comptes de l'Hydro et celles fondées sur les comptes d'allocations familiales pour les régions de petite taille (c.-à-d. moins de 4,000 habitants).

On peut évaluer l'effet marginal de l'adjonction d'un autre indicateur de population à la méthode de corrélation des différences en examinant la variation de la précision de l'estimation selon qu'on utilise ou non cet indicateur. D'après les tableaux 1 et 2, il semblerait que l'adjonction des données de l'Hydro améliore du point de vue statistique la précision de

Tableau 1

Comparaison des erreurs d'estimation entre les sources de données pour les municipalités de la Colombie-Britannique - 1981

Source de données	EAMP[3] Ensemble	n	Population ≥ 4000 EAMP	n	Population < 4000 EAMP	n
MCD/H/F	5.53	158	2.99	88	8.72	70
MCD/H	5.16	158	4.04	88	6.58	70
MCD/F	10.46	167	4.57	92	17.69	75

$$\text{EAMP} = \left[\sum_{i=1}^{n} \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right| \right] \div n \times 100$$

où:

Y_i = population du recensement pour la région i
$\hat{Y}_i$ = population estimé pour la région i
n = nombre de régions estimées.

Tableau 2

Test pour déterminer les différences statistiquement significatives entre les erreurs absolues moyennes en pourcentage, selon certaines sources de données - 1981

Source de données	Intervalle de confiance de 95% pour la différence moyenne dans les erreurs absolues en pourcentage: Ensemble	Population ≥ 4000	Population < 4000
MCD/H/F – MCD/H	.37 ± .86	−1.05 ± .56[a]	2.14 ± 1.76[a]
MCD/H/F – MCD/F[b]	−4.86 ± 1.55[a]	−1.57 ± .85[a]	−9.00 ± 3.11[a]

[a] Différences statistiquement significatives au niveau de 5% calculées à l'aide d'un test-t bilatéral et d'échantillons appariés, en supposant une distribution normale des moyennes.

[b] Pour apparier les échantillons, sur 167 estimations possibles fondées sur les comptes d'allocations familiales, 158 seulement ont été utilisées. Le nombre des observations est le suivant: ensemble, 158; plus grand que ou égal à 4,000, 88; moins de 4,000, 70.

l'estimation à la fois dans les grandes et les petites régions. Les données des comptes d'allocations familiales, par contre, améliorent la précision pour les régions de grande taille, mais la réduit pour les régions de petite taille, sans produire un effet global statistiquement significatif.

5. INCONVÉNIENTS DES DONNÉES DE L'HYDRO DANS UN MODÈLE DE RÉGRESSION

Un des problèmes qui se posent lorsqu'on utilise les données de l'Hydro dans un modèle d'estimation de la population par régression concerne les logements vacants, ou plus précisément les écarts importants des taux de logements vacants entre l'année de base et l'année d'estimation. Cette lacune a été démontrée dans l'évaluation de 1981 portant sur les localités de la région Peace River-Liard en Colombie-Britannique. Par suite du projet de North-East Coal, les localités de la division de recensement Peace River-Liard en 1981 ont connu une

explosion de la construction lorsque les promoteurs ont construit des logements en prévision d'un afflux de population. Chaque logement, occupé ou non, ou même en construction, devait être muni d'un compteur qui était peut-être peu utilisé, mais qui n'en demeurait pas moins actif, et par le fait même compris dans les calculs. Par conséquent, la variation de la proportion de compteurs par rapport à la population entre 1976 et 1981 a été surévaluée, ce qui a entraîné une surestimation de la population de 1981 pour de nombreuses localités de Peace River-Liard.

Un autre inconvénient de l'utilisation des données de l'Hydro découle de la possibilité d'une conversion, dans le cas des immeubles d'habitation, d'un compteur unique en un réseau de compteurs multiples. Cette situation peut se présenter lorsqu'un ancien immeuble d'appartements, par exemple, qui possédait un seul compteur, est rénové ou transformé en logements à compteurs individuels. Ce changement entraînerait une surestimation de la population dans un modèle de régression, s'il devait survenir à un moment quelconque entre l'année de base et l'année d'estimation.

Enfin, certains problèmes apparaîtront si on utilise les données de l'Hydro pour des régions en mutation, c'est-à-dire où le rapport entre la population et les compteurs résidentiels se modifie. On peut citer comme exemple en Colombie-Britannique le lieu de villégiature de Whistler. Il y a quinze ans cette municipalité était constituée essentiellement de chalets d'hiver construits aux abords d'une pente de ski. Au cours de la dernière décennie, cependant, elle s'est transformée en un lieu résidentiel à longueur d'année. Par conséquent, le nombre de personnes par compteur de l'Hydro, qui à l'origine était très faible par rapport à la moyenne de la Colombie-Britannique, se rapproche maintenant de la norme. Comme pour le cas des logements vacants, l'utilisation des données de l'Hydro pour estimer la population dans un territoire de ce genre donnerait probablement lieu à des erreurs supérieures à la moyenne.

La solution aux trois problèmes mentionnés ci-haut est la suppression des comptes correspondant à une utilisation mensuelle ou bimensuelle faible qui par le fait même représentent des logements considérés vacants. Le Bureau central de la statistique étudie actuellement la possibilité d'appliquer cette mesure dans le cas des données provenant de la B.C. Hydro. Si c'est possible, nous espérons disposer de l'ensemble de données améliorées pour la comparaison avec les résultats du recensement de 1986. Pour l'instant, à titre de solution partielle, les données de l'Hydro pour les secteurs qui en 1981 affichaient un rapport faible ou élevé de personnes par compteur comparativement à la norme provinciale (c.-à-d. moins de 2 ou plus de 5) ne sont pas utilisées.

Une dernière lacune possible est le recours à des sources d'information externes et différentes. Dans le passé, cette situation n'a généralement pas posé de problème. Toutefois, il y a eu quelques rares cas qui ont suscité une remise en question de la qualité des données des compteurs recueillies sur le terrain. Il pourrait s'agir d'une modification des limites d'une municipalité non reflétée dans les données relatives aux compteurs ou de l'adjonction aux données d'un certain type de comptes non résidentiels (par exemple pour les lampadaires). Il importe donc de vérifier soigneusement les données.

6. CONCLUSIONS

Les avantages et les inconvénients de l'utilisation des données de l'Hydro dans le modèle d'estimation par régression de la population en Colombie-Britannique sont les suivants:

Avantages

a) Les données de l'Hydro, lorsqu'elles sont utilisées dans un modèle de régression, produisent une erreur absolue moyenne en pourcentage plus faible que les données sur les allocations familiales, pour les petites régions.

b) Les données sont obtenues de chaque fournisseur sous une forme déjà agrégée par municipalité. Le principal avantage est que les modifications des limites municipales, qui se produisent régulièrement, paraissent dans les données sans entraîner de travail supplémentaire pour le Bureau.
c) La majorité des données peuvent être obtenues sous forme ordinolingue et accompagnées du code postal. Cela permet de les convertir facilement en fonction de régions géographiques autres que les municipalités lorsqu'un triage est effectué par le Fichier principal de conversion des codes postaux du Bureau.
d) Les données peuvent être obtenues gratuitement de chacun des fournisseurs dans un délai relativement court (2 à 3 semaines).

Inconvénients

a) Les écarts dans les taux de logements vacants entre l'année de base et l'année d'estimation produisent une distorsion des estimations.
b) Les immeubles d'habitation qui passent d'un simple compteur à des compteurs multiples parfois entre l'année de base et l'année d'estimation produisent une surestimation.
c) Les régions en mutation, c'est-à-dire qui passent par exemple d'une population saisonnière à une population "stable", introduisent un biais dans les estimations.
d) Les données proviennent de sources externes et différentes, ce qui pourrait causer éventuellement des problèmes sur le plan de la qualité et de la comparabilité des données, et produire également une situation où l'ordre de priorité du programme des estimations de la population du Bureau serait subordonné aux besoins administratifs d'un organisme externe.

BIBLIOGRAPHIE

BUREAU CENTRAL DE LA STATISTIQUE (1982). British Columbia municipal population estimation methodology. Document non publié, Bureau central de la statistique, ministère du Développement de l'industrie et de la petite entreprise, gouvernement de la Colombie-Britannique.

McRAE, D. (1985). A regression approach to small area population estimation. Communication présentée au colloque international sur les statistiques régionales, Ottawa, Canada, 22-24 mai 1985.

McRAE, D. (1982). British Columbia small area estimation model – 1981 municipal and census division evaluation. Document non publié, Bureau central de la statistique, ministère du Développement de l'industrie et de la petite entreprise, gouvernement de la Colombie-Britannique.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 229–236
Statistique Canada

Estimation de la répartition par âge et par sexe de la population totale des petites régions[1]

DAVID S. O'NEIL et CHRIS D. McINTOSH[2]

RÉSUMÉ

Ce document décrit une méthode de production d'estimations par âge et par sexe de l'état de la population des petites régions, à partir d'estimations de la population totale, de données sur les naissances et les décès et d'estimations chronologiques de la migration nette résiduelle. On y présente également une évaluation fondée sur les chiffres du recensement de 1981 concernant les divisions de recensement et les districts scolaires de la Colombie-Britannique.

MOTS CLÉS: Estimations de la population par âge et par sexe; petites régions; migration nette résiduelle.

1. INTRODUCTION

Le Bureau central de la statistique produit actuellement des estimations postcensitaires de la population pour diverses régions infraprovinciales à l'aide d'une méthode de régression (Bureau central de la statistique, 1982). En plus des estimations de la population totale par petite région, il produit des estimations par âge et par sexe.

Ce document expose la méthode permettant à partir de l'estimation de la population totale, d'obtenir des estimations par âge et par sexe pour les régions infraprovinciales de la Colombie-Britannique.

2. VUE D'ENSEMBLE

La méthode retenue pour déterminer la population des petites régions par sexe et par année d'âge comporte deux volets.

Le premier volet consiste à examiner les données chronologiques sur la migration nette résiduelle à partir des chiffres de recensement afin d'établir un certain nombre de répartitions de la migration par sexe et par année d'âge pour chaque petite région (Shryock et Siegal 1980).

Le second volet de la méthode consiste à faire vieillir la population de base pour chaque sexe, à ajouter les naissances et à retrancher les décès, ce qui donne pour chaque région une nouvelle répartition de la population que l'on désigne la population de base naturelle. En retranchant la population de base naturelle de la population totale estimée selon le sexe, on obtient un résidu qui est égal à la migration nette selon le sexe si la population dénombrée et les données de l'état civil sont exacts pour les deux périodes. Ce résidu par sexe est distribué par année d'âge au moyen d'une répartition chronologique, puis il est ajouté à la population de base naturelle, ce qui donne l'estimation de la population par âge et par sexe d'une région pour la période suivante.

À cause des courts délais de production des données d'entrée, les estimations de la population totale peuvent être calculées quatre mois après la date de référence (1er juin) et les répartitions par âge et par sexe, un ou deux mois plus tard.

---

[1] Version abrégée du document présenté à la réunion du Comité fédéral-provincial sur la démographie, Ottawa, novembre 28-29, 1985.

[2] D.S. O'Neil, SRL Sociometrics Resources Ltd., et C.M. McIntosh, Intersoft Resources Ltd., Bureau central de la statistique, ministère du développement de l'industrie et de la petite entreprise de la Colombie-Britannique, 2e étage, 1405, Douglas Street, Victoria, Colombie-Britannique, Canada, V8W 3C1.
Les opinions exprimées dans ce document sont celles des auteurs et ne représentent pas nécessairement le point de vue du gouvernement de la Colombie-Britannique.

3. RÉPARTITION CHRONOLOGIQUES DE LA MIGRATION NETTE

Pour chacun des 74 districts scolaires de la Colombie-Britannique, les estimations de la migration nette par âge et par sexe ont été calculées, par la méthode des résidus, pour les périodes de 1961-1966, 1966-1971 et 1971-1976. C'est ce qu'on appelle les répartitions chronologiques par petite région.

Si l'on examine ces répartitions de la migration nette par petite région, on découvre qu'elles sont extrêmement instables dans le temps. Pour réduire les effets de cette instabilité, un certain nombre de mesures ont été prises.

Premièrement, les répartitions de la migration par petite région ont été classées selon qu'elles ont eu lieu au cours d'une période de migration totale nette positive ou d'une période de migration totale nette négative. On a découvert que les répartitions par âge de la migration résiduelle pour de nombreuses régions diffèrent selon que la migration est positive ou négative.

On a ensuite regroupé les petites régions présentant des répartitions de migration semblables, pour ensuite calculer les répartitions de la migration nette positive et de la migration nette négative. Elles ont été désignées répartitions chronologiques par groupe. L'analyse de certains groupes d'âge (selon la méthode SPSS/PC) a permis de regrouper les répartitions chronologiques de la migration des petites régions. Cette méthode a permis de grouper la plupart des régions en trois grappes, et huit groupes étaient formés uniquement d'une région chacun.

Une fois les régions ainsi regroupées, on a déterminé les répartitions de la migration positive et les répartitions de la migration négative pour les périodes de migration positive et négative les plus récentes.

4. ESTIMATIONS DE LA POPULATION DES PETITES RÉGIONS SELON LE SEXE ET L'ANNÉE D'ÂGE

Comme il a été mentionné dans la section 3, les répartitions chronologiques de la migration nette calculées par la méthode des résidus varient considérablement pour certaines régions. Il semble que cela soit imputable à deux facteurs. Premièrement, de nombreuses régions à l'étude ont des économies de ressources déficientes qui affichent de grandes fluctuations, ce qui entraîne des mouvements des taux de migration. Deuxièmement, le calcul d'une distribution en percentiles d'un paramètre tel que la migration nette ayant des valeurs positives, négatives ou nulles introduit un certain degré d'instabilité.

Pour prévenir la construction d'une répartition chronologique de la migration nette qui ne soit pas représentative de la situation courante dans l'année d'estimation, on a calculé cinq répartitions chronologiques, par sexe, qui ont été ventilées par année d'âge. Une description de ces répartitions est donnée ci-après.

1) La première répartition de la migration qui a été choisie est la répartition chronologique par petite région de chaque petite région dont le signe est le même que la migration nette vers cette région.
2) La deuxième répartition est la répartition chronologique par groupe du groupe auquel la petite région appartient et dont le signe est le même que celui de la migration nette vers cette région.
3) On a obtenu la troisième répartition en additionnant *séparément* la migration, pour la période la plus récente, de toutes les petites régions possédant une migration nette positive et une migration nette négative, puis en calculant la répartition par âge.
4) La quatrième répartition est celle de la population de base naturelle pour chaque petite région.
5) La cinquième et dernière répartition est la répartition par âge des migrants vers la Colombie-Britannique en général. Comme la migration vers la Colombie-Britannique a été positive au cours de toutes les années visées, cette répartition est positive. On l'a néanmoins utilisée, peu importe que la migration vers une petite région ait été positive ou négative.

Il n'est pas toujours possible de calculer toutes les cinq répartitions. Ainsi, on ne peut le faire si une petite région n'a jamais eu de migration nette négative dans le passé mais si elle en a une pour l'année visée. Dans un tel cas, on se sert uniquement des répartitions que l'on peut calculer.

Des essais basés sur les données du recensement de 1981 ont montré que, des cinq répartitions de la migration nette décrites plus haut, la première (la répartition chronologique par petite région) a l'erreur absolue moyenne en pourcentage la plus faible pour tous les districts scolaires et tous les groupes d'âge; elle est suivie de la deuxième (répartition chronologique par groupe), puis de la troisième, etc. Cependant, bien que la première répartition ait l'EAMP la plus faible, elle n'a pas produit le taux d'erreur le plus bas dans chaque cas. Aussi, une technique de sélection a été appliquée pour remplacer la répartition de la population produite au numéro 1 par celle produite aux numéros 2, 3, 4 ou 5 uniquement dans les cas où la répartition de la population produite au numéro 1 était considérée comme n'étant pas fidèle à la répartition de la population pour l'année choisie. La technique de sélection suivante a été conçue à partir des résultats des tests lasée sur le recensement de 1981.

Une fois qu'on a calculé toutes les répartitions possibles, on ajoute chacune à la population de base naturelle et on obtient jusqu'à cinq estimations possibles de la population d'une petite région, selon le sexe et l'année d'âge, pour la période suivante. On examine alors ces estimations de la population par âge et par sexe afin de déterminer celle qui produit le moins de changements dans la structure par âge de la petite région par rapport à l'année précédente. Pour ce faire, on calcule d'abord la différence moyenne en pourcentage entre les structures par âge de chacune des cinq populations possibles au temps $t+1$ et celles de la population au temps t. On calcule ensuite l'écart-type des moyennes obtenues et on indique au moyen d'un code la répartition qui affiche l'écart-type le plus bas. Si l'écart-type produit à partir de la répartition chronologique par petite région est beaucoup plus grand que l'écart-type minimum (c.-à-d. la répartition signalée), cette répartition chronologique est rejetée. On applique le même procédé à la répartition chronologique par groupe, et ainsi de suite jusqu'à ce qu'une des cinq populations possibles soit retenue.

Une fois la population optimale au temps $t+1$ calculée pour toutes les petites régions, il reste à faire deux derniers rajustements. En premier lieu, on a remplacé les chiffres relatifs aux groupes d'âge 0-14 ans par ceux provenant des allocations familiales, puis ajusté proportionnellement les populations des autres groupes d'âge afin que la population totale de chaque petite région reste constante. En second lieu, on a ajusté proportionnellement la population pour s'assurer que la répartition par âge de la somme de toutes les estimations relatives aux populations des petites régions corresponde à la répartition par âge établie par Statistique Canada pour la Colombie-Britannique.

5. ÉVALUATION DE LA MÉTHODE

Les tableaux présentés plus loin indiquent les taux d'erreur relatifs aux estimations de la population au 1er juin 1981 et répartis en groupes d'âge quinquennaux de 0 à 70 ans et plus, pour 74 districts scolaires de la Colombie-Britannique et 29 divisions de recensement. On a obtenu les estimations de la population par âge et par sexe pour les divisions de recensement en regroupant les estimations de la population pour les districts scolaires.

Pour établir l'exactitude des estimations de la population des petites régions par âge et par sexe obtenues par la méthode décrite plus haut, on a produit des estimations de la population de 1981 par sexe et par groupe d'âge quinquennal de 0 à 70 ans et plus pour 74 districts scolaires, puis on a comparé les résultats obtenus avec les chiffres du recensement de 1981. On a eu recours à deux mesures sommaires pour évaluer l'efficacité des estimations de la population par âge et par sexe. Il s'agit de l'erreur absolue moyenne en pourcentage (EAMP) et l'indice de mauvaise distribution (IMD). L'EAMP est défini comme suit:

$$EAMP = 100 \times \left[\sum_{i=1}^{N} \left| (P_{Ei} - P_{Ai})/P_{Ai} \right| \right] / N$$

où P_{Ei} est la population estimée pour le groupe d'âge i, P_{Ai} est la population d'après le recensement pour le groupe d'âge i et N est le nombre de cellules. L'IMD est défini comme suit:

$$IMD = 100 \times \frac{1}{2}\left[\sum_{i=1}^{N}(\left|P_{Ai} - P_{Ei}\right|)\right] / \sum_{i=1}^{N} P_{Ai}$$

où P_{Ai} est la population réelle pour le groupe d'âge i et P_{Ei} est la population estimée pour le groupe d'âge i.

Comme l'indique le tableau 1, l'erreur absolue moyenne en pourcentage pour tous les groupes d'âge et toutes les régions est de 6.20% et l'IMD, de 1.95%. L'erreur moyenne en pourcentage pour les hommes est très semblable à celle pour les femmes (EAMP de 7% dans les deux cas et IMD de 2.15% pour les hommes et de 2.08% pour les femmes).

Du point de vue de l'âge, les taux d'erreur les plus élevés figurent dans les groupes des 20-29 ans et des 60-69 ans. Il faut également noter que la répartition par âge des erreurs est différente chez les hommes et chez les femmes. Chez les hommes, les taux d'erreur les plus élevés semblent se trouver dans les groupes d'âge supérieurs, tandis que chez les femmes ils sont dans les groupes très mobiles des 20-29 ans.

Tableau 1

Pourcentage d'erreurs entre les estimations et les chiffres du recensement de 1981, selon le groupe d'âge, pour tous les districts scolaires

Âge	Total		Hommes		Femmes	
	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)
0-4	3.33	.96	3.94	1.21	3.62	1.04
5-9	2.80	.76	3.28	0.88	3.62	1.02
10-14	2.33	.64	3.54	0.84	2.88	0.87
15-19	5.20	2.01	5.68	2.01	6.18	2.24
20-24	13.32	4.77	13.50	4.62	14.54	5.12
25-29	8.31	4.07	8.42	3.70	9.41	4.65
30-34	5.02	2.12	5.42	2.45	5.72	2.06
35-39	4.88	1.33	5.73	1.62	5.38	1.34
40-44	4.52	1.33	5.84	1.51	4.67	1.52
45-49	3.60	1.22	4.47	1.37	4.78	1.49
50-54	5.66	1.33	5.86	1.48	6.68	1.54
55-59	6.11	1.72	6.19	1.78	7.82	1.97
60-64	8.86	2.44	10.35	2.95	8.91	2.17
65-69	10.60	2.66	12.53	3.52	11.44	2.30
70+	8.49	1.95	10.19	2.35	9.33	1.94
Moyenne	6.20	1.95	7.00	2.15	7.00	2.08

Comme l'indique le tableau 2, les régions peu peuplées ont en général un pourcentage d'erreur plus élevé. Le pourcentage élevé d'erreurs dans les petites régions peut être attribué à l'instabilité des économies de ressources, laquelle se reflète dans la répartition de la migration nette.

Au niveau des divisions de recensement, on constate qu'il se dégage une tendance semblable. Comme le montre le tableau 3, l'erreur absolue moyenne en pourcentage pour toutes les régions et tous les groupes d'âge est de 4.83%, (5.19% pour les hommes et 5.60% pour les femmes). L'IMD est de 1.27% au total, de 1.41% pour les hommes et de 1.35% pour les femmes. L'erreur est encore une fois bimodale, atteignant des sommets dans les groupes des 20-29 ans et des 60-69 ans. En outre, les femmes affichent des taux d'erreur plus élevés que les hommes dans les groupes des 20-29 ans, et des taux d'erreur plus bas dans les groupes des 60-69 ans.

Tableau 2

Pourcentage d'erreurs pour tous les districts scolaires selon la taille de la population

Taille de population	Hommes		Femmes		Total	
	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)
0-9,999	8.87	3.16	10.14	3.89	10.27	3.65
10,000-24,999	6.07	2.47	6.92	2.96	6.62	2.58
25,000+	3.66	1.67	3.92	1.78	4.09	1.78
Moyenne pour les districts scolaires	6.20	1.95	7.00	2.15	7.00	2.08

Tableau 3

Pourcentage d'erreurs entre les estimations et les chiffres du recensement de 1981, selon le groupe d'âge, pour toutes les divisions de recensement

Groupe d'âge	Total		Hommes		Femmes	
	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)
0-4	2.37	0.54	3.20	0.76	2.28	0.58
5-9	1.52	0.50	1.71	0.55	2.13	0.68
10-14	1.69	0.39	2.75	0.57	2.50	0.60
15-19	3.81	1.39	3.79	1.30	4.68	1.63
20-24	9.83	3.07	9.30	2.91	10.90	3.41
25-29	7.02	3.04	7.30	2.87	8.09	3.37
30-34	3.28	1.29	3.31	1.43	3.85	1.25
35-39	3.34	0.66	3.06	0.57	4.21	0.88
40-44	3.86	0.88	4.29	1.01	4.16	0.90
45-49	2.91	0.70	3.20	0.75	3.75	0.83
50-54	4.82	0.64	4.41	0.75	6.10	0.86
55-59	5.49	1.34	5.36	1.55	6.94	1.30
60-64	7.88	1.95	8.37	2.29	7.94	1.74
65-69	8.48	1.89	10.30	2.67	9.79	1.43
70+	6.16	0.81	7.46	1.20	6.73	0.71
Moyenne	4.83	1.27	5.19	1.41	5.60	1.35

Le tableau 4 (pourcentage d'erreurs dans les divisions de recensement selon la taille de la population) indique qu'à la suite de l'agrégation au niveau de régions infraprovinciales plus grandes, on obtient une amélioration des taux d'erreur. Le tableau 7 montre qu'il existe une relation négative entre les taux d'erreur et la taille de la population, à l'échelon de la division de recensement.

La comparaison des tableaux 5 et 6 confirme l'amélioration des taux d'erreurs lorsqu'on agrège en fonction de cellules par âge et sexe plus grandes. Bien qu'il faille être prudent dans l'utilisation d'estimations régionales par âge et sexe, les auteurs estiment qu'elles sont fiables.

Tableau 4

Pourcentage d'erreurs pour toutes les divisions de recensement selon la taille de la population

Taille de population	Total		Hommes		Femmes		N
	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)	
0-39,000	7.22	1.94	7.55	2.13	8.79	2.29	10
40,000-59,999	4.32	1.82	5.03	2.14	4.91	1.83	10
60,000+	2.51	0.87	2.73	0.98	2.84	0.90	9
Moyenne des divisions de recensement	4.83	1.27	5.19	1.41	5.60	1.35	29

Tableau 5

Districts scolaires
Nombre d'estimations par intervalle d'erreur

	Intervalle d'erreur absolue moyenne en pourcentage				
	< 5	5 à 10	10 à 15	15+	Total
N° de cellules	674	239	101	96	1110
Pourcentage	61%	22%	9%	9%	100%

Tableau 6

Divisions de recensement
Nombre d'estimations par intervalle d'erreur

	Intervalle d'erreur absolue moyenne en pourcentage				
	< 5	5 à 10	10 à 15	15+	Total
N° de cellues	306	77	25	27	435
Pourcentage	70%	18%	6%	6%	100%

6. CONCLUSIONS

La méthode décrite dans ce document présente des avantages dans le cas des régions pour lesquelles il existe de bonnes sources de données chronologiques sur la population et de statistiques de l'état civil. On estime que l'application d'une méthode comportant des estimations de la migration nette est relativement simple, qu'elle produit des taux d'erreur acceptables et permet la production d'estimations par âge et sexe rapidement après la date de référence. Certes, il serait idéal d'établir des estimations de la migration interne et externe, mais on ne dispose pas actuellement d'assez d'information sur les mouvements migratoires dans les petites régions de la Colombie-Britannique. Une autre amélioration qui est envisagée a trait à l'utilisation des chiffres relatifs à la sécurité de la vieillesse pour accroître la stabilité et la précision des estimations concernant les groupes d'âge supérieurs.

REMERCIEMENTS

Les auteurs tiennent à remercier Don McRae, Steve Miller, Ravi Verma, Garnett Picot et Paul Knapp ainsi que tous leurs prédécesseurs pour leur apport et leur soutien à la construction du système de ventilation des estimations.

Tableau 7

Pourcentage d'erreurs entre les estimations et les chiffres du recensement de 1981, selon la division de recensement, pour tous les groupes d'âge

Division du recensement	Population totale	Total		Hommes		Femmes	
		EAMP (%)	IMD (%)	EAMP (%)	IMD (%)	EAMP (%)	IMD (%)
1000 East Kootenay	53,725	4.24	2.04	5.24	2.29	3.88	2.15
3000 Central Kootenay	52,045	4.00	2.18	4.03	2.13	5.06	1.69
5000 Kootenay-Boundary	33,235	2.32	1.23	2.34	1.18	3.21	1.68
7000 Okanagan-Similkameen	57,185	5.04	2.64	6.02	3.08	4.72	2.49
9000 Fraser-Cheem	56,930	3.12	1.60	3.33	1.78	4.15	2.08
11000 Central Fraser Valley	115,015	3.14	1.43	3.46	1.52	3.65	1.81
13000 Dowdney-Alouette	62,000	2.10	1.15	2.56	1.23	2.26	1.32
15000 Greater Vancouver	1,168,700	1.63	0.94	1.68	0.93	1.67	0.98
17000 Capital	249,475	1.64	0.87	2.31	1.21	1.18	0.61
19000 Cowichen Valley	45,315	3.09	1.66	3.36	1.69	3.85	2.08
21000 Nanaimo	84,815	3.07	1.58	3.40	1.74	3.22	1.66
23000 Alberni-Clayoquot	32,560	2.75	1.36	2.88	1.27	3.27	1.68
25000 Comox-Strathcona	68,620	1.44	0.80	1.85	0.87	2.85	1.50
27000 Powell River	19,050	5.36	2.58	5.06	2.44	6.18	3.03
29000 Sunshine Coast	16,625	4.84	2.57	6.79	3.58	5.65	2.81
31000 Squamish-Liltooet	18,925	1.82	0.99	2.56	1.37	3.10	1.58
33000 Thompson-Nicola	102,430	2.13	1.10	2.07	0.10	2.65	1.37
35000 Central Okanagan	85,235	3.96	1.93	3.91	1.88	4.32	2.14
37000 North Okanagan	69,033	5.26	2.52	6.44	3.06	5.05	2.50
39000 Columbia-Shuswap	45,425	3.04	1.63	3.56	1.84	2.99	1.66
41000 Cariboo	58,810	3.18	1.93	3.90	2.18	3.42	2.06
43000 Mount Waddington	14,675	8.96	3.04	5.13	1.59	17.77	5.49
45000 Central Coast	3,050	17.99	7.62	21.62	8.86	14.92	7.34
47000 Skeena-Queen Charlotte	24,030	4.82	2.09	5.70	2.58	4.61	1.84
49000 Kitimat-Stikina	41,790	6.26	1.99	4.99	1.66	8.59	2.78
51000 Bulkley-Nechako	38,310	6.23	2.31	5.76	2.10	6.83	2.57
53000 Fraser-Fort George	89,430	3.50	1.41	3.39	1.25	3.72	1.68
55000 Peace River-Liard	55,340	8.00	2.95	9.43	3.65	7.34	2.83
57000 Stikine	2,685	17.15	6.89	17.89	6.88	22.39	8.35
Erreur moyenne		4.83	2.17	5.19	2.31	5.60	2.51

BIBLIOGRAPHIE

BUREAU CENTRAL DE LA STATISTIQUE (1982). British Columbia Municipal Population Estimation Methodology. Document non publié. Ministère du Développement de l'industrie et de la petite entreprise, Victoria.

SPSS INC. (1984). *Statistical Package for the Social Sciences/PC*. Chicago, B265-B280.

SHRYOCK, H.S., et SIEGAL, J.S. (1980). *The Methods and Materials of Demography*. 2, U.S. Bureau of the Census, 628-630.

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 237–245
Statistique Canada

Estimation de la population par âge et par sexe des divisions de recensement et des régions métropolitaines de recensement[1]

RAVI B.P. VERMA, K.G. BASAVARAJAPPA, et ROSEMARY K. BENDER[2]

RÉSUMÉ

Une méthode d'estimation de la population par année d'âge et par sexe a été élaborée pour les petites régions (divisions de recensement et régions métropolitaines de recensement). Pour améliorer leur fiabilité, les estimations démographiques par année d'âge ont été classées par groupe de cinq années d'âge et seules ces données peuvent être diffusées. Elles sont fondées sur la composition par âge et par sexe de la population selon le dernier recensement, les naissances par sexe, les décès par année d'âge et par sexe, les estimations des migrations par âge et par sexe et les bénéficiaires d'allocations familiales âgés de 1 à 14 ans.

MOTS CLÉS: Méthode des composantes par cohorte; erreur absolue moyenne; indice de dissemblance; facteur de séparation.

1. INTRODUCTION

Le présent document décrit la méthode utilisée pour produire des estimations de la population par âge et par sexe des petites régions (divisions de recensement et régions métropolitaines de recensement), donne les résultats de l'examen des méthodes d'estimation et, enfin, traite des facteurs qui influent sur la qualité des estimations. Selon les chiffres du recensement de 1981, la population des 266 divisions de recensement variait de 2,000 à 2,000,000 d'habitants et celle des 24 régions métropolitaines de recensement, de 100,000 à 3,000,000 d'habitants. La section 2 décrit les méthodes d'estimation et les principales sources de données. La section 3 donne les résultats de l'évaluation des estimations relatives à la migration et à la population.

2. MÉTHODES

Dans cette section, nous décrivons les méthodes d'estimation et la préparation des données d'entrée de base.

2.1 Méthode des composantes par cohorte

Pour chaque division de recensement (DR) et chaque région métropolitaine de recensement (RMR), nous utilisons la méthode des composantes par cohorte pour estimer la population par âge. Les équations correspondantes sont les suivantes:

Pour l'âge 0, $$P_0^{t+1} = N - f_0 D_0 + \tfrac{1}{2} M_0 \qquad (1)$$

Pour l'âge 1, $$P_1^{t+1} = P_0^t - \left[(1-f_0)D_0 + \tfrac{1}{2}D_1\right] + \tfrac{1}{2}(M_0 + M_1) \qquad (2)$$

Pour les âges 2 à 84, $$P_{a+1}^{t+1} = P_a^t - \tfrac{1}{2}(D_a + D_{a+1}) + \tfrac{1}{2}(M_a + M_{a+1}) \qquad (3)$$

Pour les âges 85 et plus, $$P_{85+}^{t+1} = P_{84+}^t - \tfrac{1}{2}D_{84} - D_{85+} + \tfrac{1}{2}M_{84} + M_{85+} \qquad (4)$$

---

[1] Version abrégée du document présenté aux réunions du Comité fédéral-provincial sur la démographie tenues les 28 et 29 novembre 1985 à Statistique Canada, Ottawa, Canada. Cette recherche a été faite en collaboration avec le personnel du Programme de données régionales de Statistique Canada.

[2] Ravi B.P. Verma, K.G. Basavarajappa et Rosemary K. Bender, Division de la démographie, Statistique Canada, 4e étage, Immeuble Jean Talon, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

où f_0 = facteur de séparation pour les décès à l'âge 0

M_a = migrants nets d'âge a entre l'année t et l'année $t+1$

N = naissances entre l'année t et l'année $t+1$

D_a = décès à l'âge a entre l'année t et l'année $t+1$

P_a^t = population d'âge a l'année t.

La méthode des composantes par cohorte (publication n° 91-210 au catalogue) est également utilisée par Statistique Canada pour produire des estimations provinciales et par la Colombie-Britannique pour obtenir des estimations par âge de la population des divisions de recensement, des districts scolaires et des districts de santé (Central Statistics Bureau, 1980).

2.2 Préparation des données d'entrée de base

Puisque nous proposons de produire des estimations postcensitaires provisoires de la population dans les 8 mois de la date de référence, nous ne pouvons utiliser les données définitives sur les composantes de l'accroissement démographique de la population parce que celles-ci ne sont connues qu'après un délai de 18 à 24 mois. En conséquence, il faut utiliser des estimations pour chaque composante.

Naissances et décès

Les estimations provisoires des naissances par sexe pour l'année t sont obtenues en multipliant la répartition provinciale en pourcentage par sexe des naissances dans les petites régions pour l'année $t-1$ par le total provisoire des naissances dans la province pour l'année t. De la même façon, on obtient les estimations provisoires des décès par âge et par sexe pour l'année t en multipliant la répartition provinciale des décès par âge et par sexe dans les petites régions pour l'année $t-1$ par le total provisoire des décès dans la province pour l'année t. Enfin, on convertit les décès en décès par cohorte en posant comme hypothèse que les dates de naissance de ceux qui meurent et le nombre de décès sont répartis uniformément sur 12 mois à l'exception des décès à l'âge 0. Les formules correspondantes sont les suivantes:

Pour l'âge 0:
Décès par cohorte (0) = décès (0) × 0.89
Pour l'âge 1,
Décés par cohorte (1) = [décès (0) × 0.11] + [décès (1) × 0.5]
Pour les âges 2 à 84,
Décès par cohorte (âge) = [décès (âge-1) × 0.5] + [décès (âge) × 0.5]
Décès par cohorte (âge 85 et plus) = décès (84) × 0.5 + décès (85 et plus).

Dans les formules ci-dessus, les facteurs de séparation (f) sont de 0.89 pour l'âge 0, de 0.11 pour l'âge 1 et de 0.5 pour tous les autres âges.

Migration nette résiduelle

Tout d'abord, les estimations postcensitaires de la population totale des DR et des RMR calculées par la méthode de régression sont réparties par sexe selon la répartition par sexe

obtenue à partir des données du dernier recensement. Cette méthode de régression qui produit des estimations emboîtées est décrite dans la publication n° 91-211 au catalogue de Statistique Canada. Pour chaque sexe, la migration nette résiduelle est calculée au moyen de la différence entre les variations démographiques et l'accroissement naturel. Cette migration est ensuite répartie pour chaque région par groupe d'âge de 5 ans au moyen de données sur les migrations par âge provenant de trois sources: migration nette résiduelle selon les recensements de 1976 et 1981, données sur les migrations tirées du fichier de l'impôt sur le revenu et réponses à la question du recensement de 1981 sur la mobilité, "Où habitiez-vous il y a 5 ans, c'est-à-dire le 1 juin 1976?". À partir des réponses données à cette question, on peut calculer le nombre de migrants entrants et sortants de chaque petite région. Enfin, la migration par groupe d'âge de 5 ans est répartie en migration par année d'âge au moyen de multiplicateurs de SPRAGUE. Avant d'appliquer ces multiplicateurs, il faut d'abord décomposer la migration nette résiduelle en migration d'entrée et en migration de sortie. En utilisant comme référence les données fiscales pour le calcul des migrations d'entrée et de sortie, on refait ensuite ce calcul pour chaque groupe d'âge de 5 ans.

$$\text{Migration d'entrée résiduelle} = \frac{\text{migration d'entrée selon les données de l'impôt sur le revenu}}{\text{migration nette selon les données de l'impôt sur le revenu}} \times \text{migration nette résiduelle}$$

$$\text{Migration de sortie résiduelle} = \text{migration d'entrée résiduelle} - \text{migration nette résiduelle}$$

Quand les ratios décrits ci-dessus sont utilisés, d'importants problèmes se présentent lorsque la migration nette fractionnée n'a pas le même signe que la migration nette calculée à partir des données fiscales. Dans ce cas, on conserve le signe de la migration nette fractionnée, mais les entrées et les sorties qui en résultent sont échangées de façon à produire le signe approprié. Cette opération est fondée sur l'hypothèse de l'ampleur égale d'un renversement du flux migratoire.

2.3 Chiffres tirés du fichier des bénéficiaires d'allocations familiales âgés de 1 à 14 ans

Les estimations de la population produites au moyen de la méthode des composantes et des cohortes pour les groupes d'âge 1-4, 5-9 et 10-14 ans sont remplacées par les chiffres sur les bénéficiaires d'allocations familiales des mêmes âges qu'il est facile d'obtenir, pour les DR et les RMR, 3 ou 4 mois après la date de référence. Le programme des allocations familiales est universel au Canada et les chiffres obtenus à partir du fichier des allocations familiales sont jugés complets à toutes fins pratiques. Puisque ces données ne sont pas disponibles par sexe, elles sont réparties par sexe selon la répartition par sexe obtenue à partir des données du dernier recensement.

2.4 Ajustements pour assurer la cohérence des estimations provinciales et des estimations par division de recensement

Les estimations postcensitaires emboîtées de la population totale de chaque DR et de chaque RMR sont diffusées dans les 6 mois de la date de référence. De plus, les estimations provinciales sont aussi produites par âge et par sexe dans à peu près le même délai. Les estimations de la population par âge et par sexe obtenues de la façon décrite ci-dessus pour les DR de chaque province sont comparées aux estimations de la population totale des divisions de recensement et aux estimations provinciales par âge et par sexe, et les différences sont distribuées proportionnellement. Pour les régions métropolitaines de recensement, les totaux par âge et par sexe ne sont corrigés qu'en fonction des estimations de la population totale des RMR.

3. ÉVALUATION

L'évaluation est faite en fonction de trois critères: i) la précision, ii) les délais de production et iii) la cohérence. Chacun des critères est examiné en détail dans la partie qui suit.

3.1 Précision

La précision des estimations de la population par âge et par sexe dépend dans une large mesure de la qualité des estimations de la répartition par âge et par sexe des migrants, car les données sur les décès par âge et par sexe sont jugées satisfaisantes. Par conséquent, l'évaluation des estimations de la population par âge et par sexe jette indirectement de la lumière sur l'exactitude des estimations des migrations par âge et par sexe. Pour évaluer l'exactitude des estimations de la population, on compare ces estimations aux chiffres de recensement correspondants.

Pour chaque DR et chaque RMR, les trois séries d'estimations de la population par âge et par sexe, au 1 juin 1981, qui ont été produites à l'aide de la répartition par âge des migrants établie à partir des trois sources, la migration résiduelle (1976-1981), la mobilité (1981) et les données annuelles du fichier sur les impôts et les chiffres des fichiers des allocations familiales, de la façon décrite dans les sections 2.1 à 2.4, ont été comparées aux chiffres du recensement de 1981. Les différences observées ont été considérées comme des erreurs et, pour chaque petite région, on a calculé un indice sommaire appelé "erreur absolue moyenne" (EAM) en prenant la moyenne arithmétique des erreurs en pourcentage (sans tenant compte du signe) pour les seize groupes d'âge de cinq ans. Plus la valeur de cet indice est petite, plus les estimations sont exactes. Au tableau 1, un classement des DR par province et des RMR est présenté pour quatre niveaux d'erreur absolue moyenne: moins de 3%, 3-5%, 5-10% et plus de 10%. Dans l'ensemble, il ressort de ce tableau que la répartition par âge des migrants fondée sur les données sur la migration résiduelle produit de meilleures estimations. Pour les personnes de sexe masculin, environ 66% des divisions de recensement présentent une EAM inférieure à 5%. Dans le cas des personnes de sexe féminin, ce pourcentage est légèrement supérieur, soit 69%. En revanche, on observe des pourcentages inférieurs dans les estimations de la population produites à partir des données sur la mobilité (55 et 57% respectivement pour les hommes et les femmes) et des données sur les migrations obtenues à partir des données fiscales (9 et 19% respectivement pour les hommes et les femmes).

Pour les RMR également, la répartition par âge des migrants calculée à partir des données sur la migration résiduelle semble produire de meilleures estimations. La proportion de RMR où l'EAM est inférieure à 3% était de 58% dans le cas des hommes et de 79% dans le cas des femmes. La répartition fondée sur la mobilité et celle qui est fondée sur les données fiscales viennent respectivement au deuxième et au troisième rang. Cela vaut tant pour les hommes que pour les femmes.

Dans toutes les provinces, à l'exception de l'Île-du-Prince-Édouard, la précision des trois séries de répartition par âge des migrants semble bonne. Cette observation vaut tant pour les deux sexes. Cependant, dans certains cas, la répartition par âge des migrants fondée sur les données sur la migration résiduelle et la répartition par âge des migrants fondée sur les données sur la mobilité semblent donner des résultats semblables. Dans le cas des hommes, on a observé une similitude entre les résultats des deux méthodes dans trois provinces, soit Terre-Neuve, le Nouveau-Brunswick et la Colombie-Britannique, alors que pour les femmes, c'était le cas seulement au Nouveau-Brunswick.

Il convient de souligner que, pour la répartition par âge des migrants établie au moyen de la méthode fondée sur les données sur la migration résiduelle, on utilise les répartitions par âge calculées à partir des recensements de 1976 et 1981. Par conséquent, on peut s'attendre que les estimations de la population au 3 juin 1981, produites à partir de la répartition

Tableau 1

Répartition des divisions de recensement et des RMR selon la pécision des estimations de la population par âge, 1981

Provinces	Méthodes d'évaluation de la migration	Niveau d'erreur absolue moyenne (%) par sexe							
		Masculin				Féminin			
		Moins de 3	3-5	5-10	10+	Moins de 3	3-5	5-10	10+
Terre-Neuve	R	8	2	0	0	10	0	0	0
	M	2	7	1	0	3	3	3	1
	F	0	0	5	5	1	1	3	5
Île-du-Prince-Édouard	R	1	1	1	0	2	0	0	1
	M	1	2	0	0	1	2	0	0
	F	0	0	1	2	0	0	2	1
Nouvelle-Écosse	R	8	4	3	3	8	5	3	2
	M	3	6	4	5	5	6	2	5
	F	0	2	8	8	1	3	11	3
Nouveau-Brunswick	R	10	1	2	2	7	4	3	1
	M	4	7	3	1	3	8	3	1
	F	0	2	5	8	0	4	5	6
Québec	R	13	27	23	13	24	18	19	15
	M	12	26	26	12	17	23	21	15
	F	1	5	37	33	3	13	32	28
Ontario	R	30	8	8	7	37	5	4	7
	M	8	16	18	11	21	10	10	12
	F	0	8	34	11	4	10	31	8
Manitoba	R	4	6	8	5	4	6	8	5
	M	1	5	12	5	0	6	7	10
	F	0	1	8	14	0	1	4	18
Saskatchewan	R	10	5	1	2	9	5	2	2
	M	1	11	5	1	4	5	5	4
	F	1	1	10	6	1	1	12	4
Alberta	R	9	3	1	2	8	3	3	1
	M	5	5	3	2	5	4	5	1
	F	0	2	5	8	0	3	5	7
Colombie-Britannique	R	19	3	3	4	23	1	1	4
	M	9	13	2	5	14	8	3	4
	F	0	0	13	16	0	3	16	10
RMR	R	14	8	2	0	19	3	2	0
	M	2	17	5	0	10	9	4	1
	F	1	7	13	3	1	10	12	1

Note: R: Répartition par âge des migrants fondée sur la méthode des résidus, 1976-1981.
M: Répartition par âge des migrants fondée sur la mobilité, 1981.
F: Répartition par âge des migrants fondée sur les données fiscales annuelles.
Source: Division de la désmographie, Statistique Canada, 1985.

par âge des migrants calculée au moyen de la méthode fondée sur les données sur la migration résiduelle, soient comparables aux estimations de la population fondées sur la répartition par âge calculée à partir du recensement de 1981. On ne peut donc pas conclure que la répartition par âge des migrants fondée sur les données sur la migration résiduelle produit de meilleures estimations que la répartition par âge des migrants fondée sur les réponses à la question sur la mobilité ou de meilleures estimations que la répartition par âge des migrants calculée à partir des données fiscales.

Le tableau 2 montre la répartition en pourcentage des DR et des RMR pour lesquelles on enregistre des observations aberrantes. On considère comme observation aberrante une EAM de plus de 10% dans le cas des DR et de plus de 5% dans le cas des RMR. Les données sont classées par sexe et selon les trois types de répartition par âge des migrants. Comme on s'y attendait, aussi bien pour les hommes que pour les femmes, le pourcentage de cas d'observations aberrantes est en général peu élevé lorsque les estimations sont produites au moyen de la répartition par âge des migrants calculée selon la méthode fondée sur les données sur la migration résiduelle. Par contre, le pourcentage de cas d'observations aberrantes tend à être élevé lorsque les estimations sont produites au moyen de la répartition des migrants fondée sur les données fiscales.

Stabilité temporelle des trois séries d'estimations dans les années postcensitaires, 1982-1984

Comme il n'y a pas de répartitions par âge standard avec lesquelles les estimations de la population peuvent être comparées pendant les années postcensitaires, nous comparons entre elles les trois séries d'estimations de la population par âge et par sexe pour examiner la stabilité temporelle entre elles. Un indice sommaire, l'indice de dissemblance, correspondant à la moitié de la somme des écarts absolus entre les estimations de la population produites à partir de deux distributions (en pourcentage) par âge des migrants, est utilisé à cette fin. L'indice varie de 0 à 100. Plus la valeur de l'indice est faible, plus les deux distributions comparées sont semblables. Les petites régions sont classées en trois niveaux de dissemblance: i) le plus petit niveau de différence, c.-à-d. des indices de 0 à 5%, ii) le niveau moyen de différence, indices de 5 à 10%, et iii) les observations aberrantes, c'est-à-dire les régions dont l'indice de dissemblance est égal ou supérieur à 10%. Le classement des DR selon la valeur de l'indice de dissemblance figure au tableau 3 et celui des RMR, au tableau 4.

Si on examine le tableau 3, il semble que les trois types de répartition par âge de la population tendent à être semblables et qu'en moyenne, un très grand pourcentage de cas, environ 90%, se situent dans la catégorie des faibles indices de dissemblance (0 à 5%), alors que seulement 7% se situent dans la catégorie des indices de 5 à 10%.

Nous avons aussi examiné, pour les dix provinces et leur total de population, le pourcentage de cas où le niveau de différence est très élevé (indice de dissemblance excédant 10%). Chez les hommes, les pourcentages de cas extrêmes sont faibles, soit de 3 à 5%, entre la répartition par âge calculée selon la méthode fondée sur les données sur la migration résiduelle et la répartition par âge calculée à partir des données sur la mobilité. Chez les femmes, on note une proportion relativement plus élevée d'observations aberrantes. Pour les deux autres comparaisons, soit la comparaison entre les estimations produites selon la méthode fondée sur les données sur la migration résiduelle et les estimations produites selon la méthode fondée sur les données fiscales et la comparaison entre les estimations produites selon la méthode fondée sur les données sur la mobilité et les estimations produites selon la méthode fondée sur les données fiscales, on trouve une proportion légèrement plus importante de cas d'observations aberrantes. Les résultats sont semblables pour les régions métropolitaines de recensement (voir tableau 4).

En conclusion, on peut dire que, même si les trois répartitions par âge des migrations (fondées sur les données sur la migration résiduelle, sur les données sur la mobilité et sur les données fiscales) sont différentes, les répartitions par âge de la population qui en résultent se ressemblent beaucoup.

3.2 Délai de production

Par délai de production, on entend la production d'estimations dans le délai le plus court possible après la date de référence. En utilisant les chiffres provisoires de la population

Tableau 2

Pourcentage des divisions de recensement par province et des RMR présentant des observations aberrantes[a], 1981

Provinces	Hommes			Femmes		
	R	M	R	R	M	R
Terre-Neuve	0	0	50	0	10	50
Île-du-Prince-Édouard	0	0	67	33	0	33
Nouvelle-Écosse	17	28	44	11	28	17
Nouveau-Brunswick	13	7	53	7	7	40
Québec	17	16	43	20	20	37
Ontario	13	21	21	13	23	15
Manitoba	22	22	61	22	43	78
Saskatchewan	11	6	33	11	22	22
Alberta	13	13	53	7	7	47
Colombie-Britannique	14	17	55	14	14	34
Total	15	16	43	15	20	35
RMR	8	21	67	8	21	54

Note: R: Répartitionpar âge des migrants fondée sur la méthode des résidue.
M: Répartition par âge des migrants fondée sur la mobilité.
F: Répartition par âge des migrants fondée sur les données fiscales.
[a] Les observations aberrantes correspondent aux DR qui affichent une EAM de plus de 10% et aux RMR dont l'EAM est supérieur à 5%.
Source: Tableau 1.

Tableau 3

Répartition des divisions de recensement selon le niveau de dissemblance obtenu en comparant les répartitions par âge de la population calculées à partir des trois sources de données: données sur la migration résiduelle, données sur la mobilité et données sur la migration tirées des données fiscales, 1982-1984

Année/	Masculin			Féminin		
Indice de dissemblance	Migration résiduelle et mobilité	Migration résiduelle et données fiscales	Mobilité et données fiscales	Migration résiduelle et mobilité	Migration résiduelle et données fiscales	Mobilité et données fiscales
ANNÉE 1982						
0-5	245	237	242	240	241	234
5-10	7	13	10	8	5	11
10+	8	10	8	12	14	15
Total	260	260	260	260	260	260
ANNÉE 1983						
0-5	235	221	223	230	229	223
5-10	11	18	21	10	13	13
10+	14	21	16	20	18	24
Total	260	260	260	260	260	260
ANNÉE 1984						
0-5	240	226	229	235	233	231
5-10	11	16	14	15	13	12
10+	9	18	17	10	14	17
Total	260	260	260	260	260	260

Source: Division de la démographie, Statistique Canada, octobre 1985.

Tableau 4

Répartition des régions métropolitlaines de recensement selon le niveau de dissemblance obtenu en comparant les répartitions par âge de la population calculées à partir des trois sources de données: données sur la migration résiduelle, données sur la mobilité et données sur la migration tirées des données fiscales, 1982-1984

Année/	Masculin			Féminin		
Indice de dissemblance	Migration résiduelle et mobilité	Migration résiduelle et données fiscales	Mobilité et données fiscales	Migration résiduelle et mobilité	Migration résiduelle et données fiscales	Mobilité et données fiscales
ANNÉE 1982						
0-3	24	24	24	23	22	23
3-5	0	0	0	0	1	0
5+	0	0	0	1	1	1
Total	24	24	24	24	24	24
ANNÉE 1983						
0-3	22	23	22	21	20	20
3-5	2	0	0	1	2	3
5+	0	1	2	2	2	1
Total	24	24	24	24	24	24
ANNÉE 1984						
0-3	21	21	21	21	20	20
3-5	2	2	2	0	1	0
5+	1	1	1	3	3	4
Total	24	24	24	24	24	24

Source: Division de la démographie, Statistique Canada, octobre 1985.

(dans ce cas-ci, les estimations emboîtées de la population), produits dans les six mois après la date de référence, on pourrait établir, avec les estimations des naissances, des décès par âge et des migrations nettes fractionnées par âge, de la façon décrite dans les parties 2.1 à 2.4, les estimations de la population par âge et par sexe pour les DR et les RMR, au plus tard huit mois après la date de référence.

3.3 Cohérence

Par cohérence, on entend l'homogénéité des sources d'ensembles de données utilisées pour produire les estimations à divers niveaux de désagrégation administrative ou à divers niveaux de répartition géographique et l'uniformité des méthodes d'estimation. Même s'il faut utiliser dans certains cas une méthode différente, il est de loin préférable d'utiliser toujours la même méthode pour assurer la cohérence méthodologique entre les divers niveaux de désagrégation géographique.

Au niveau provincial ainsi qu'au niveau des DR et des RMR, les sources de données sont les mêmes en ce qui a trait aux naissances, aux décès c'est-à-dire, les statistiques de l'état civil. Au niveau des données sur la migration, les sources sont également les mêmes, à savoir les données fiscales et les données sur la mobilité obtenues à partir du recensement de 1981, pour tous les niveaux de désagrégation géographique. Une autre série de données, soit la série de données sur la répartition par âge des migrants fondée sur la méthode des données sur la migration résiduelle obtenues à partir de deux recensements consécutifs, est toutefois également utilisée.

A tous les niveaux géographiques on utilise comme méthodologie celle des composantes par cohorte.

4. CONCLUSION

Trois séries d'estimations par âge et par sexe ont été produites, pour les DR et les RMR, au moyen de la méthode des composantes par cohorte. Pour chaque série, on utilise une composante différente de la migration par âge et par sexe: i) pour la première, le fichier des données fiscales, ii) pour la deuxième, les données sur la mobilité obtenues à partir du dernier recensement, et iii) pour la troisième, les données sur la migration résiduelle obtenues à partir de deux recensements consécutifs.

Bien que les trois répartitions par âge des migrants diffèrent, les estimations de la population par âge et par sexe qu'ils produisent se ressemblent beaucoup. Chaque série comporte ses propres hypothèses. L'utilisation de la répartition par âge des migrants fondée sur les données sur la migration résiduelle pour produire des estimations postcensitaires suppose que la répartition par âge demeure la même tout au long de la période d'estimation. L'utilisation des données sur la mobilité par âge et par sexe pour les années postcensitaires repose sur la même hypothèse. L'utilisation des données fiscales suppose que la répartition par âge et par sexe demeure la même entre deux années consécutives. Ces séries de données ne mesurent toutefois pas le même type de mouvement. Les données sur la migration résiduelle mesurent seulement le mouvement net entre deux recensements consécutifs (c'est-à-dire, dans ce cas-ci, les recensements de 1976 et de 1981). Les données obtenues à partir des réponses à la question sur la mobilité mesurent également le résultat de mouvements sur une période de cinq ans. Par contre, les fichiers de données fiscales indiquent le mouvement survenu pendant une période d'environ 12 mois. À partir des comparaisons exposées dans le présent document, il est impossible de conclure qu'une série de données sur les migrations produit de meilleures estimations de la population qu'une autre. Une évaluation plus satisfaisante des trois séries d'estimations ne pourra être faite que lorsque les résultats du prochain recensement seront connus.

BIBLIOGRAPHIE

CENTRAL STATISTICS BUREAU (1980), *Population Estimates by Age Groups for School Districts*, 1977-1980. Rapport technique non publié, gouvernement de la Colombie-Britannique.

NORRIS, D., et STANDISH, L. (1983). Rapport technique sur la production des données migratoires à partir des dossiers des impôts. Rapport technique, Division de l'exploitation des données administratives, Statistique Canada.

STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population suivant l'état matrimonial, l'âge, le sexe et composantes de l'accroissement, Canada et provinces*, vol. 2, deuxième édition. N° 91-210 au catalogue, Ottawa: ministère des Approvisionnements et Services.

STATISTIQUE CANADA (Annuel). *Estimations annuelles postcensitaires de la population des divisions et régions métropolitaines de recensement (méthode de régression)*. N° 91-211 au catalogue, Ottawa: ministère des Approvisionnements et Services.

STONE, LEROY O. (1980). Evaluating the Relative Accuracy and Significance of Net Migration Estimates. *Demography*, 4, 310-330

Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 247–251
Statistique Canada

Estimations de la population des petites régions: expérience de Statistique Canada[1]

ROSEMARY K. BENDER[2]

RÉSUMÉ

À Statistique Canada, les estimations de la population des divisions et régions métropolitaines de recensement sont calculées à l'aide de la méthode des composantes et d'un modèle de régression qui produit des estimations emboîtées. Le présent document décrit les travaux effectués dans ce domaine relativement à la période 1981-1985 et expose particulièrement les difficultés qu'a posées l'utilisation des données sur les bénéficiaires d'allocations familiales.

MOTS CLÉS: Estimations emboîtées; estimations de la méthode des composantes; bénéficiaires d'allocations familiales; fichiers de codes postaux.

1. INTRODUCTION

Pour estimer la population des divisions de recensement (DR) et des régions métropolitaines de recensement (RMR), Statistique Canada utilise actuellement la méthode des composantes et un modèle de régression. Les estimations par régression pour les années 1982, 1983 et 1985 ont été publiées selon les délais prévus dans le numéro 91-211 au catalogue, mais celles de 1984 ont pu être diffusées seulement en mars 1985 à cause d'un retard dans la collecte des données sur les allocations familiales. En outre, comme on le verra plus loin, ces données présentaient des lacunes sur le plan de la fiabilité. Par conséquent, comme les estimations de la population des RMR fondées sur cette information n'étaient pas acceptables, il a fallu appliquer une autre méthode.

Les estimations de la population des DR et des RMR, calculées par la méthode des composantes, ont été publiées dans le numéro 91-212 au catalogue comme prévu relativement aux années 1982 et 1983; les estimations relatives à 1984 devraient paraître en avril 1986. L'évaluation de ce genre d'estimations produites jusqu'à maintenant révèle que la qualité de ces données est bonne.

2. MODIFICATIONS

Depuis la mise en oeuvre du programme d'estimations par régression de la population des DR et RMR, en 1982, certaines modifications ont dû être apportées aux données et aux méthodes de travail:

- Pour le calcul des estimations de 1983 relatives à la DR de Chicoutimi et à la RMR de Chicoutimi-Jonquière, les données sur les allocations familiales ont été corrigées en fonction du niveau de croissance enregistré l'année précédente. La source du problème était liée aux codes postaux utilisés pour recueillir les données sur les allocations familiales.
- En 1984, les estimations relatives à 17 divisions de recensement ont été imputées en fonction d'estimations provisoires calculées par la méthode des composantes.
- En 1984, il a été décidé de calculer les estimations de la population de la RMR de Calgary à partir des résultats du recensement effectué dans cette ville, et d'étendre ce projet pour toute la période 1981-1986 visée.

[1] Version abrégée du document présenté aux réunions du Comité fédéral-provincial sur la démographie, les 28 et 29 novembre 1985, à Ottawa, Canada.

[2] Rosemary K. Bender, Division de la démographie, Direction du recensement et de la statistique démographie, Statistique Canada, 4e étage, Immeuble Jean-Talon, Parc Tunney, Ottawa, Ontario, Canada K1A 0T6.

- En 1984 également, une nouvelle méthode a été mise au point pour toutes les RMR hors celle de Calgary; cette méthode permet d'agréger les estimations par régression relatives aux divisions de recensement. Ce projet sera appliqué pour toute la période 1981-1986 visée.

Les difficultés que la production de ces estimations a soulevées sont exposées en détail dans les prochaines sections.

3. COLLECTE DES DONNÉES NÉCESSAIRES AU CALCUL DES ESTIMATIONS PAR RÉGRESSION

Il y a eu des retards dans la production des estimations de 1984 à cause de la difficulté d'obtenir les données sur les bénéficiaires d'allocations familiales auprès de Santé et Bien-être Canada, et d'utiliser les bons fichiers de conversion des codes postaux qui sont essentiels au traitement de ces données.

i) *Données sur les allocations familiales*

Le nombre de bénéficiaires d'allocations familiales, en date du 1 juin, est habituellement connu à la mi-septembre de chaque année. Cependant, les données de 1984 n'ont pas été fournies à temps par Santé et Bien-être Canada à cause de la décentralisation des opérations régionales du programme en Ontario. Il y a eu également des problèmes d'extraction, dans les fichiers provinciaux, de données sur la date d'entrée en vigueur des transferts, de même que des codes de raison des transferts interrégionaux. Les données de 1984 fournies à Statistique Canada ont été présentées sous une forme non épurée, en novembre. Des mesures correctives ont été prises par Santé et Bien-être Canada, et les données sur les allocations familiales au 1er juin 1985 ont été diffusées comme prévu.

ii) *Fichiers des codes postaux*

Un code postal est attribué à chaque donnée sur les bénéficiaires d'allocations familiales recueillies auprès de Santé et Bien-être Canada. Aussi, pour dénombrer les enfants bénéficiaires d'allocations familiales dans chaque DR et RMR, il faut créer un fichier qui regroupe les codes postaux par DR et RMR. Pour cela il faut utiliser un fichier principal qui contient tous les codes postaux au Canada, de même que des codes géographiques précis qui permettent d'attribuer les codes postaux selon chaque niveau de répartition géographique.

Des difficultés imprévues se sont produites dans l'utilisation des codes et, dans certains cas, elles ont eu des conséquences graves. Pour la production de ces estimations, il est important que les fichiers de codes postaux que Santé et Bien-être Canada utilise chaque année soient compatibles avec celui que Statistique Canada a construit en fonction du modèle de régression. Le seul changement qu'on devrait apporter à ce fichier est l'ajout de nouveaux codes postaux. Le transfert de codes postaux d'une région à une autre peut entraîner des modifications de la composition de la population qui ne se sont pas vraiment produites.

Aussi, les problèmes qui se sont posés viennent du fait que, depuis la construction du modèle de régression de Statistique Canada, divers ministères et divisions ont créé des fichiers de codes postaux. En 1982 et 1983, c'est la Division de l'exploitation des données administratives (DEDA) de Statistique Canada qui l'a fait; l'année suivante, la Division des normes de Statistique Canada a été chargée de cette tâche et, en 1985, c'est le ministère de la Santé et du Bien-être qui a pris la relève. Chacun a adopté une méthode particulière pour constituer ces fichiers, de sorte que les données sur les allocations familiales n'étaient pas uniformes d'une année à l'autre. Cette situation a créé deux genres de problèmes, dont un a été résolu. Le second, toutefois, persistera pendant toute la période postcensitaire de 1981 à 1986.

La première cause de difficultés a été le transfert de codes postaux d'une région à une autre. Le fichier principal des codes a été créé par la Division des normes de Statistique Canada.

Dans certains cas, le code géographique d'une DR ou RMR est inexact ou inexistant. Cette situation se produit la plupart du temps avec les codes ruraux dans les DR, parce que les codes postaux correspondent souvent aux codes d'identification des bureaux de poste qui desservent de grands territoires à travers plus qu'une DR. L'inclusion des codes géographiques des RMR est assez récente et leur fiabilité s'accroît chaque année. Par conséquent, notre première présomption selon laquelle le fichier des codes postaux serait uniforme d'une année à l'autre n'était pas tout à fait juste.

Les premiers fichiers de Statistique Canada ont été créés par la Division de l'exploitation des données administratives (DEDA) qui a modifié sa version du fichier principal avant de procéder au regroupement des données. En 1984, la Division des normes a pris la relève pour la production de ce fichier. Lorsque les conséquences de ce transfert de responsabilité sont devenues évidentes, on a conçu en collaboration avec la DEDA un moyen de faire concorder le fichier principal original avec le dernier fichier principal créé par la Division des normes, en introduisant uniquement les nouveaux codes postaux et en ne tenant pas compte des corrections apportées à des codes DR ou RMR. Il faut admettre qu'en procédant de cette façon on n'obtient pas le fichier de codes postaux le plus exact qui soit. Cependant, pour la production des estimations, il importe de connaître les variations dans les proportions du nombre d'enfants bénéficiaires d'allocations familiales. Le fait d'utiliser certains codes postaux inexacts mais uniformes a pour effet d'inclure des enfants dans une région ou de les exclure d'une autre région de bénéficiaires. Les résultats de ces calculs ne sont pas très différents de ceux qu'on obtiendrait à partir de codes postaux exacts, mais ils seraient différents si on incluait ou excluait subitement ces enfants dans les calculs.

L'introduction uniquement des nouveaux codes postaux dans le fichier de Statistique Canada a amélioré considérablement la qualité des données de 1984 sur les allocations familiales au niveau des divisions de recensement. On a dû imputer les estimations par régression de seulement 17 DR sur 231 (sans les DR de la C.-B.,parce que cette province produit ses propres estimations). À cause des retards dans la collecte des données, cela a permis d'utiliser les estimations provisoires calculées à l'aide de la méthode des composantes. Dans le cas des régions métropolitaines de recensement, il y avait encore des incohérences qui sont imputables à un autre genre de problème.

En effet, lorsque les codes postaux sont groupés par DR et RMR, ils sont également répartis par intervalle de codes. Par exemple, si les codes postaux A1A1A1, A1A1A2, A1A1A3 et A1A1A4 correspondent tous au même code de RMR, alors ils sont groupés dans l'intervalle A1A1A1–A1A1A4. Toutefois, lorsqu'on traite la série de plus de 600,000 codes postaux, il faut poser certaines hypothèses selon le logiciel utilisé. Si, dans l'exemple donné plus haut, le code A1A1A2 n'apparaissait pas, le programme créerait quand même le même intervalle, car il supposerait que si le code A1A1A2 existait il aurait le même code de RMR que les autres codes postaux de cet intervalle. Ce genre de suppositon pourrait modifier les données sur les allocations familiales traitées relativement à chaque région. En outre, l'application de divers logiciels dans une année pourrait produire des incohérences graves.

À notre avis, il s'agit là de la principale cause de la mauvaise qualité des données sur les bénéficiaires d'allocations familiales au niveau des RMR. Les logiciels que la DEDA et la Division des normes utilisent sont différents et, ce qui n'arrange pas les choses, depuis 1985, le programme relève de Santé et Bien-être Canada qui utilise lui aussi un autre logiciel. Il faut donc laisser de côté ce genre de données et élaborer une autre méthode d'estimation de la population des RMR.

4. NOUVELLE MÉTHODE POUR LES RMR

Les estimations de la population des RMR pour 1982 et 1983 étaient des estimations emboîtées du genre de celles relatives aux divisions de recensement. Dans le programme

d évaluation des estimations de 1984, les estimations d'un bon nombre de régions métropolitainnes de recensement ont été jugées incompatibles avec celles tirées d'autres sources et non conformes aux tendances passées. Comme il a été mentionné plus haut, les difficultés qui ont été soulevées tenaient plus à la qualité des données des fichiers qu'aux méthodes de calcul.

Une fois pris en considération le manque d'uniformité des estimations et les observations faites par les points de contact dans les provinces, il a été décidé d'adopter une autre méthode, celle qui avait été conçue pour estimer diverses composantes de l'accroissement démographique de la population. Selon cette méthode, on agrège les estimations par régression relatives aux divisions de recensement, à l'aide du ratio de la population d'une RMR et de celle des DR qui la chevauchent, en fonction des estimations de l'année précédente établies par la méthode des composantes. Par la comparaison des estimations relatives à 1981 calculées selon cette méthode et des chiffres de population des régions métropolitaines de recensement selon le recensement de 1981, on obtient une erreur absolue moyenne de 1.3%, comparativement à 2.3% selon la méthode de calcul des estimations emboîtées.

Pour assurer un degré d'uniformité des données pour la période de 1981-1986, cette nouvelle méthode a été appliquée aux calculs des estimations de la population des RMR pour les années 1982 à 1985, et elle sera utilisée également pour 1986. Ainsi, des estimations de la population des RMR autres que celle de Calgary sont établies par agrégation des estimations emboîtées relatives aux divisions de recensement, et les estimations relatives à Calgary (voir explication plus bas) sont fondées sur les résultats du recensement annuel mené par cette ville.

En 1984, les estimations emboîtées relatives à la RMR de Calgary pour 1982 et 1983 ont été trop élevées par rapport aux chiffres du recensement annuel de Calgary. Les estimations calculées par la méthode des composantes ont servi à corriger les estimations emboîtées relatives à cette ville. Il a été décidé de publier des estimations fondées sur les chiffres du recensement de Calgary, en extrapolant les données d'avril au 1 juin. Cela est conforme à la ligne de conduite de Statistique Canada, selon laquelle, lorsqu'il y a un recensement de la population, les chiffres obtenus doivent être utilisés de préférence aux estimations calculées à l'aide d'une méthode indirecte, à moins qu'on n'ait des motifs évidents de douter de la validité des chiffres du recensement.

5. COMPARAISON D'AUTRES SOURCES DE DONNÉES

Chaque fois que cela est possible, les estimations par régression et les estimations calculées par la méthode des composantes sont comparées avec l'information d'autres sources possibles. En effet, la Saskatchewan et l'Alberta nous fournissent des données sur le nombre de bénéficiaires de leur programme de soins de santé; cette information est utilisée dans le modèle de régression. Toutefois, ces estimations font l'objet d'une évaluation qui vise à vérifier leur uniformité par rapport aux données sur les bénéficiaires d'allocations familiales et sur les tendances passées. Dans la plupart des cas, ces estimations étaient compatibles, et les cas de divergences ont pu être attribués aux difficultés liées aux données sur les allocations familiales.

Le Bureau de la statistique du Québec produit des estimations annuelles de la population relatives aux régions administratives, lesquelles sont des subdivisions des DR du Québec. Ces estimations sont comparables à celles de l'agence fédérale, sauf celles relatives à la DR du Nouveau-Québec. En effet, cette division qui est située dans le nord du Québec comprend essentiellement des territoires non municipalisés, ce qui complique l'estimation de la population. Le B.S.Q. utilise habituellement les estimations de Statistique Canada, quoique, pour 1984, il ait imputé ses estimations pour le Nouveau-Québec.

Statistique Canada est toujours reconnaissant de l'information fournie par les utilisateurs qui disposent de données particulières sur des localités.

6. CONCLUSION

En règle générale, les méthodes utilisées pour produire les estimations de la population des divisions et régions métropolitaines de recensement ont été très efficaces. Cependant, dans le cas des estimations par régression, il a fallu imputer les estimations de certaines DR à l'aide d'autres genres de données et réviser la méthode de calcul des estimations relatives aux RMR, à cause des problèmes d'obtention des données d'entrée.

Les problèmes qui se sont posés avaient trait pour la plupart à l'utilisation des données sur les bénéficiaires d'allocations familiales et des fichiers de codes postaux. Une bonne partie de ces problèmes, toutefois, ont pu être réglés. Par ailleurs, il est possible que, du fait que Santé et Bien-être Canada soit maintenant chargé de créer les fichiers de codes postaux, il y ait encore des cas d'incohérence dans les données de 1986 et qu'il faille procéder à une évaluation rigoureuse.

Malgré ces difficultés, les estimations relatives à 1986 seront produites à l'aide de la méthode de régression qui sera quelque peu remaniée. Si, toutefois, on décidait de continuer d'appliquer les méthodes actuelles pour la période 1986-1991, il faudrait tout d'abord s'assurer que des fichiers de codes postaux compatibles seront traités par le même ministère ou service durant toute la période visée.

COLLABORATION À LA RÉDACTION

En conformité avec la politique rédactionnelle de Techniques d'enquête, les articles publiés dans le volume 11 (numéro 1 et 2) ont fait l'objet de critiques, sauf celui de M. Wilk et les exposés présentés à la conférence du comité fédéral-provincial de la démographie, tenue à Ottawa les 28 et 29 novembre 1985. (Nous faisons paraître des versions condensées de ces textes dans la présente édition pour mieux renseigner les lecteurs sur les nouveautés méthodologiques dans un domaine important de leurs applications.)

Les responsables de la revue Techniques d'enquête désirent remercier les personnes suivantes qui ont accepté de faire la critique des articles présentés au cours de l'année dernière:

D.A. Binder
Y.P. Chaubey
G.H. Choudhry
E.B. Dagum
J. Gambino
G.B. Gray
M.A. Hidiroglou
S.K. McKenzie
M. Morry
D.A. Pierce
B. Quenneville
C.E. Särndal
A. Satin
A. Singh
J. Tourigny
A. Van Barren

Statistique Canada | Statistics Canada

ÉLABORATION ET CONCEPTION DES QUESTIONNAIRES D'ENQUÊTE

Le succès d'une enquête dépend dans une grande mesure de la conception du questionnaire. Ou, il est étonnant de constater qu'on néglige très souvent d'accorder suffisamment d'attention, de soin et de ressources à l'élaboration des questionnaires d'enquête. En conséquence, bien des enquêtes ne produisent pas tous les résultats qu'on pourrait en attendre.

Par son introduction et ses quatre chapitres portant sur la conception, l'élaboration, la production et l'évaluation des questionnaires, la publication *Élaboration et conception des questionnaires d'enquête*, vise à promouvoir une conception soignée des questionnaires et à servir d'instrument de référence et de formation. On y traite de questions aussi diverses que la formulation des questions, la présentation, la disposition, etc., au moyen d'exemples d'enquêtes fédérales effectuées récemment. Une liste récapitulative et une bibliographie complètent ces 132 pages (15 cm. x 23 cm.).

Table des matières

- Introduction
- Élaboration du questionnaire
- Conception du questionnaire
  - Qualité des données
  - Combinaison de sujets
  - Transposition des concepts en termes opérationnels
  - Le questionnaire et la formule d'enquête
  - Formulation des questions
  - Ordre des parties du questionnaire
  - Questions ouvertes et questions fermées
  - Échelles d'attitudes
- Production du questionnaire
  - Présentation
  - Codage et saisie des données
  - Impression
  - Administration du questionnaire
- Mise à l'essai et évaluation du questionnaire
- Liste récapitulative des étapes de la conception d'un questionnaire
- Bibliographie

BON DE COMMANDE

PF 02923

Postez à:
Publications-Vente et service
Statistique Canada
Ottawa, K1A 0T6

(Lettres moulées s.v.p.)

Compagnie:
Service:
Attention:
Adresse:
Ville: Tél.:
Province: Code postal:

☐ Numéro de la commande (inclure s.v.p.)

☐ Paiement inclus $

PORTEZ À MON COMPTE:

☐ MASTERCARD ☐ VISA ☐ Statistique Canada

N° de compte:

Date d'expiration

☐ Facturez-moi plus tard

Numéro de référence du client:

Signature:

N° au catalogue		Titre	Quantité	Prix	Total
12-519F		**Élaboration et conception des questionnaires d'enquête**		**$25 au Canada $26 autres pays**	

Les chèques ou mandats doivent être établis à l'ordre du Receveur général du Canada/Publication, en dollars canadiens ou l'équivalent.

Canada

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; 1, l).

3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour le première fois.)

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.