Direction des études analytiques

Documents de recherche

Statistique Canada Statistics Canada

Canada

L'APPARIEMENT DE DONNÉES ÉCHANTILLONNALES ET ADMINISTRATIVES EN VUE D'ÉTUDIER LES DÉTERMINANTS DE LA SANTÉ

par

P. David, J.-M. Berthelot, et C. Mustard

N° 58

Groupe de la modélisation et de l'analyse de
l'information sur la santé
Direction des études analytiques
Statistique Canada
1993

Les auteurs assument seuls la responsabilité du contenu de l'analyse, lequel ne représente pas nécessairement les vues ou les politiques de Statistique Canada

Also available in English

L'appariement de données échantillonnales et administratives en vue d'étudier les déterminants de la santé

Pierre David [1], Jean-Marie Berthelot [2], Cam Mustard [3], D.Sc.

RÉSUMÉ

La recherche actuelle dans le domaine de la santé étudie un vaste éventail de facteurs qui influencent la santé des individus et l'utilisation des services de santé. Elle confirme l'existence de la relation entre les facteurs socio-économiques et la santé, tout en améliorant notre compréhension des phénomènes qui sous-tendent cette relation. L'article présente un projet pilote qui réunira pour la première fois au Canada des microdonnées transversales portant sur le niveau de santé et sur les caractéristiques socio-économiques des individus, et des données longitudinales détaillées portant sur l'utilisation de services de santé par un échantillon représentatif de la population d'une province. L'article décrit principalement les méthodes d'appariement probabiliste utilisées pour combiner des données censitaires et administratives.

MOTS CLÉS : Appariement probabiliste; recensement; statut socio-économique; santé; confidentialité.

1. INTRODUCTION

Quelques études ont clairement démontré qu'il existe un lien entre le statut socio-économique d'un individu et la probabilité de son décès au cours d'une période donnée (par exemple, Wolfson et coll. 1993, Marmot 1986, Wilkins et coll. 1991). D'autres études ont démontré que la prévalence de certaines maladies varie beaucoup en fonction des caractéristiques socio-économiques du secteur dans lequel un individu demeure (Anderson et coll. 1993, Dougherty et coll. 1990, Gentleman et coll. 1991). De plus, quelques enquêtes canadiennes ont déjà fourni des données transversales sur l'état de santé et le statut socio-économique des individus, ainsi qu'une information de base sur l'utilisation des services de santé. Cependant, à notre connaissance, il n'existe pas au Canada de base de données longitudinales qui combine une information exhaustive sur la santé, sur l'utilisation de services de santé et sur le statut socio-économique des individus. Ainsi, Statistique Canada et le Manitoba Centre for Health Policy and Evaluation ont conjointement mis sur pied un projet pilote visant à évaluer la possibilité de créer une telle base de données *à partir de données existantes*.

L'objectif premier de ce projet pilote est d'évaluer la faisabilité de réunir les trois sources de données suivantes : le recensement de la population de 1986, l'enquête sur la santé et les limitations d'activités de 1986-1987 (ESLA), et le fichier longitudinal sur les soins de santé de la Manitoba Health Services Commission (MHSC). La base de données résultant de cette

[1] Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, K1A 0T6
[2] Division des études sociales et économiques, Statistique Canada, Ottawa, K1A 0T6
[3] Manitoba Centre for Health Policy and Evaluation, Department of Community Health Sciences, Faculty of Medicine, University of Manitoba

combinaison permettra aux chercheurs d'explorer d'importantes nouvelles avenues en matière de déterminants de la santé.

Le recensement de la population de 1986 fournira une composante socio-économique détaillée qui comprendra des variables telles la composition de la famille, les caractéristiques du logement, le type d'occupation, l'origine ethnique, la langue maternelle, ainsi que plusieurs variables ayant trait au revenu et au niveau d'éducation. L'ESLA de 1986-1987 est une enquête post-censitaire canadienne qui s'adresse aux individus limités par le type ou la quantité d'activités quotidiennes qu'ils peuvent exercer en raison de leur santé. Cet ensemble de données contient de l'information sur la santé et les limitations fonctionnelles des individus ainsi que sur le type d'emploi, le niveau d'éducation, le transport, le logement et les loisirs. L'enquête étant de type auto-déclaré, les données représentent la situation des répondants de leur point de vue plutôt que d'un point de vue administratif ou clinique. Le fichier longitudinal de la MHSC contient de l'information sur les visites à l'hôpital, les diagnostics, les interventions chirurgicales, les soins de santé reçus à la maison, la date et la cause des décès, ainsi que d'autres données sur l'utilisation des soins de santé. De nombreuses études innovatrices en recherche sur les soins de santé ont utilisé ce fichier (par exemple, L.L. Roos et coll. 1987, N.P. Roos et coll. 1987, Shapiro et coll. 1984).

Conformément aux politiques des organismes collaborant à ce projet, certaines procédures ont été entreprises avant d'apparier ces ensembles de données. Elles comprennent des consultations avec le Commissaire à la vie privée du Canada, le Faculty Committee on the Use of Human Subjects in Research de l'université du Manitoba, et le Comité sur la confidentialité et la législation de Statistique Canada. De plus, le Comité d'accès et de confidentialité de la MHSC a été informé du projet.

Suite à ces consultations et selon les politiques formelles de Statistiqe Canada, le ministre responsable de Statistique Canada a autorisé l'appariement tel que proposé : il s'agit d'un projet pilote qui vise à évaluer la faisabilité et l'utilité de l'appariement; le nom et l'adresse des individus ne servira pas à effectuer l'appariement; l'appariement sera effectué entièrement dans les locaux de Statistique Canada par des personnes assermentées par la loi sur la statistique; seul un échantillon de 20,000 individus appariés servira aux fins de recherche et d'analyse; enfin, l'accès aux données finales sera étroitement contrôlé selon les dispositions de la loi sur la statistique. De plus, une lettre d'entente entre Statistique Canada, l'université du Manitoba et le ministère de la santé du Manitoba couvre déjà toute utilisation de la base de données finale.

2. MÉTHODOLOGIE

Le projet d'appariement comporte deux étapes principales. La première étape consiste à assortir par paires les individus appartenant à trois sources de données distinctes. La deuxième étape consiste à choisir un échantillon de 20,000 enregistrements appariés afin de créer la base de données finale. Le système Canlink développé à Statistique Canada a été utilisé pour l'étape d'assortiment. Il s'agit d'un logiciel d'appariement statistique qui assortit les enregistrements de deux ensembles de données en utilisant la puissance discriminante des variables. Le système

pondère les paires d'enregistrements en fonction du degré de concordance des valeurs observées et tient aussi compte de la probabilité de concordances aléatoires. La méthode sous-jacente s'appuie sur la théorie de Fellegi-Sunter (1969). Cette section traite de la méthodologie employée lors de l'étape d'assortiment des fichiers du recensement et du Manitoba. Ces fichiers sont l'échantillon 2B du recensement de 1986 couvrant la province du Manitoba et un sous-ensemble du fichier des personnes inscrites à la MHSC en juin 1986.

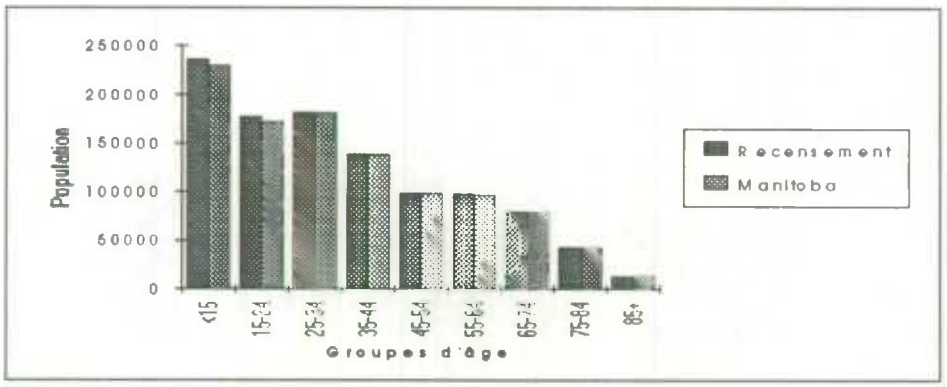
2.1 Ensembles de données

L'échantillon de l'ESLA de 1986-1987 ayant été tiré de l'échantillon 2B du recensement (Dolson et coll. 1987), tous les enregistrements de l'ESLA sont déjà assortis à ceux du recensement. Ainsi, l'étape d'assortiment ne met en jeu que les deux ensembles de données suivants :

1. Un sous-ensemble de variables tirées de l'échantillon 2B du recensement de 1986. Environ vingt pourcent des ménages canadiens sont couverts par le 2B. Ces ménages répondent à la version longue du questionnaire du recensement. Les variables utilisées lors de l'étape d'assortiment sont les suivantes : le code postal résidentiel, le mois et l'année de naissance, le sexe, la taille de la famille, la structure de la famille (adulte seul ou en couple, avec ou sans enfant), le statut familial (petit-enfant, enfant, conjoint marié ou en union libre, parent), la mobilité entre les recensements de 1981 et 1986, et le statut autochtone. Il est à noter que le nom et l'adresse des individus ne sont pas utilisés.

2. Le fichier tiré du registre de la MHSC. Ce fichier couvre tous les résidents du Manitoba qui étaient inscrits au programme universel d'assurance maladie en date du mois de juin 1986. Le fichier contient des informations sur le mois et l'année de naissance, le sexe, la structure familiale et le code postal résidentiel des personnes inscrites. Comme il s'agit d'un fichier longitudinal, on peut l'utiliser pour décrire la mobilité des individus et les changements de structure familiale dans le temps. Le registre est reconnu comme étant équivalent au recensement en termes d'exactitude de l'information qu'il contient à propos de la taille et de la structure de la population (Roos et coll. 1993). Le graphique suivant illustre ceci.

Figure 1. Population selon le recensement de 1986 et le registre du Manitoba



2.2 Formation des paires

Les enregistrements individuels à assortir proviennent donc des fichiers suivants : un dérivé du fichier 2B du recensement de 1986 (contenant les enregistrements de 261,861 individus vivant au Manitoba) et un dérivé du registre administratif du Manitoba (contenant les enregistrements de 1,047,443 personnes). Le nombre de paires d'individus que l'on peut former en prenant une personne de chaque ensemble est le produit de ces deux quantités, c'est-à-dire plus de 274 milliards. L'étape d'assortiment consiste à identifier les bonnes paires d'individus, c'est-à-dire les paires pour lesquelles les enregistrements des deux fichiers correspondent aux mêmes individus. Une fois les bonnes paires identifiées, un échantillon de 20,000 paires sera tiré afin de constituer le noyau de la base de données finale. Les variables analytiques d'intérêt contenues sur les fichiers originaux seront alors extraites et rattachées uniquement à cet échantillon.

La formation et l'évaluation de 274 milliards de paires d'individus seraient très coûteuses. De plus, on sait que l'ensemble de ces paires contiendrait tout au plus 261,861 paires valides, c'est-à-dire moins de 0.0001% de l'ensemble. Ainsi, il serait opérationnellement inefficace de former et d'examiner toutes les paires possibles. La stratégie adoptée pour identifier les bonnes paires consiste à morceller les deux ensembles de données en blocs et à ne former que les paires d'individus appartenant au même bloc.

Après avoir examiné diverses définitions possibles de bloc, nous avons défini un bloc comme étant un ensemble de quatre caractéristiques individuelles, soit le sexe, l'année de naissance, le mois de naissance et le code postal d'une personne. Ceci veut dire que l'on ne forme que les paires d'individus pour lesquels ces quatre variables concordent parfaitement. Ceci a généré un grand nombre de petits blocs, chacun contenant entre 1 et 22 enregistrements. Cette approche qui définit d'abord des blocs et qui examine ensuite les paires potentielles est plus efficace en terme d'assortiment que la méthode qui consiste simplement à évaluer toutes les paires possibles.

2.3 Pondération des paires

Pour chacune des paires, les variables d'assortiment sont comparées une à une, générant des poids proportionnels au degré de concordance entre les valeurs observées. Calculés à partir de probabilités estimées et a priori, ces poids sont généralement élevés dans le cas d'une concordance exacte, faibles dans le cas d'une discordance, et intermédiaires dans le cas d'une concordance partielle (lorsque celle-ci est définie). Ensuite l'addition de tous ces poids, appelés poids de comparaison, donne le poids total d'une paire. Ce poids total est un indicateur de la vraisemblance de la validité d'une paire, c'est-à-dire que le poids total est proportionnel à la probabilité que les enregistrements correspondent au même individu.

Les poids de comparaison sont calculés à partir du rapport des probabilités conditionnelles de concordance et de discordance chez les bonnes et les mauvaises paires (Statistique Canada 1989, David 1992). Pour la variable de comparaison i et le résultat j, le rapport des probabilités conditionnelles est le suivant :

$$(1)\ R_{ij} = \frac{P(résultat\ j \mid bonne\ paire)}{P(résultat\ j \mid mauvaise\ paire)}$$

Comme il est plus pratique d'utiliser des fonctions additives et des nombres entiers, le poids de comparaison pour la variable i et le résultat j est défini comme suit :

$$(2)\ W_{ij} = INT(\ 10 \times LOG_2\ (R_{ij})\ )$$

Le rapport des probabilités pour une paire donnée est égal à la multiplication des rapports R_{ij} pour toutes les variables étant donnés les résultats observés pour cette paire. Par conséquent, le poids total d'une paire est la somme des poids de comparaison étant donnés les résultats observés pour cette paire.

Les concordances qui sont plus probables chez les bonnes paires que chez les mauvaises paires reçoivent des poids positifs puisque le rapport R_{ij} est supérieur à 1 dans ce cas. De plus, par définition, les variables qui présentent un nombre élevé de valeurs possibles ont une puissance discriminante supérieure et génèrent des poids de concordance plus élevés. Le tableau 1, où le dénominateur de R_{ij} est estimé par la probabilité d'observer le résultat j chez les paires formées de façon aléatoire, illustre ce fait. Le numérateur du rapport est estimé de façon itérative à l'aide d'échantillons de paires jugées comme étant de bonnes paires. Les numérateurs présentés dans le tableau 1 ne servent qu'à illustrer le calcul des poids et ne sont pas de véritables estimations des probabilités correspondantes.

Tableau 1. Poids de comparaison pour le sexe et le mois de naissance de l'enfant le plus jeune

	Sexe de l'enfant	Mois de naissance de l'enfant
$R_{iC} = \frac{P(Concordance \mid bonne\ paire)}{P(Concordance \mid mauvaise\ paire)}$	$\frac{0.984}{1/2} = 1.97$	$\frac{0.980}{1/12} = 11.76$
Poids de concordance = $INT\ (10 \times LOG_2\ (R_{iC})\ )$	9	35
$R_{iD} = \frac{P(Discordance \mid bonne\ paire)}{P(Discordance \mid mauvaise\ paire)}$	$\frac{0.016}{1/2} = 0.03$	$\frac{0.020}{11/12} = 0.02$
Poids de discordance = $INT\ (10 \times LOG_2\ (R_{iD})\ )$	-49	-55

Dans le tableau 2, quelques variables illustrent la méthode de pondération, bien qu'en réalité on utilise plus de variables pour pondérer les paires. Dans cet exemple, les deux individus étant mariés, la paire reçoit un poids de concordance pour la variable état civil. La taille de la famille différant d'une unité est un exemple de concordance partielle; ceci engendre un poids moins grand que s'il y avait eu concordance exacte. Ensuite on observe que la concordance des années de naissance du conjoint génère un poids important puisqu'il s'agit d'une variable ayant une forte puissance discriminante. La discordance au niveau du mois de naissance du conjoint entraîne un

poids négatif. Finalement, la somme de ces poids de comparaison donnerait à cette paire fictive un poids total de 25.

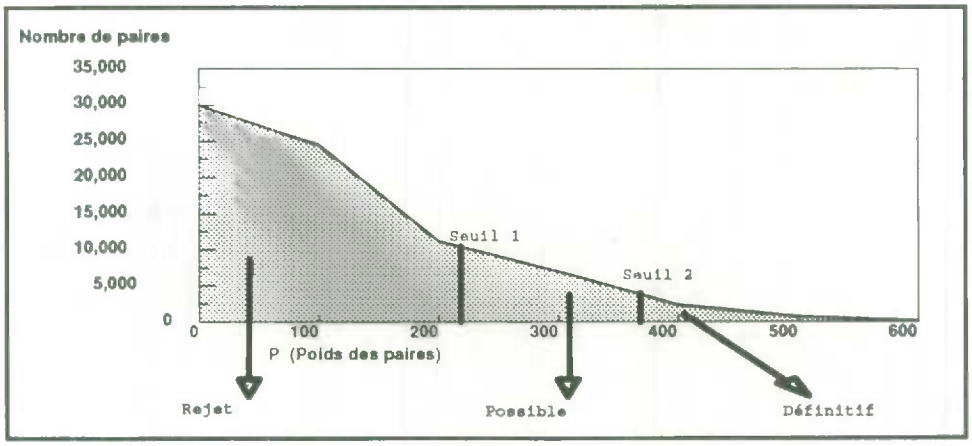
Tableau 2. Exemple de pondération

	État civil	Taille de la famille	Année de naissance du conjoint	Mois de naissance du conjoint	Total
Recensement	marié	4	1956	10	
Manitoba	marié	3	1956	4	
Poids	9	2	45	-31	25

Après avoir examiné le contenu des deux sources de données à apparier, les variables suivantes ont été retenues pour la pondération des paires : l'état civil et le statut autochtone de l'individu; le mois et l'année de naissance du conjoint; le sexe, le mois et l'année de naissance de l'enfant le plus jeune (s'il y a lieu); et finalement, la taille, la structure et la mobilité géographique de la famille.

Une fois que toutes les paires sont pondérées, elles sont classifiées en trois groupes selon leur poids total : un groupe de rejet, un groupe d'acceptation possible et un groupe d'acceptation définitive. Les seuils servant à délimiter ces groupes sont fixés après examen d'un échantillon représentatif de paires pondérées. Ils répartissent généralement les paires en trois groupes relativement homogènes. De plus, si la pondération est adéquate, les paires seront placées en ordre croissant selon la vraisemblance qu'il s'agisse de bonnes paires. La figure 2 illustre un cas fictif mais typique.

Figure 2. Distribution des paires selon leur poids total



Une grande proportion de paires ont un poids inférieur au seuil 1. Ces paires sont formées d'individus qui concordent au niveau des variables de bloc mais qui concordent peu ou pas du tout au niveau des variables de comparaison. La majorité de ces paires ne sont pas bonnes et elles peuvent être rejetées avec confiance. L'objectif principal de cette pondération est d'ailleurs d'éliminer les paires qui sont manifestement mauvaises. C'est le seuil 1 qui joue le rôle le plus important pour accomplir cette tâche : s'il est placé trop bas, de mauvaises paires seront classées comme possibles, s'il est placé trop haut, des paires qui peuvent être bonnes seront rejetées.

Pour un petit nombre de paires, la plupart des variables concordent, indiquant une forte probabilité que les deux individus soient les mêmes. La majorité de ces paires sont clairement bonnes et ont un poids supérieur au seuil 2.

La validité des paires dont le poids se situe entre les seuils 1 et 2 est incertaine. Certaines paires ne sont pas de bonnes paires : il s'agit de personnes qui ont des caractéristiques très semblables mais qui ne sont pas les mêmes. Les autres paires sont bonnes mais contiennent des erreurs qui font que leur poids total n'est pas aussi élevé qu'il devrait. De façon générale, des erreurs dans les données et des différences conceptuelles entre les bases de données rendent souvent deux enregistrements d'un même individu difficiles à reconnaître. Regardons brièvement trois types d'erreurs qui brouillent les données et qui compliquent l'assortiment des individus.

1. L'imprécision du répondant et les erreurs de saisie sont deux cas qui engendrent des données erronées. Par exemple, le répondant rapporte 1954 comme année de naissance au lieu de 1953, ou encore le commis à la saisie inscrit 12 au lieu de 2 comme mois de naissance. Aussi sophistiquées et efficaces que puissent être les techniques de saisie, il est difficile d'éliminer complètement ce type d'erreur qui peut avoir de graves répercussions. Par exemple, une erreur dans une variable de bloc fait qu'un enregistrement ne sera même pas comparé à sa véritable contrepartie, à moins que les deux fichiers ne contiennent la même erreur, ce qui est en général très improbable.

2. Les erreurs de mise à jour surviennent lorsque les données sont recueillies ou corrigées à différents moments. Par exemple, les données du recensement ont été recueillies à une date spécifique (le 3 juin 1986), alors que le registre du Manitoba est généralement mis à jour à tous les six mois. Des dates de référence différentes entraînent inévitablement des écarts entre les données. Par exemple, un individu peut être célibataire sur le fichier du recensement et marié sur celui du Manitoba dans l'éventualité où le mariage aurait eu lieu entre la date du recensement et la mise à jour du registre du Manitoba.

3. Le troisième type d'erreur a trait au cadre conceptuel inhérent aux bases de données à apparier. Par exemple, le recensement et le registre du Manitoba définissent différemment la famille. Même si les familles du recensement ont été remodelées de façon à reproduire le plus fidèlement possible la structure utilisée par le registre du Manitoba, certains écarts peuvent subsister, réduisant ainsi la probabilité de former de bonnes paires.

2.4 Pondération basée sur la fréquence

L'objectif de la pondération basée sur la fréquence est de raffiner le classement des paires en utilisant des poids proportionnels à la rareté de la valeur sur laquelle deux enregistrements concordent. Ce type de pondération est plus coûteux à utiliser que le premier car il associe un poids spécifique à chaque valeur concordante. C'est pourquoi il n'est utilisé qu'une fois les nombreuses mauvaises paires rejetées.

Tableau 3. Exemple de pondération basée sur la fréquence

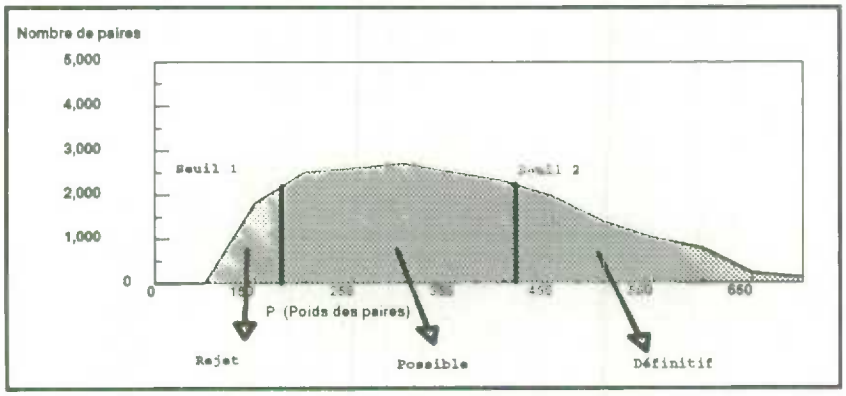
	État civil	Taille de la famille	Année de naissance du conjoint	Mois de naissance du conjoint	Total
Recensement	marié	2	1944	10	
Manitoba	marié	2	1944	10	
Poids	5	20	50	35	110

	État civil	Taille de la famille	Année de naissance du conjoint	Mois de naissance du conjoint	Total
Recensement	marié	10	1899	12	
Manitoba	marié	10	1899	12	
Poids	5	100	170	35	310

Dans le tableau 3, on voit qu'une concordance sur une taille de famille égale à dix reçoit plus de poids qu'une concordance sur une taille de famile égale à deux. On voit aussi qu'une année de naissance rare génère plus de poids qu'une année de naissance plus commune. Par ailleurs, certaines variables dont les valeurs sont jugées relativement équiprobables (par exemple, le mois de naissance du conjoint) recoivent un poids fixe comme lors de la première pondération.

La pondération basée sur la fréquence engendre une nouvelle distribution des paires tel qu'illustré à la figure 3. On fixe alors de nouveaux seuils à l'aide d'un échantillon de paires nouvellement pondérées et on obtient la classification finale des paires.

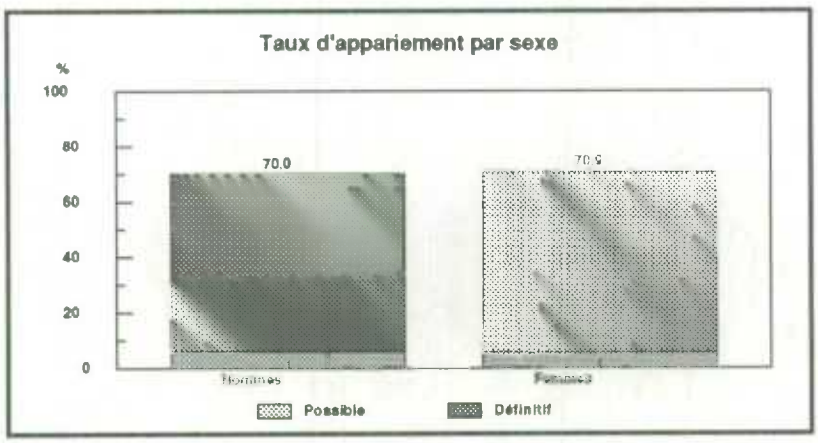
Figure 3. Distribution des paires suite à la pondération basée sur la fréquence



3. RÉSULTATS

Globalement, 70.4% des individus du recensement ont été appariés à un individu sur le fichier du Manitoba. La figure 4 montre qu'il y a peu de différence entre les hommes et les femmes. On voit aussi qu'environ 6% des individus font partie de paires classées comme possibles. Même si ces paires présentent des poids moindres, il se peut que les enregistrements qui constituent ces paires correspondent aux mêmes individus (dans les cas où les faibles poids s'expliquent par des données erronées ou incompatibles). Il se peut aussi qu'il s'agisse d'individus semblables mais différents. Ce taux relativement bas de 6% reflète l'utilisation d'un seuil 1 assez élevé afin de limiter le nombre de paires possibles. Habituellement, cette stratégie engendre le rejet de quelques bonnes paires, mais permet de conserver peu de mauvaises paires. La qualité des paires retenues devrait donc être assez bonne.

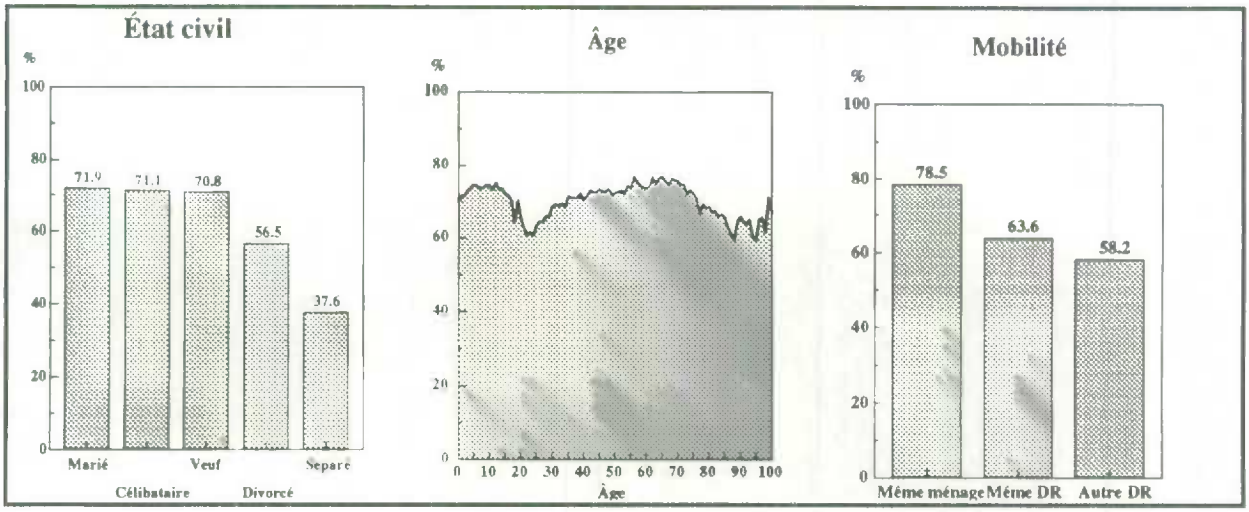
Figure 4. Taux d'appariement du fichier du recensement



3.1 La mobilité

Les facteurs qui ont eu le plus d'influence sur le taux d'appariement sont reliés à la mobilité géographique des individus. Ainsi, les groupes d'individus suivants ont été plus difficiles à apparier : les jeunes adultes (entre 20 et 25 ans), les personnes ayant changé de lieu de résidence entre les recensements de 1981 et 1986, ainsi que les personnes divorcées ou séparées. Chez ces groupes, les changements fréquents d'adresse et de structure familiale rendent la concordance entre les deux sources de données plus difficile que chez les groupes moins mobiles. En effet, comme les données du recensement sont en date du 3 juin 1986 précisément et que les données du Manitoba sont en date du 31 décembre 1986 pour la plupart des variables, un décalage de l'information est plus probable chez les individus mobiles. La figure 5 illustre les taux d'appariement selon quelques unes de ces variables.

Figure 5. Taux d'appariement du fichier du recensement selon diverses variables



DR : division de recensement, une aire géographique utilisée par le recensement. Le territoire du Manitoba compte vingt-trois divisions de recensement.

On observe un faible taux d'appariement chez les personnes séparées. Ceci s'explique par la mobilité inhérente au phénomène de la séparation et aussi par le décalage de l'information entre les deux sources de données.

L'effet de l'âge sur le taux d'appariement n'est pas surprenant. Les enfants de moins de quinze ans et les adultes entre trente et soixante ans connaissent de meilleurs taux étant donné leur situation plus stable. Chez les individus de plus de 85 ans, on trouve davantage de variabilité dans les données à cause du phénomène de l'entrée en institution et du petit nombre de cas.

Chez les individus qui n'ont pas déménagé entre les recensements de 1981 et 1986 (même ménage), on aurait pu s'attendre à un taux d'appariement encore meilleur. Le taux de 78.5% est peut-être une indication qu'il y a un plafond d'appariement d'environ 80% en utilisant la

méthodologie décrite jusqu'ici, étant donné que les fichiers ne sont pas absolument exempts d'erreurs.

3.2 Le code postal

Sur le fichier du recensement, le code postal est en date du 3 juin 1986. À l'origine, 6% des enregistrements n'avaient pas de code postal. Dans ces cas, la Division de la géographie de Statistique Canada a généré des codes postaux à l'aide d'une méthode éprouvée qui tient compte de la relation entre la géographie du recensement et les codes postaux. L'utilisation de ce code postal dérivé a donné de bons résultats en fournissant 4% des appariements.

Sur le fichier du Manitoba, les codes postaux sont en date du 31 décembre de chaque année. Pour apparier les fichiers, le code postal de 1986 a servi de code postal de base. Trois autres codes postaux ont servi d'alternatives : celui de 1985, celui de 1987 et un autre code postal daté de 1986 dans le cas des individus ayant une adresse alternative cette année là. L'utilisation des codes postaux alternatifs sur le fichier du Manitoba a fourni 7% des appariements. En conclusion, les deux méthodes ont été utiles, générant des appariements qu'il aurait été impossible d'obtenir en n'utilisant que les codes postaux initiaux.

3.3 La réconciliation des familles

Après avoir apparié 70.4% des individus du recensement à l'aide de Canlink, nous avons examiné les familles du recensement et du Manitoba pour lesquelles une seule personne n'avait pas été appariée. Lorsque les personnes non appariées des familles correspondantes étaient semblables (âge à 5 ans près et même sexe), on a formé une paire définitive. Cette procédure a ajouté près de 2% d'appariements, conduisant à un taux global de 72.1%.

4. PHASE DEUX

Bien que le taux d'appariement de 72.1% soit passablement bon, les caractéristiques relativement différentes des individus non appariés suggéraient d'essayer une deuxième vague d'appariement. L'analyse de cette deuxième phase n'est pas complétée.

4.1 Travail préliminaire

Pour la deuxième phase de l'appariement, on a inclus tous les individus qui n'avaient pas été appariés et aussi quelques individus dont l'appariement n'était pas complètement satisfaisant. C'est le cas des groupes suivants :

1. Les personnes qui vivent seules et dont l'appariement était classé comme possible (3,390 enregistrements du recensement). Le nombre de variables à comparer étant très limité pour ces gens, il est difficile de juger de la validité des paires.

2. Les familles incomplètes (62,888 enregistrements du recensement). Tous les membres des familles où au moins une personne n'était pas appariée ont été inclus dans la phase deux.

3. Certaines familles complètes (13,076 enregistrements du recensement). Chez les familles du recensement pour lesquelles on a apparié des individus appartenant à plus d'une famille du Manitoba, tous les membres de chaque fichier ont été inclus dans la phase deux.

Lors de la première phase, les quatre variables définissant le bloc (sexe, mois de naissance, année de naissance et code postal) devaient concorder exactement pour qu'un individu soit comparé à sa véritable contrepartie et qu'ils aient la possibilité d'être appariés. Une seule erreur sur le mois de naissance par exemple empêchait la formation de la bonne paire pour cet individu et ne pouvait donc pas générer un appariement valide.

Pour la deuxième phase, la définition des blocs a été élargie afin de former davantage de paires d'individus. L'année et le mois de naissance exacts ont été remplacés par l'âge de la personne, ce qui permet de comparer un individu à un plus grand nombre de candidats. De plus, l'aire couverte par la variable géographique en milieu urbain a été agrandie de deux à trois fois, le secteur de dénombrement du recensement remplaçant le code postal.

Les mêmes variables de comparaison ont été utilisées à l'exception de ces quelques changements : le statut autochtone a été laissé de côté à cause de problèmes de définition entre les deux sources de données; l'année et le mois de naissance ont été utilisées comme variables de comparaison; le code postal a été utilisé comme variable de comparaison en milieu urbain; enfin, la structure familiale a été modifiée afin de rendre les définitions de petit-enfant et d'union libre plus comparables.

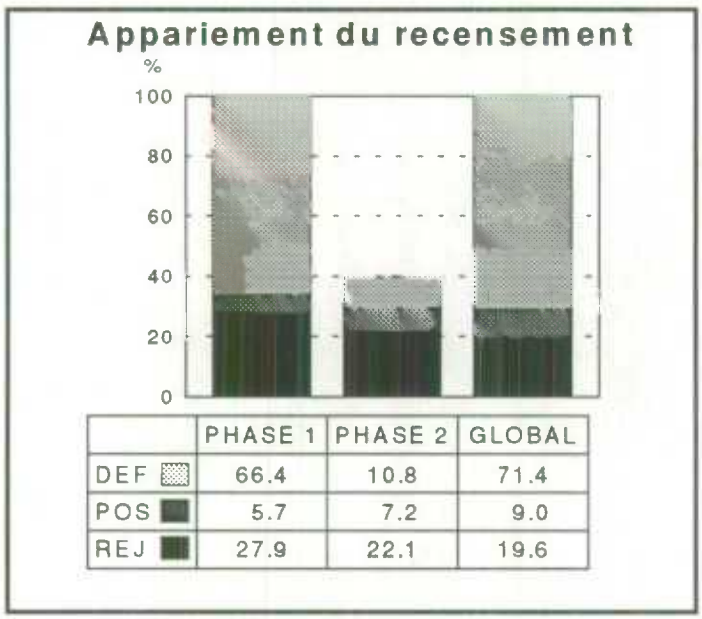
4.2 Résultats

Globalement, 45% des individus du recensement inclus dans cette deuxième vague ont été appariés à un individu sur le fichier du Manitoba. Considérant que les meilleures paires avaient déjà été formées lors de la phase 1 et qu'elles étaient exclues de cette phase-ci, ce taux s'avère satisfaisant.

Parmi les appariements de la phase 2, on remarque une grande proportion de paires possibles en milieu urbain. Tel qu'attendu, on les trouve surtout chez les jeunes adultes célibataires pour qui on a peu d'information à comparer hormis les variables de bloc (c'est-à-dire que la taille de la famille est presque toujours égale à 1, l'état civil est presque toujours «marié», il n'y a pas d'information à comparer à propos d'un conjoint ou d'un enfant, etc.).

La figure 6 montre le taux d'appariement du fichier du recensement lors des phases 1 et 2, ainsi qu'une projection globale incorporant les appariements des deux phases.

Figure 6. Taux d'appariement du fichier du recensement



	PHASE 1	PHASE 2	GLOBAL
DEF	66.4	10.8	71.4
POS	5.7	7.2	9.0
REJ	27.9	22.1	19.6

5. CONCLUSION

En conclusion, la méthodologie présentée dans cet article permet d'apparier environ 80% du fichier du recensement (voir figure 6 : 71.4% + 9.0%) au fichier de l'université du Manitoba en utilisant principalement l'âge, le sexe, le code postal, la taille et la structure de la famille. Il reste quelques ajustements à faire au niveau de l'appariement qui pourraient gonfler ce taux d'un ou deux pourcent. Par exemple, on pourrait compléter l'appariement des familles où un seul individu

n'a pas été apparié (comme à la suite de la phase 1). Ce travail ainsi que l'analyse des appariements de la phase 2 seront effectués sous peu et compléteront l'étude de l'appariement de nos deux sources de données.

Le taux de 80% s'avère satisfaisant en comparaison du taux de réponse typique de plusieurs enquêtes. Par exemple, les taux de réponse de l'enquête sur l'alimentation en Nouvelle-Écosse sont de 79.7% parmi les répondants repérés et de 60.0% pour l'échantillon total (MacLean 1993). L'enquête Manitoba Heart Health a connu pour sa part des taux de réponse de 77.1% parmi les répondants repérés et de 60.8% pour l'échantillon total (Young et coll. 1991).

Évidemment, en considérant les divers types d'erreurs qui peuvent affliger un appariement de grande envergure, on comprend qu'il n'est pas réaliste de s'attendre à un taux de 100%. Des données erronées, des décalages dans la collecte ou la mise à jour de l'information, ainsi que des différences conceptuelles entre les ensembles à apparier influencent inévitablement le taux de succès de tout appariement statistique. Par ailleurs, bien que les individus non appariés présentent des caractéristiques relativement différentes des individus appariés, une information socio-démographique très riche est disponible du recensement à propos de cette population non appariée. Ces renseignements pourront servir à sélectionner un échantillon d'appariements représentatif de la population entière en vue d'effectuer des analyses de qualité.

Parmi les activités à venir, une évaluation de la qualité des appariements permettra de confirmer globalement la validité des paires retenues et notamment de trancher la question des paires «possibles». La méthode d'évaluation prévue consiste à prendre un échantillon de mille ou deux milles paires afin de comparer les noms et adresses sur les deux sources de données. Cette information ne servirait pas à déterminer la validité d'appariements spécifiques, mais seulement à estimer les taux d'appariement véritables à des niveaux agrégés.

Ensuite, on sélectionnera un échantillon de vingt milles individus appariés qui soit représentatif de la population que l'on veut étudier. On ajoutera aux enregistrements appariés les variables d'intérêt portant sur la santé et sur le statut socio-économique. Finalement, il faudra organiser les données en une base unique qui permette d'effectuer des analyses sur les relations entre le statut socio-économique, la santé et l'utilisation des services de santé.

REMERCIEMENTS

Les auteurs tiennent à remercier les personnes suivantes pour leur importante et généreuse contribution à ce travail : Yves Béland, Christian Houle, Sheila Krawchuck et Gurupdesh Pandher, Division des méthodes d'enquêtes sociales, Statistique Canada; John Armstrong et Jackie Mayda, Division des méthodes d'enquêtes-entreprises, Statistique Canada; J. Patrick Nicol, Shelley Derksen et Leonard McWilliam, Manitoba Centre for Health Policy and Evaluation. Pour leur initiative dans ce projet ainsi que pour la confiance et le soutien qu'ils nous ont témoigné, nous tenons aussi à remercier : Michael Wolfson, Direction des études analytiques, Statistique Canada, et Leslie Roos, Université du Manitoba.

BIBLIOGRAPHIE

ANDERSON, G., GRUMBACH, K., LUTT, H., ROOS, L.L., et MUSTARD, C. (1993). Use of coronary artery bypass surgery in the United States and Canada: influence of age and income. *Journal of the American Medical Association*, 269, 1661-1666.

DAVID, P. (1992). Methods for calculating probabilities and weights. Appendice 1 du rapport interne du 17 novembre 1992, Statistique Canada.

DOLSON, D., McCLEAN, K., MORIN, J.-P., et THÉBERGE, A. (1987). Plan d'échantillonnage pour l'enquête sur la santé et les limitations d'activités. *Techniques d'enquête*, 13(1), 101-117.

DOUGHERTY, G., PLESS, I.B., et WILKINS, R. (1990). Social class and the occurence of traffic injuries and death in urban children. *Canadian Journal Of Public Health*, 81, 204-209.

FELLEGI, I.P. et SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

GENTLEMAN, J.F., WILKINS, R., NAIR, C., et BEAULIEU, S. (1991). An analysis of frequencies of surgical procedures in Canada. *Health Reports*, 3(4), 291-309.

MACLEAN, ?.?. (1993). Report of the Nova Scotia Nutrition Survey. Nova Scotia Heart Health Program, Department of Health, Government of Nova Scotia.

MARMOT, M.G. (1986). Social inequalities in mortality: the social environment. Dans *Class and Health, Research and Longitudinal Data*, (Éd. R.G. Wilkinson). London: Tavistock Publications.

ROOS, L.L., MUSTARD, C.A., NICOL, J.P., COMM, B., McLERRAN, D.F., MALENKA, D.J., YOUNG, T.K., et COHEN, M.M. (1993). Registries and administrative data: organization and accuracy. *Medical Care*, 31(3), 201-212.

ROOS, L.L., NICOL, J.P., et CAGEORGE, S.M. (1987). Using administrative data for longitudinal research: comparisons with primary data collection. *Journal of Chronical Diseases*, 40(1), 41-49.

ROOS, N.P., MONTGOMERY, P., et ROOS, L.L. (1987). Health care utilization in the years prior to death. *The Milbank Quarterly*, 65(2), 231-254.

SHAPIRO, E. et ROOS, L.L. (1984). Using health care: rural/urban differences among the Manitoba elderly. *The Gerontologist*, 24(3), 270-274.

Statistique Canada, Division du développement de systèmes (1989). Generalized iterative record linkage system weights.

WILKINS, R., ADAMS, O., et BRANCKER, A. (1991). Changes in mortality by income in urban Canada from 1971 to 1986. *Health Reports*, 1(2), 137-174.

WOLFSON, M.C., ROWE, G., GENTLEMAN, J.F., et TOMIAK, M. (1993). Career earnings and death: a longitudinal analysis of older canadian men. *Journal of Gerontology: Social Sciences*.

YOUNG, T.K., GELSKEY, D.E., MACDONALD, S.M., HOOK, E., et HAMILTON, S (1991). The Manitoba heart health survey: technical report.

DIRECTION DES ÉTUDES ANALYTIQUES
DOCUMENTS DE RECHERCHE

N°

1. *Réaction comportementale dans le contexte d'une simulation micro-analytique socio-économique,* ***Lars osberg***

2. *Chômage et formation,* ***Garnett Picot***

3. *Des pensions aux personnes au foyer et leur répartition sur la durée du cycle de vie,* ***Michael Wolfson***

4. *La modélisation des profils d'emploi des Canadiens au cours de leur existence,* ***Garnett Picot***

5. *Perte d'un emploi et adaption au marché du travail dans l'économie canadienne,* ***Garnett Picot et Ted Wannell***

6. *Système de statistiques relatives à la santé: proposition d'un nouveau cadre théorique visant l'intégration de données relatives à la santé,* ***Michael C. Wolfson***

7. *Projet-pilote de raccordement micro-cacro pour le secteur des ménages au Canada,* ***Hans J. Adler et Michael C. Wolfson***

8. *Notes sur les groupements de société et l'impôt sur le revenu au Canada,* ***Michael C. Wolfson***

9. *L'expansion de la classe moyenne: données canadiennes sur le débat sur la déqualification,* ***John Myles***

10. *La montée des conglomérats,* ***Jorge Niosi***

11. *Analyse énergétique du commerce extérieur canadien: 1971 et 1976,* ***K.E. Hamilton***

12. *Taux nets et bruts de concentration des terres,* ***Ray D. Bollman et Philip Ehrensaft***

13. *Tables de mortalité en l'absence d'une cause pour le Canada (1921 à 1981): une méthode d'analyse de la transition épidémiologique,* ***Dhruva Nagnur et Michael Nagrodski***

14. *Distribution de la fréquence d'occurence des sous-séquences de nucléotides, d'après leur capacité de chevauchement,* ***Jane F. Gentleman et Ronald C. Mullin***

15. *L'immigration et le caractère ethnolinguistique du Canada et du Quebec,* ***Réjean Lachapelle***

16. *Intégration de la ferme au marché extérieur et travail hors ferme des membres des ménage agricoles,* ***Ray D. Bollman et Pamela Smith***

17. *Les salaires et les emplois au cours des années 1980: évolutin des salaires des jeunes et déclin de la classe moyenne,* ***J. Myles, G. Picot et T. Wannell***

18. *Profil des exploitants agricoles dotés d'un ordinateur,* ***Ray D. Bollman***

19. *Répartitions des risques de mortalité: une analyse de tables de mortalité,* ***Geoff Rowe***

20. *La classification par industrie dans le recensement canadien des manufactures: vérification automatisée à l'aide des données sur les produits,* ***John S. Crysdale***

21. *Consommation, revenus et retraite,* ***A.L. Robb et J.B. Burbridge***

22. *Le renouvellement des emplois dans le secteur manufacturier au Canada,* ***John R. Baldwin et Paul K. Gorecki***

23. *La Dynamique des marchés concurrentiels,* ***John R. Baldwin et Paul K. Gorecki***

    A. *Entrée et sortie d'entreprises dans le secteur manufacturier au Canada*
    B. *Mobilité à l'intérieur des branches d'activité dans le secteur manufacturier au Canada*
    C. *Mesure de l'entrée et de la sortie dans le secteur manufacturier au Canada: méthodologie*
    D. *Effet de la libre concurrence sur la productivité: rôle de la rotation des entreprises et des usines*
    E. *Les fusions et le processur concurrentiel*
    F. *À venir*
    G. *Lews statistiques de concentration comme prédicteurs du degré de concurrence*
    H. *Le rapport entre la mobilité et la concentration dans le secteur manufacturier au Canada*

24. *Améliorations apportées au SAS de l'ordinateur central en vue de faciliter l'analyse exploratoire des données,* ***Richard Johnson et Jane F. Gentleman***

25. *Aspects de l'évolution du marché du travail au Canada: mutations intersectorielles et roulement de la main-d'oeuvre,* ***John R. Baldwin et Paul K. Gorecki***

26. *L'écart persistant: étude de la différence dans les gains des hommes et des femmes qui ont récemment reçu un diplôme d'études postsecondaires,* ***Ted Wannell***

27. *Estimation des pertes de sol sur les terres agricoles à partir des données du recensement de l'agriculture sur les superficies cultivées,* ***Douglas F. Trant***

28. *Les bons et les mauvais emplois et le déclin de la classe moyenne: 1967-1986,* ***Garnett Picot, John Myles, et Ted Wannell***

29. *Données longitudinales sur la carrière relatives à certaines cohortes de fonctinnaires,* ***Garnett Picot et Ted Wannell***

30. *L'incidence des revenus sur la mortalité sur une période de vingt-cinq ans,* ***Michael Wolfson, Geoff Rowe, Jane F. Gentleman et Monica Tomiak***

31. *Réaction des entreprises à l'incertitude des prix: la stabilisation tripartite et l'industrie des bovins dans l'ouest du Canada,* ***Theodore M. Horbulyk***

32. *Méthodes de lissage pour microdonnées longitudinales simulées,* ***Jane F. Gentleman, Dale Robertson et Monica Tomiak***

33. *Tendances des investissements directs canadiens à l'étranger,* ***Paul K. Gorecki***

34. *POHEM - une approche inédite pour l'estimation de l'espérance de vie corrigée en fonction de l'état de santé,* ***Michael C. Wolfson***

35. *Emploi et taille des entreprises au Canada: les petites entreprises offrent-elles des salaires inférieurs?,* ***René Morissette***

36. *Distinguer les caractéristiques des acquisitions étrangères en haute technologie dans le secteur manufacturier canadien,* ***John R. Baldwin et Paul K. Gorecki***

37. *Efficience des branches d'activité et roulement des établissements dans le secteur canadien de la fabrication,* ***John R. Baldwin***

38. *Le vieillissement de la génération du baby boom: effets sur le secteur public du Canada,* ***Brian B. Murphy et Michael C. Wolfson***

39. *Tendances dans la répartition de l'emploi selon la taille des employeurs: données canadiennes,récentes,* ***Ted Wannell***

40. *Les petites collectivités du Canada atlantique: structure industrielle et caractéristiques du marchédu travail au début des années 80,* ***Garnett Picot et John Heath***

41. *La répartition des impôts et des transferts fédéraux et provinciaux dans le Canada rural,* ***Brian B. Murphy***

42. *Les multinationales étrangères et les fusions au Canada,* ***John Baldwin et Richard Caves***

43. *Recours répétés à l'assurance-chômage,* ***Miles Corak***

44. *POHEM -- Un cadre permettant d'expliquer et de modéliser la santé de populations humaines,* ***Michael C. Wolfson***

45. *Analyse de modèle de l'espérance de vie en santé de la population: une approche fondée sur la microsimulation,* ***Michael C. Wolfson et Kenneth G. Manton***

46. *Revenue de carrière et décès: une analyse longitudinale de la population âgée masculine du Canada,* ***Michael C. Wolfson, Geoff Rowe, Jane Gentleman et Monica Tomiak***

47. *La modélisation des profils d'emploi des canadiens au cours de leur existence,* ***Miles Corak***

48. *La dynamique du mouvement des entreprises et le processus concurrentiel,* ***John Baldwin***

49. *Élaboration de données-panel longitudinales à partir de registres des entreprises: Observations du Canada,* ***John Baldwin, Richard Dupuy et William Penner***

50. *Le calcul de l'espérance de vie ajustée sur la santé pour une province canadienne à l'aide d'une fonction d'utilité multiattibut: Un premier essai,* ***J.-M. Berthelot, R. Roberge et M. C. Wolfson***

51. *Mesure de la robustesse des barrières à l'entrée,* ***J. R. Baldwin et M. Rafiquzzaman***

52. *Les multinationales au Canada : Caractéristiques et facteurs déterminants,* ***Paul K. Gorecki***

53. *La persistance du chômage : Dans quelle mesure l'attribuer aux presttions d'assurance-chômage de prolongation fondée sur le taux de chômage régional,* ***Miles Corak et Stephen Jones***

54. *Variations cycliques de la durée des périodes de chômage,* ***Miles Corak***

55. *Licenciements et travailleurs déplacés: Variations cycliques, secteurs les plus touchés et expériences après le licenciement,* ***Garnett Picot et Wendy Pyper***

56. *La durée du chômage en période d'expansion et de récession,* ***Miles Corak***

57. *Obtenir un emploi en 1989-1990 au Canada,* ***René Morissette***

58. *L'appariement de données échantillonnales et administratives en vue d'étudier les déterminants de la santé,* ***P. David, J.-M. Berthelot et C. Mustard***

Pour de plus amples renseignements, s'adresser au Président, Comité d'études des publications, Direction des études analytiques, Édifice, R.H. Coats, 24ième étage, Statistique Canada, Parc Tunney, Ottawa, Ontario, K1A 0T6, (613) 951-8213.

Barcode on previous page.