LES UTILISATIONS STATISTIQUES DES DONNÉES ADMINISTRATIVES

Recueil

SYMPOSIUM INTERNATIONAL

23-25 NOVEMBRE, 1987

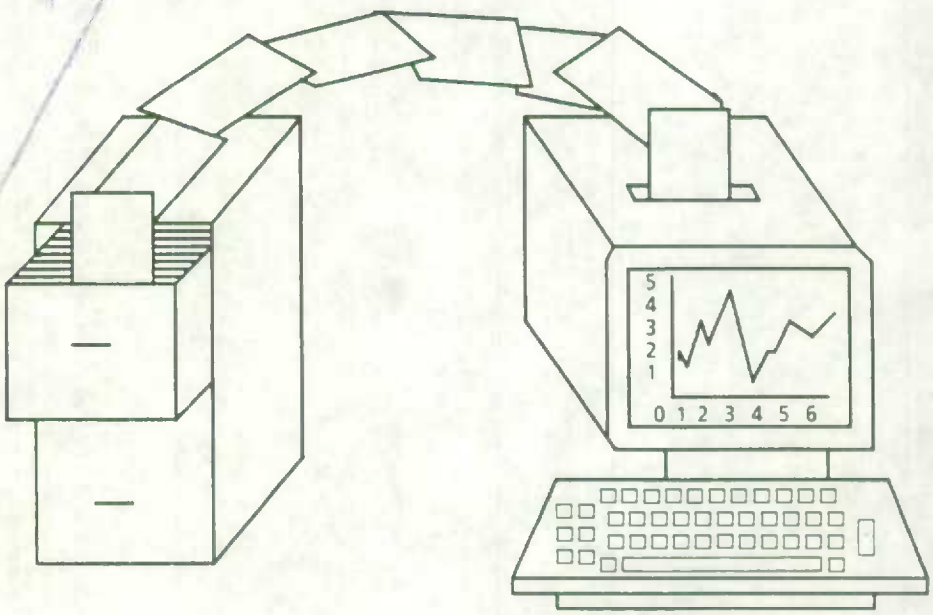
Organisé par Statistique Canada

Statistique Canada Statistics Canada

Canada

LES UTILISATIONS STATISTIQUES DES DONNÉES ADMINISTRATIVES UN SYMPOSIUM INTERNATIONAL

RÉDIGÉ PAR
J.W. COOMBS et M.P. SINGH



STATISTIQUE CANADA
DECEMBRE 1988
OTTAWA

PRIX: Canada $35.00
Autres pays: $42.00
Paiement en dollars canadiens ou l'équivalent

This publication is available in English

PREFACE

Le Symposium sur les utilisations statistiques des données administratives était le quatrième d'une série de symposiums annuels sur les questions méthodologiques organisés par Statistique Canada. Les symposiums antérieurs ont porté sur l'analyse des données d'enquête (1984), la statistique des petites régions (1985) et le méthodologie des données manquantes dans les enquêtes (1986).

Dans de nombreux pays, les organismes statistiques sont confrontés au défi de fournir des données sur une base plus fréquente et plus détaillée pour répondre à la demande d'une part et de supporter un fardeau de réponses plus lourd et des contraintes budgétaires accrues d'autre part. Ceci explique l'utilisation plus courante de données administratives dans le contexte statistique. Le symposium visait à permettre aux spécialistes des secteurs universitaire, privé et public du Canada, des États-Unis et d'autres pays, de se réunir et de partager leur expérience. Plus de trois cents personnes ont participé à la conférence. On y a présenté six communications sollicitées, un débat de spécialistes, et quatre communications offertes. Le présent document renferme une version révisée de diverses communications soumises.

Le présent document est structuré par séances en dix sections. On retrouve à la Section I les expériences et les préoccupations organisationnelles qui se rattachent à l'utilisation accrue des données administratives par divers pays. On y discute de questions et de défis techniques et administratifs, de problèmes relatifs à la confidentialité et à la protection des données, et de la possibilité de mener des recensements de la population basés sur des registres. On y passe également en revue les tendances et les défis récents liés aux divers facteurs qui influent sur les progrès en ce sens.

Dans les cinq communications de la Section II, on fait état de plusieurs façons d'améliorer le modèle Fellegi-Sunter, considéré comme le modèle principal de couplage des données, et de ses applications dans la recherche transversale et longitudinale dans le contexte des registres des entreprises et de la recherche

démographique. La Section III renferme des communications portant sur l'utilisation des données administratives à diverses étapes des opérations d'enquête, notamment la construction de la base de sondage, l'ajout ou le remplacement des données d'enquête, l'amélioration des méthodes d'estimation, l'analyse et l'évaluation des données et l'amélioration des bases de données grâce à l'utilisation combinée de données découlant de plusieurs enquêtes et fichiers administratifs connexes.

Les quatre communications de la Section IV portent sur la comparaison des avantages et des inconvénients des divers fichiers administratifs en ce qui a trait à la production de données statistiques et leur utilisation en vue de l'évaluation ou de l'amélioration de la qualité des données d'enquête. À la Section V, on fait état des expériences et des études ayant trait à l'utilisation des données administratives dans les recensements de la population, des conséquences des modifications des fichiers de données administratives sur les programmes permanents et la pertinence de ces données.

Les Sections VI à IX renferment des communications offertes et traitent d'un éventail de questions, notamment l'utilisation des données administratives dans le contexte de l'estimation des statistiques des petites régions et des composantes des variations démographiques, de l'ajustement en fonction de la non-réponse, de la création d'une base de données dynamiques à l'intention des étudiants, de la détermination et de la mise à jour de la probabilité de sélection, pour des entités économiques, de la production et de l'analyse de statistiques criminelles et médicales et du programme de contrôle de la qualité.

Enfin, la Section X renferme les contributions de quatre spécialistes qui ont participé au débat. Cette séance traite des défis qu'il faudra relever en ce qui a trait aux faits nouveaux dans le domaine de la méthodologie, à la protection de la vie privée, à la pertinence et à la qualité des données.

Ce symposium a été magnifiquement orchestré par J.W. Coombs et M.P. Singh (présidents), N. Kopustas, F. Mayda et C. Patrick, de même que plusieurs personnes ressources de Statistique Canada.

Le contenu des communications qui figurent dans le présent rapport relève de la responsabilité des auteurs. Aucune tentative systématique de révision des communications n'a été faite. Des contrôles ont toutefois été effectués à l'étape du regroupement des communications pour relever les omissions flagrantes. Tout problème majeur détecté a été porté à la connaissance de l'auteur. On a laissé aux rédacteurs le privilège d'effectuer des remaniements mineurs du texte. Le mode de présentation des communications est généralement conforme à la disposition et au style de **Techniques d'enquête,** une publication de Statistique Canada qui traîte de l'élaboration et des applications des méthodes d'enquête.

Remerciements: Les rédacteurs voudraient remercier Judy Clarke, Josée Dufresne, Dula Edirisinghe, Lucie Gagné, Myra Kent et Christine Larabie pour leur apport précieux à la préparation du document. Il faut également remercier les auteurs et leur personnel de soutien qui ont préparé les manuscrits selon le mode de présentation requis et qui nous ont fourni leurs disquettes.

Nous tenons à remercier tout particulièrement Christine Larabie et Frank Mayda pour leur persévérance et leur collaboration.

Gordon J. Brackstone
Statisticien en chef adjoint
Statistique Canada

SYMPOSIUM SUR LES UTILISATIONS STATISTIQUES

DES DONNÉES ADMINISTRATIVES

TABLE DE MATIÈRES PAGE

SESSION III: Approches intégrées à l'élaboration des données
Président: B. Petrie, Statistique Canada

SESSION IV: Évaluation de la qualité
Présidente: N.P. Gendreau, Bureau de la Statistique du Québec

SESSION V: Les dossiers administratifs comme autre source de données
Président: F. Scheuren, U.S. Internal Revenue Service

COMMUNICATIONS OFFERTES

SESSION VI: Président: M.P. Singh, Statistique Canada

SESSION VII: Président: Daniel Kazprzyk, U.S. Bureau of the Census

SESSION VIII: Président: John Coombs, Statistique Canada

SESSION IX: Président: Geoff Lee, Australian Bureau of Statistics

SESSION X: Discussion en panel
Président: G.J. Brackstone, Statistique Canada

ALLOCUTION D'OUVERTURE

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

ALLOCUTION D'OUVERTURE

I.P. FELLEGI[1]

Au nom de Statistique Canada, je voudrais vous souhaiter la bienvenue au symposium international sur les utilisations statistiques des données administratives, le quatrième symposium annuel parrainé par Statistique Canada portant sur les questions méthodologiques.

Le présent symposium abordera l'étude des développements, des défis et des questions se rapportant à l'acquisition et à l'utilisation grandissante des dossiers administratifs à des fins statistiques. Je suis heureux de constater la présence de participants du milieu universitaire et des secteurs privés et publics qui représentent le Canada, les États-Unis et plusieurs autres pays.

Je suis convaincu que les renseignements et les idées qui seront échangés au cours des trois prochains jours seront profitables pour les observateurs ainsi que pour les participants.

Le système statistique canadien repose beaucoup sur l'utilisation de dossiers administratifs pour produire des statistiques. En ce qui concerne un certain nombre de dossiers administratifs, la principale raison historique de leur utilisation, et celle qui domine encore aujourdhui, est qu'ils constituent pratiquement la seule source de statistiques de base. Ainsi, les statistiques des importations et des exportations reposent sur les déclarations douanières, et les statistiques démographiques et de mortalité reposent sur les registres des naissances, des mariages et des décès.

Même lorsque les dossiers administratifs ne constituent pas la seule source d'information, ils complètent souvent les autres sources de données d'une façon efficace. Par exemple, l'information au sujet des soins de santé qui est obtenue des dossiers des hôpitaux est grandement enrichie et non remplacée par l'information sur la santé et les maladies de la population en général qui provient des enquêtes sur les ménages.

Au cours des dernières années, l'usage des dossiers administratifs par Statistique Canada a connu une augmentation considérable, de telle sorte que dans certains cas les fichiers administratifs ont remplacé les enquêtes. Il y a plusieurs raisons importantes derrière cette évolution. L'un des premiers motifs pour l'usage accru des dossiers administratifs a été le besoin de réduire le fardeau de réponse, particulièrement en ce qui concerne les petites entreprises. Les résultats de la réduction du fardeau de réponse sont impressionnants et découlent, dans une grande mesure, de l'accessibilité aux dossiers sur les impôts sur le revenu par Statistique Canada. En fait, de 1978 à 1985, le fardeau de réponse globale des entreprises a été réduit de moitié, et aujourd'hui, 85 pour cent des petits détaillants n'ont plus à compléter de questionnaires annuels de Statistique Canada. Les données sont entièrement obtenues des dossiers fiscaux.

Une autre raison importante qui justifie l'utilisation des dossiers administratifs à Statistique Canada est qu'ils on été créés pour d'autres usages mais peuvent être utilisés à des fins statistiques à un coût marginal, ce qu'il est important de prendre en considération en tout temps, mais plus particulièrement en période de compressions budgétaires. L'exploitation judicieuse des dossiers administratifs, l'accent étant mis à la fois sur "exploitation" et sur "judicieuse", est l'un des moyens qui nous a permis, depuis 1975, de supporter une réduction de notre budget de près de 30 pour cent tout en restant capable de maintenir notre production, et même de l'augmenter dans certains secteurs importants.

Une troisième raison d'importance, cette dernière n'ayant rien à voir avec l'ordre de présentation, qui explique le fait que nous dépendons de plus en plus des dossiers

[1] I.P. Fellegi, Statisticien en chef, Statistique Canada, Parc Tunney's 26-A. Édifice R.H. Coats, Ottawa, Ontario. K1A 0T6

administratifs, est qu'ils ont la propriété de fournir la base nécessaire à la seule méthode rentable de maintenir une liste des entreprises qui soit à la fois exhaustive et sans double compte laquelle à son tour constitue la pierre angulaire de nos statistiques économiques. En effet, il est juste d'affirmer que sans l'apport des dossiers sur les déductions fiscales et sur les retenues sur la paie, un programme de statistiques économiques d'excellente qualité, bien intégré et rentable, ne pourrait pas survivre.

Un quatrième motif qui entraine notre dépendance croissante des dossiers administratifs est le fait que nos clients manifestent beaucoup d'intérêt à l'égard des détails infra-provinciaux de nos résultats statistiques. Nous avons institué une division qui s'occupe de l'élaboration de ce genre de renseignements, et une fois de plus, les dossiers administratifs sont la source de leur travail.

Enfin, les secteurs sociaux, de la santé et certaines politiques économiques exigent la production de quantités croissantes de données longitudinales. Il est à la fois difficile et coûteux de répondre à ce besoin au moyen d'enquêtes directes. Même si nous disposons de quelques-unes de ces enquêtes, notre dépendance envers les dossiers administratifs pour obtenir ce genre d'information est disproportionnée. L'exploitation de ces dossiers entraîne souvent leur couplage.

Malgré les progrès considérables qui ont été réalisés depuis le début des années 70 dans l'utilisation des dossiers administratifs à des fins statistiques il reste un certain nombre de questions et de défis qui ont sans aucun doute empêché d'atteindre l'utilisation optimale des dossiers administratifs. Par exemple:

1. Les dossiers administratifs sont souvent couverts par de multiples juridictions donc les problèmes de qualité, de normalisation, de comparabilité, de cohérence, de couverture, etc. sont forcément complexes;
2. Les données administratives sont recueillies à des fins très précises; elles ne peuvent donc répondre adéquatement à d'autres besoins statistiques. Il est important d'encourager l'esprit de coopération entre les administrateurs et les statisticiens, grâce à une compréhension mutuelle des problèmes, des objectifs et des contributions de chacun;
3. Toutes les données administratives ne sont pas disponibles sur un support exploitable par une machine, ce qui est indispensable pour les traiter à l'aide d'une système informatique;
4. Nous sommes conscients de la façon négative dont le public perçoit l'utilisation des renseignements confidentiels contenus dans les dossiers administratifs. Les statisticiens sont confrontés à l'importante tâche de convaincre le public que la confidentialité de ces renseignements est parfaitement protégée et que nous faisons preuve de rigueur et de prudence en ce qui concerne ces préoccupations tout à fait naturelles : nous ne procédons à la combinaison de dossiers que pour produire des résultats statistiques, un type d'utilisation qui devrait être clairement distinct de l'application d'autres "expéditions de repêchage" administratif. Même lorsqu'il s'agit d'utilisation à des fins statistiques nous devrions disposer de procédures de révision rigoureuses et vérifiables afin de nous assurer que nous ne combinons ces données que lorsque le bien public qui résulte de ces nouvelles informations statistiques l'emporte manifestment sur l'intrusion dans la vie privée que leur création entraîne.

Nous avons, de façon individuelle et collective, accumulé une quantité considérable d'information et acquis une grande expérience en ce qui a trait à l'utilisation des données administratives. Le temps est venu de discuter de leurs utilisations et des problèmes connexes dans le cadre d'un collogue. Si l'on en juge par la qualité des articles et le grand nombre de participants qui ont choisi de prendre part à ce congrés, nombreux sont ceux qui appuient l'initiative. Je suis certain que le personnel de Statistique Canada profitera de la tenue de ce symposium, grâce à la stimulation qu'il tire de ces rencontres, à l'échange de renseignements sur des expériences pertinentes et, peut-être, à l'acquisition de quelques nouvelles idées importantes.

Je souhaite à tous que cet événement soit productif et agréable.

SESSION I: COMMUNICATIONS SOLLICITÉES

THÈMES DE POLITIQUE ET EXPÉRIENCE ORGANISATIONNELLE

Président: J. Ryten, Statistique Canada

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

UTILISATIONS STATISTIQUES DES DONNÉES ADMINISTRATIVES: QUESTIONS ET DÉFIS

G.J. BRACKSTONE[1]

RÉSUMÉ

Le système statistique du Canada utilise dans une large mesure les dossiers administratifs pour produire sur une base régulière des statistiques nationales et infra-nationales. Après avoir illustré cette dépendance, on examine les thèmes et défis qui entourent cinq aspects différents de ce sujet. D'abord, une revue des différentes utilisations possibles des dossiers administratifs précède une discussion pour déterminer si des restrictions de politique sur l'utilisation statistique des données administratives devraient exister. Ensuite, on étudie les questions du droit d'accès aux dossiers administratifs. Puis le document traite de questions se rattachant à la qualité ou à la validité des dossiers administratifs à des fins statistiques et leur rôle final dans le système statistique. La question de savoir de quelle façon les organismes statistiques peuvent ou doivent influencer l'utilité statistique des dossiers administratifs est ensuite examinée. Enfin, on aborde quelques questions importantes se rattachant à la perception du public face à l'utilisation des dossiers administratifs par les organismes statistiques et à ses préoccupations à propos du respect de la vie privée et de la confidentialité.

1. INTRODUCTION

L'utilisation de données administratives à des fins statistiques est un sujet en vogue à l'heure actuelle. Qu'est-ce qui a suscité ce renouvellement d'intérêt? Il s'agit effectivement d'un "renouvellement" puisque les statistiques officielles étaient à l'origine puisées dans les systèmes administratifs. Le mot recensement vient de son utilisation originale d'imposition, alors qu'une bonne partie des statistiques officielles recueillies dépendaient des données obtenues de processus administratifs. Les recensements statistiques ont été grandement utilisés depuis le début du 19e siècle, mais les enquêtes-échantillons ne le sont que depuis la Seconde Guerre mondiale. Les ouvrages intéressants de Hansen, Dalenius et Tepping (1985) et de Hansen (1987) traitent de l'évolution des enquêtes-échantillons et de leur acceptation en tant qu'instrument de statistiques officielles, alors que Bjerve (1985) raconte l'historique de l'utilisation des données administratives. Pourquoi les statisticiens redécouvrent-ils aujourd'hui les dossiers administratifs?

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

L'automatisation de nombreux programmes administratifs au cours des années 60 et 70 a entraîné une utilisation accrue des documents administratifs obtenus à des fins statistiques. Cependant, trois facteurs principaux ont suscité l'intérêt manifesté récemment pour les dossiers administratifs. Tout d'abord, les compressions budgétaires poussent à trouver d'autres moyens de recueillir les données afin de remplacer les enquêtes statistiques et les recensements qui sont relativement coûteux. Deuxièmement, la préoccupation de plus en plus grande de réduire le fardeau de réponse des répondants incite à trouver des solutions de rechange. Troisièmement, l'augmentation de la demande de données régionales, données qui ne peuvent pas être recueillies à partir d'enquêtes-échantillons, justifie la consultation de dossiers administratifs comme source de référence.

Les progrès technologiques, qui ont facilité la manipulation d'importants fichiers administratifs et qui en ont réduit les coûts au cours des dernières années, ont également encouragé l'utilisation accrue de dossiers administratifs.

Dans le présent document, nous examinons d'abord les caractéristiques distinctives des dossiers administratifs (section 2) et nous les classons selon leur objet premier (section 3). De nombreux documents présentés au symposium traitent des différentes utilisations des dossiers administratifs. La section 4 donne un aperçu de ces méthodes. La section 5 porte sur les caractéristiques de qualité, un critère important qui détermine l'utilisation des dossiers administratifs. L'accès aux dossiers administratifs et les façons d'en accroître utilité à des fins statistiques sont discutés à la section 6. Les questions importantes concernant les perceptions du public par rapport à l'utilisation des dossiers administratifs et ses préoccupations sont traitées à la section 7.

2. DÉFINITION DES DOSSIERS ADMINISTRATIFS

Il est utile de s'arrêter un moment sur le sens exact de "données administratives". Ce terme est couramment utilisé en statistique (et c'est aussi la base pour le titre du symposium), mais il est rarement défini. On sait généralement que la distinction entre les données recueillies à des fins administratives et à des fins statistiques est très importante du point de vue des renseignements personnels et de la confidentialité. Par exemple, le comité spécial des renseignements personnels et de la confidentialité de l'American Statistical Association apporte la précision suivante dans son rapport (ASA, 1977):

> "Un dossier administratif est recueilli et maintenu afin de prendre les dispositions par rapport à une personne ou une autre entité ou de contrôler ses actions ... Par ailleurs, le dossier statistique a un but entièrement différent; il vise à connaître les dimensions, les tendances et les relations entre les groupes de personnes ou d'autres entités. L'identification individuelle d'un dossier statistique et de son contenu est gardée confidentielle, à l'exception des personnes qui s'occupent de la collecte et du rassemblement des données agrégées. Le dossier d'une personne ne sert pas à déterminer toute mesure qui touche cette dernière, sauf par la contribution de ce dossier aux agrégats, moyennes ou mesures de relations en statistique. L'essence même de l'analyse statistique réside dans le fait que l'identité des unités individuelles est immatérielle. Les personnes ne doivent pas être identifiables dans les données d'un système statistique."

Cette distinction entre la raison d'être des documents administratifs et des documents statistiques semble claire. Cependant, lorsqu'il s'agit de cataloguer les utilisations des dossiers administratifs à des fins statistiques au sein d'un organisme de statistique, certaines ambiguïtés subsistent. Nous reconnaissons clairement que des sources telles que les fichiers de déclarations d'impôt et les actes de naissance constituent des dossiers

administratifs qui sont ensuite traités comme des dossiers statistiques à des fins d'agrégation et d'analyse. Il n'en est pas ainsi dans tous les cas.

Le statisticien peut avoir accès aux dossiers administratifs de plusieurs façons différentes, par exemple:

(a) Les dossiers administratifs peuvent être recueillis par l'organisme administratif et fournis à l'organisme statistique comme un fichier de dossiers individuels;

(b) Les dossiers administratifs peuvent être recueillis par l'organisme administratif et fournis à l'organisme statistique sous une forme agrégée;

(c) L'organisme statistique peut mener une enquête (ou un recensement) des organismes administratifs locaux (par exemple les municipalités, les commissions scolaires) afin de recueillir des données concernant les unités individuelles (par exemple les logements, les étudiants) qu'ils administrent. Ces données peuvent être fournies sous forme de dossiers individuels ou agrégées pour chaque secteur de compétence local.

On peut immédiatement dire que les catégories (a) et (b) sont des exemples de dossiers utilisés à des fins administratives, alors que la catégorie (c) n'est pas clairement identifiable. Par un processus administratif, des données des unités individuelles sont recueillies, alors que par un processus statistique (enquête ou recensement), des données sont fournies à un organisme statistique. Si les dossiers des unités individuelles se rendent jusqu'à l'organisme statistique, on peut considérer que cela constitue une utilisation statistique des dossiers administratifs à des fins statistiques. Cependant, s'ils arrivent sous forme agrégée, en quoi diffèrent-ils, par exemple, d'une enquête des entreprises concernant leur main-d'oeuvre, données qui sont probablement rassemblées dans l'entreprise par un processus administratif? L'élément-clé porterait-il sur la distinction entre les microdonnées ou les données agrégées? Le fait que l'organisme administratif soit du secteur public ou non a-t-il une importance?

En ce qui a trait à l'utilisation statistique, il est peut-être moins important d'avoir une définition rigide que de bien comprendre les caractéristiques qui distinguent les données administratives des données statistiques. Nous suggérons les caractéristiques suivantes:

(i) l'agent qui fournit les données à l'organisme statistique et l'unité à laquelle elles se rapportent sont différentes (contrairement à la plupart des enquêtes statistiques);

(ii) les données étaient à l'origine recueillies à des fins précises et non statistiques, ce qui peut influer sur le traitement de l'unité de référence;

(iii) l'objectif est une couverture complète (100%) de la population visée;

(iv) le contrôle des méthodes de collecte et de traitement des données administratives relève de l'organisme administratif.

Chacune de ces caractéristiques touche celles des dossiers administratifs et a une incidence sur la façon dont les dossiers administratifs sont utilisés dans un système statistique.

3. TYPES DE DOSSIERS ADMINISTRATIFS

L'absence d'une définition explicite des dossiers administratifs ne nous empêche pas de les classer selon leur but premier. La raison pour laquelle les dossiers administratifs étaient à l'origine constitués peut avoir d'importantes répercussions sur leur utilité à des

fins statistiques, en ce qui concerne leur taux de couverture leur contenu et l'exactitude des données particulières qu'ils contiennent. Voici les principales catégories que l'on peut distinguer:

(i) Les dossiers tenus pour contrôler la circulation des biens et des personnes aux frontières (par exemple les importations, les exportations, l'immigration)

(ii) Les dossiers tenus en vertu de l'obligation légale d'enregistrer certains événements (par exemple la statistique de l'état civil, les constitutions d'entreprises en société, l'octroi de licences)

(iii) Les dossiers nécessaires à la prestation d'avantages ou à l'administration d'engagements (par exemple les impôts, l'assurance-chômage, l'assurance-maladie, les allocations familiales)

(iv) Les dossiers nécessaires à l'administration des établissements publics (par exemple les écoles, les hôpitaux, les prisons)

(v) Les dossiers qui découlent de la réglementation de certaines activités économiques par le gouvernement (par exemple le transport, les affaires bancaires, la radio-diffusion et la télédiffusion)

(vi) Les dossiers reliés à la prestation de services publics (par exemple l'électricité, le téléphone, l'eau).

Il existe une autre distinction entre les documents administratifs tenus au niveau national (habituellement par le gouvernement fédéral) et ceux qui le sont à un niveau infra-national (par les provinces ou les municipalités, par exemple). Pour que ces derniers soient utiles au niveau national, il faut que divers secteurs de compétence s'entendent notamment sur les définitions, les normes, la présentation des documents et la marche à suivre. Il n'est pas toujours facile de conclure de telles ententes, surtout dans des domaines qui sont constitutionnellement de la compétence des provinces.

Les dossiers administratifs peuvent être classés selon d'autres critères, c'est-à-dire la méthode de collecte (interviews personnelles, enquêtes postales, observations, etc.), le moyen d'enregistrement (sur papier ou fichier informatique), l'accessibilité (voir la section 6).

4. UTILISATION DES DOSSIERS ADMINISTRATIFS

La plus grande partie du symposium porte sur les façons dont les dossiers administratifs peuvent être utilisés à des fins statistiques. Ces utilisations peuvent être regroupées en cinq grandes catégories. La plupart des applications qui sont faites des dossiers administratifs à des fins statistiques se rattachent à l'une de ces catégories, en représentent une combinaison ou en constituent une variante.

(i) Totalisations directes

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

(ii) Remplacement des données d'enquêtes

Cette catégorie recouvre des cas d'utilisation des données administratives à la place d'enquêtes de certaines strates, par exemple, l'utilisation de données fiscales pour les petites entreprises à Statistique Canada. On pourrait élargir cette catégorie pour inclure les cas où les données administratives sont utilisées selon un processus d'imputation afin de compenser le taux de non-réponse à l'enquête.

(iii) Estimations indirectes

Cette catégorie comprend les cas où les données tirées des dossiers administratifs constituent un des entrants utilisés dans un processus d'estimation. Un exemple serait le couplage de deux déclarations d'impôt consécutives d'un particulier dans le but de produire des estimations partielles de la migration, lesquelles peuvent être pondérées à l'aide des données de recensement utilisées comme des étalons. Ces estimations de la migration sont ensuite utilisées dans le programme d'estimations démographiques de Statistique Canada, au même titre que les données administratives sur les naissances, les décès et l'immigration. Un deuxième exemple serait l'utilisation des données administratives comme source de variables auxiliaires pour l'estimation par le quotient ou par régression.

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

(iv) Bases de sondage

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

(v) Évaluation des enquêtes

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

De nombreuses applications comprennent la combinaison de l'utilisation des dossiers administratifs et d'autres méthodes de collecte telles que les enquêtes-échantillons et les recensements. Par exemple, le projet de remaniement des enquêtes-entreprises de Statistique Canada repose principalement sur l'utilisation des dossiers administratifs et des enquêtes pour l'entretien de la base et la collecte de données économiques (Colledge,

1987). Il est clair que le programme d'estimations démographiques comporte une combinaison de recensements et de dossiers administratifs, avec des enquêtes-échantillons également utilisées pour évaluer le taux de couverture du recensement. Des exemples d'approches intégrées de Statistique Canada et d'ailleurs seront donnés lors d'autres séances du symposium.

L'utilisation qui sera faite des dossiers administratifs est surtout fonction de leur contenu et de leur qualité. Nous nous pencherons ensuite sur les questions de la qualité des dossiers administratifs.

5. QUALITÉ DES DOSSIERS ADMINISTRATIFS

Pour déterminer si certains dossiers administratifs peuvent convenir à une utilisation particulière, il faut tenir compte de divers facteurs, dont voici les plus importants:

(i) le taux de couverture prévu du système administratif;

(ii) le contenu du système administratif ou les variables qu'il renferme et les concepts et les définitions qui le sous-tendent;

(iii) la qualité des données déclarées et de leur traitement dans le système administratif;

(iv) la rapidité avec laquelle les données peuvent être obtenues pour l'utilisation à des fins statistiques.

Les deux premiers facteurs se rattachent au but du programme administratif. La population visée, les renseignements recueillis et les définitions utilisées sont généralement dictés par le but administratif. Il est important de savoir dans quelle mesure les besoins statistiques pourraient influencer le taux de couverture ou le contenu des dossiers administratifs. Nous y reviendrons à la section 6.

Le troisième facteur démontre jusqu'à quel point le système administratif obtient le taux de couverture et le contenu souhaités et contribue grandement à évaluer l'utilité d'une source administrative à des fins statistiques. On ne peut généraliser la qualité des dossiers administratifs; on doit examiner la qualité des ensembles de dossiers administratifs individuellement.

Le quatrième facteur, l'actualité des données, est un point important à considérer lorsque l'on intègre l'utilisation de dossiers administratifs aux processus statistiques. On doit tenir compte du temps qui s'écoule entre la période de référence et le moment où le fichier est utilisé à des fins statistiques ainsi que de l'assurance de recevoir le fichier à temps.

Les points qui suivent résument certaines caractéristiques des systèmes de dossiers administratifs qui peuvent avoir une incidence sur leur utilité statistique:

1. Qu'est-ce qui motive les unités individuelles à s'inscrire dans un système administratif? Les programmes qui procurent des bénéfices aux personnes inscrites (par exemple les allocations familiales, l'assurance-maladie) devraient permettre un taux de couverture très complet des personnes admissibles. Il peut même y avoir des problèmes de surdénombrement lorsque le mécanisme de retrait de ceux qui ne sont plus admissibles est déficient. Il se peut que les programmes qui ne sont pas perçus comme étant avantageux pour les personnes inscrites (par exemple l'impôt, l'octroi de certaines licences) aient un taux de couverture incomplet.

2. Les dossiers administratifs ne représentent habituellement pas une bonne source de données recoupées étant donné qu'en général, ils ne recueillent que l'ensemble

limité des variables nécessaires à l'administration d'un programme. Par exemple, le fichier de déclarations d'impôt contient des données détaillées sur le revenu et renferme cependant des données limitées à d'autres égards, par exemple l'âge et le sexe, mais aucune donnée sur le niveau de scolarité ou de l'industrie.

3. Étant donné que les concepts et les définitions utilisés dans les systèmes administratifs sont conçus pour répondre aux besoins du programme, ils ne coïncideront pas nécessairement avec ceux de l'analyse sociale ou économique. L'utilisation d'un fichier de déclarations d'impôt, comme source de données sur la population d'âge actif doit être tempérée par le fait que ce ne sont pas toutes les personnes dans ce groupe d'âge qui sont des déclarants. De plus, certains groupes ayant un faible taux de couverture dans le fichier fiscal sont précisément ceux qui revêtent une importance analytique (par exemple les personnes à faible revenu, les personnes âgées). Un deuxième exemple bien connu serait les personnes qui reçoivent des prestations d'assurance-chômage et qui ne seraient pas considérées comme des chômeurs selon les normes statistiques internationales.

4. Le taux de couverture et le contenu des dossiers administratifs peuvent contenir des solutions de continuité en raison de changements de lois, de règlements ou de procédures administratives. Des changements semblables perturberont l'estimation qui consiste à étalonner les données administratives aux autres sources de données. Par exemple, l'introduction du crédit d'impôt pour enfants à la fin des années 70 a entraîné une augmentation soudaine du taux de couverture des fichiers de déclarations d'impôt.

5. Les dossiers administratifs peuvent être une source précieuse de données régionales parce qu'ils sont en quelque sorte un recensement. Cependant, pour être utile, chaque dossier doit avoir un code d'emplacement géographique précis. Les codes postaux peuvent servir à cette fin lorsque l'adresse est indiquée. Il faut s'assurer que l'adresse inscrite corresponde à l'endroit approprié. Par exemple, les déclarations d'impôt peuvent contenir l'adresse de l'escompteur d'impôt au lieu de celle du déclarant.

6. Les procédures d'assurance de la qualité des données du système administratif peuvent être très rigides pour des variables essentielles à l'administration du programme, mais beaucoup moins pour d'autres variables. Comme on l'a mentionné plus haut, les identificateurs géographiques peuvent être nébuleux.

7. Souvent, les dossiers administratifs sur les particuliers n'identifieront pas les familles ou les ménages. Dans certains cas, il est possible de combiner les personnes aux familles ou aux ménages, mais il arrive souvent que les renseignements nécessaires pour l'appariement ne soient pas inscrits.

Les caractéristiques ci-dessus peuvent avoir fait ressortir les points faibles des dossiers administratifs plutôt que leurs points forts. Dans la production de statistiques, on doit tenir compte de ces forces et faiblesses ainsi que de celles des recensements et des enquêtes-échantillons. Dans bien des cas, le programme statistique optimal a recours à une combinaison de ces différentes sources de données de façon intégrée. Le Manuel d'organisation statistique des Nations Unies résume bien ce point [4]:

> "Un programme équilibré d'amélioration des statistiques nationales implique l'exploitation des recensements, des enquêtes par sondage et des relevés administratifs. À long terme, ces trois sources de statistiques sont en grande partie complémentaires; chacune présente des avantages et souffrent de certaines limitations. Aussi la pleine exploitation de l'une d'entre elles ne rend-elle nullement superflues les deux autres."

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

D'autres questions importantes portent sur l'accès du statisticien aux systèmes de dossiers administratifs et la façon dont il peut influer leur contenu. Il ressortira clairement à la fin du symposium, si ce n'est déjà fait, combien de nombreux organismes statistiques, dont Statistique Canada, sont tributaires des documents administratifs. Il faut donc des mesures afin d'assurer l'accès permanent et accru à ces documents ainsi que des mesures pour les rendre plus utiles du point de vue statistique.

Accès

Il faut obtenir l'autorisation pour avoir accès aux dossiers administratifs. Le fondement juridique pour Statistique Canada est l'article 12 de la Loi sur la statistique (1971) qui se lit comme suit:

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

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

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

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

Influence sur les changements

Nous avons déjà mentionné l'incidence que des modifications apportées aux pratiques ou règlements administratifs auraient sur les statistiques produites. Il suffirait qu'on change le taux de couverture d'un programme, qu'on introduise une mesure incitant les personnes visées par un programme à y participer ou, au contraire, à ne plus y participer

ou bien qu'on modifie la procédure de manière à altérer la qualité ou l'exhaustivité des dossiers, et les séries chronologiques seraient interrompues. L'organisme statistique doit se méfier des changements d'origine externe et agir en conséquence.

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

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

Il s'agit là d'un domaine où un système statistique décentralisé a certains avantages sur un système centralisé comme celui au Canada. On pourrait croire qu'il est plus facile pour les statisticiens qui travaillent dans un organisme ou un ministère d'influencer les systèmes ministériels que pour ceux situés dans un organisme statistique centralisé. Peut-être que certains visiteurs des États-Unis ou d'ailleurs pourront nous fournir des précisions à ce sujet.

Dans un système statistique centralisé, il existe certains mécanismes qui peuvent ajouter un poids aux études statistiques lors de la conception et du remaniement des systèmes administratifs. Certains de ces mécanismes sont bilatéraux, alors que d'autres concernent l'ensemble du secteur public.

1. Créer des comités bilatéraux composés de hauts fonctionnaires qui examineraient les questions intéressant les deux parties, notamment les problèmes relatifs à la fourniture de données administratives;
2. Fournir à l'organisme administratif les données statistiques afin de montrer à la fois l'utilité des données et, le cas échéant, les lacunes attribuables aux pratiques administratives;
3. Fournir des conseils ou des services techniques au service statistique de l'organisme administratif;
4. Adopter une politique officielle en matière de collecte de renseignements exigeant, par exemple, que tout projet de collecte de données (dans un but statistique ou administratif) soit soumis à l'examen d'un organisme central;
5. Demander que chaque nouvelle proposition de programme soit assortie d'un plan prévoyant la façon d'obtenir les renseignements statistiques nécessaires pour contrôler et évaluer le programme en question;
6. Promouvoir l'utilisation des définitions statistiques normalisées (par exemple famille, établissement commercial, chômeur) dans les systèmes administratifs;
7. Faire en sorte que les vérificateurs des activités gouvernementales recommandent l'utilisation des dossiers administratifs comme moyen économique par excellence de collecte de renseignements;

8. Suivre une orientation politique favorisant un recours accru à certains systèmes administratifs ou la recherche de solutions de rechange aux enquêtes;

9. Supprimer les obstacles législatifs qui limitent l'accès aux dossiers administratifs ou à leur utilisation à des fins statistiques.

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

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

7. PERCEPTION DU PUBLIC

La dernière question, et peut-être la plus importante, concerne la perception que le public a de l'utilisation des dossiers administratifs. Nous pensons aux images négatives qu'évoquent les notions de vastes banques de données créées par le couplage de dossiers et qui permettent de retracer les caractéristiques et les activités des personnes. Cette catégorie comprend aussi les débats en vue de déterminer si les données devraient servir à des fins autres que celles pour lesquelles elles ont été recueillies. La protection des renseignements personnels et la nécessité de garantir la gestion et le contrôle de l'utilisation de ces renseignements sont au coeur de ces préoccupations.

Par contre, on doit faire ressortir les perceptions favorables du public concernant l'utilisation des dossiers administratifs à des fins statistiques. L'allègement du fardeau de réponse et la réduction des dépenses gouvernementales sont des effets positifs de l'utilisation de dossiers administratifs.

Pour que l'utilisation des données administratives à des fins statistiques soit acceptée, il est probablement nécessaire de remplir trois conditions principales. Tout d'abord, il faut reconnaître que les travaux statistiques constituent une utilisation secondaire légitime des dossiers administratifs. La Loi sur la protection des renseignements personnels reconnaît ce principe sous réserve. Cependant, il est douteux que le citoyen moyen fasse la différence entre l'utilisation statistique, où l'identité du titulaire de chaque dossier ne présente aucun intérêt permanent, et l'utilisation administrative où il est essentiel de savoir à qui chaque dossier se rapporte. Il serait plus facile d'expliquer cette différence et de s'en servir si l'on pouvait dire sans équivoque que, pour le genre de travaux qui sont faits en statistique, on n'a jamais besoin d'identificateurs. Malheureusement, ce n'est pas le cas. Plusieurs techniques statistiques tout aussi légitimes exigent qu'on se serve d'identificateurs au cours des manipulations intermédiaires des données. Ces techniques supposent toute une forme quelconque d'appariement des données provenant de différents fichiers ou de différentes sources, et les identificateurs permettent d'apparier correctement les dossiers ensemble. Une fois cette opération effectuée, les dossiers peuvent être rendus anonymes si aucun couplage subséquent n'est prévu. Citons, à titre d'exemples de cas où une forme d'identification est nécessaire, le recensement de la population pour veiller à ce que le taux de couverture soit exhaustive et s'en assurer par la suite, les études longitudinales pour lesquelles on utilise des dossiers administratifs, les enquêtes épidémiologiques et enfin les études d'évaluation visant à vérifier les réponses dans des questionnaires d'enquête en les comparant à des données de sources administratives. Le fait de devoir expliquer pourquoi on a besoin d'identificateurs alors que l'identité des personnes auxquelles les données se rapportent ne présente aucun intérêt est un défi de taille pour l'organisme statistique.

En deuxième lieu, il devrait y avoir un contrôle manifeste du couplage des dossiers. Statistique Canada a élaboré une politique sur le couplage des dossiers. Cette politique reconnaît les avantages du couplage des dossiers pour les programmes statistiques ainsi que les problèmes qu'engendre le couplage des dossiers sans contraintes et précise que le couplage des dossiers nécessite un bon jugement pour équilibrer ces deux facteurs.

Essentiellement, la politique stipule que Statistique Canada s'occupera du couplage des dossiers si toutes les conditions suivantes sont remplies:

- que le couplage des documents soit effectué à des fins statistiques ou de recherche qui cadrent avec le mandat de Statistique Canada;
- que les produits du couplage soient diffusés uniquement s'ils sont en conformité avec les dispositions relatives à la confidentialité de la Loi sur la statistique;
- que le couplage prouve qu'il y a réduction des coûts et allègement du fardeau de réponse ou qu'il est la seule option possible;
- que le couplage ne serve pas à des fins qui iraient à l'encontre des intérêts des particuliers en cause et que les avantages du couplage soient nettement dans l'intérêt du public;
- que le couplage ne mette pas en cause la réalisation de programmes de Statistique Canada dans l'avenir;
- que le couplage satisfasse un processus de révision et d'approbation prescrits.

Le processus de révision et d'approbation comprend nécessairement la documentation de chaque proposition de couplage de dossiers, la révision et la recommandation par le

Comité de la confidentialité et des mesures législatives de Statistique Canada et l'approbation du statisticien en chef et du ministre.

Grâce à ce processus, le couplage de dossiers est réalisé là où le besoin le justifie et dans l'intérêt du public et n'est pas effectué à la légère.

Troisièmement, la confidentialité des données des dossiers administratifs est essentielle. Il s'agit à la fois de la protection des dossiers individuels et de la vérification des agrégats afin de garantir qu'aucun tableau divulgue par inadvertance les données qui se rattachent à un particulier. La confidentialité est évidemment une nécessité pour toutes les microdonnées dans un organisme statistique et non seulement celles tirées des dossiers administratifs. Il faudrait souligner le fait que les microdonnées ne circulent que dans un sens, c'est-à-dire de l'organisme administratif à l'organisme statistique.

En plus de ces trois critères principaux, d'autres mesures peuvent être prises afin de prévenir ou de réduire les réactions défavorables du public à l'utilisation de dossiers administratifs. L'information publique peut mettre l'accent sur la confidentialité et sur les avantages qui découlent de cette utilisation, à savoir l'allègement du fardeau de réponse et des coûts. Dans toute collecte de renseignements personnels, il faudrait préciser l'utilisation statistique dont ces renseignements pourraient faire l'objet, comme l'exige la Loi sur les renseignements personnels.

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

8. CONCLUSION

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

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

rigoureux, s'est avéré d'une très grande utilité, notamment en ce qui a trait aux données sur les entreprises, aux études longitudinales du marché du travail et aux enquêtes épidémiologiques.

En raison de l'utilisation croissante des dossiers administratifs, les organismes statistiques sont de plus en plus tributaires des autres organismes pour la fourniture continue des données dont ils ont besoin pour leurs programmes. Quelles que soient les lois et les politiques qui régissent les activités de l'organisme statistique, l'établissement d'ententes de collaboration étroite avec les organismes fournisseurs est extrêmement important. La capacité de l'organisme statistique d'influer sur la conception ou la refonte de systèmes administratifs repose sur une compréhension mutuelle des besoins des deux organismes intéressés. Si l'on pouvait adopter, à l'échelle de l'appareil gouvernemental, une politique ou un principe donnant à Statistique Canada voix au chapitre de la conception des systèmes administratifs ou, plus généralement, de la façon de satisfaire les besoins statistiques des nouveaux programmes, cela faciliterait la tâche de l'organisme statistique mais ne remplacerait pas la création d'un climat de collaboration étroite avec les organismes administratifs.

Les séances du symposium donneront un bon aperçu des diverses questions concernant l'utilisation des dossiers administratifs. En plus des questions de politique globales sur lesquelles nous nous sommes penchés, nous traiterons lors d'autres séances du couplage des dossiers, de la qualité des données et de l'utilisation des dossiers administratifs conjointement avec les données de recensement ou d'enquêtes ou en remplacement partiel de ces dernières. Le panel qui aura lieu mercredi se concentrera sur les renseignements personnels et la perception du public.

Les systèmes administratifs contiennent tellement de données que le véritable défi des statisticiens est d'influencer et de faire en sorte que ces données fassent partie intégrante de ces systèmes. Il en est ainsi dans certains domaines dans certains pays, mais il reste encore beaucoup à faire et de nombreux avantages à tirer à ce niveau.

BIBLIOGRAPHIE

American Statistical Association (1977). Report of the Ad Hoc Committee on Privacy and Confidentiality. *The American Statistician*. 31, 59-78.

Bjerve, P. J. (1985). "International Trends in Official Statistics" in A Celebration of Statistics, *The ISI Centenary Volume*. Ed. Atkinson, A.C. and Fienberg, S.E.

Colledge, M.J. (1987). "The Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada" Proceedings of the Third Annual Research Conference, March 1987, U.S. Bureau of the Census, Washington, D.C.

Hansen, M. "Some History and Reminiscences on Survey Sampling". *Statistical Science*, May 1987, Vol. 2 No. 2, 180-190.

Hansen, M. H., Dalenius, Tore, and Tepping, Benjamin J. (1985). "The Development of Sample Surveys of Finite Populations" in A Celebration of Statistics, *The ISI Centenary Volume*. Ed. Atkinson, A.C. and Fienberg, S.E.

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

Nations Unies (1980), "Manuel d'organisation statistique" volume 1, Étude de l'organisation des services nationaux de statistiques et des problèmes connexes de gestion.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

L'EXPÉRIENCE EUROPÉENNE RELATIVE À L'UTILISATION DES DONNÉES ADMINISTRATIVES POUR RECENSER LA POPULATION: QUESTIONS D'ORDRE POLITIQUE

PHILIP REDFERN[1]

RÉSUMÉ

L'expérience de quatre pays scandinaves fait ressortir les avantages et les inconvénients des recensements de la population réalisés à partir de registres et montre comment on pourrait remédier aux inconvénients. Dans d'autres pays, les tenants de cette façon de procéder se heurtent à des obstacles: soit on n'y possède pas les systèmes de données nécessaires ou de la qualité voulue, soit le public y voit une menace à la vie privée et s'interroge sur le pouvoir de l'État. Ces questions se situent bien au-delà du domaine de la statistique: elles sont d'ordre politique et administratif. Dans cette communication, la situation dans deux pays, le Royaume-Uni et l'Australie, est examinée. Au Royaume-Uni lorsque, par le passé, il a été question d'établir un registre de population en période de paix, ce genre de tentative a échoué, et l'opinion publique actuelle y est toujours hostile. Le gouvernement a néamoins entrepris une réforme controversée des taxes locales, qui suppose la création de nouveaux registres. En Australie, le gouvernement déposa un projet de loi visant à introduire une carte d'identité nationale fondée sur un registre central et invoqua des arguments politiques clairs à l'appui de ce projet de loi; plutard cette loi fût rétractée. La conclusion est que les questions soulevées par la réforme des systèmes de données méritent un examen approfondi, et quelques raisons pour lesquelles les statisticiens devraient prendre une part prépondérante au débat sont avancées.

1. INTRODUCTION

Cette communication s'inspire d'une étude sur les différentes façons de procéder à un recensement de la population que j'ai réalisée pour le compte de l'Office statistique des communautés européennes (Redfern 1987). Pour cette étude, je me suis penché sur l'expérience des douze pays membres de la CEE et sur celle du Canada, de la Suède et des États-Unis. Il en est ressorti que les enquêtes par sondage peuvent compléter mais non remplacer les recensements pour la bonne raison quelles ne permettent pas d'obtenir des statistiques régionales fiables. L'utilisation d'un questionnaire long et d'un questionnaire abrégé au recensement du Canada et à celui des États-Unis est un exemple important de sondage complétant un dénombrement intégral. Il est question que la Norvège ait quant à

[1] Philip Redfern, Royaume-Uni, 17 Fulwith Close, Harrogate, North Yorshire, Angleterre. HG2 8HP.

elle recours à une enquête par sondage pour compléter les données tirées de registres se rapportant à la population entière. (Section 3.3).

Il est possible de produire des données régionales à partir de registres contenant les adresses des personnes qui y sont inscrites. Si les domaines d'intérêt du recensement sont bien couverts par les registres (du point de vue des définitions, du champ d'observation, de l'exactitude et des périodes de référence) et que les registres peuvent être reliés entre eux, il est alors possible de créer, pour chaque individu, un enregistrement semblable à une déclaration de recensement et ainsi réaliser un recensement à partir des registres. Il s'agit essentiellement de recycler les données administratives à des fins statistiques. Les pressions que constituent les coûts et le fardeau de remplir les questionnaires habituels de recensement ont amené quatre pays scandinaves (le Danemark, la Finlande, la Norvège et la Suède) à adopter cette approche totalement ou en partie.

Les données administratives peuvent étayer un recensement **classique** de diverses façons (Redfern 1987), mais c'est leur utilisation dans un recensement **réalisé à partir de registres** qui sera le thème principal de cette communication. La section 2 décrit les registres dont on a besoin pour effectuer un tel recensement et la section 3 met en évidence les points communs et les différences entre les façons dont les quatre pays scandinaves ont choisi de procéder à cet égard. La section 4 examine ensuite les obstacles que les autres pays rencontreraient s'ils voulaient améliorer leurs systèmes de registres afin de pouvoir s'en servir pour réaliser des recensements, et il ressort de cette analyse que les problèmes soulevés sont davantage d'ordre administratif et politique que statistique.

Ce sont ces questions d'intérêt plus général qui constituent le second grand thème de la communication. La section 5 examine en détail la situation dans un pays où, pour des raisons politiques et idéologiques, les dossiers administratifs ne sont pas reliés entre eux au moyen d'un registre de population: le Royaume-Uni. La section 6 décrit une initiative récente en Australie, en vue d'améliorer les dossiers administratifs. Enfin, la section 7 résume les arguments politiques invoqués par les partisans et les adversaires du couplage des données administratives au moyen de registres de population et présente quelques raisons pour lesquelles les statisticiens devraient prendre une part prépondérante au débat sur la question.

2. REGISTRES NÉCESSAIRES À LA RÉALISATION D'UN RECENSEMENT

2.1 Registres de population

Pour réaliser un recensement à partir de registres, il faut, comme point de départ essentiel, un registre de population qui contienne des numéros de référence personnels et des adresses. Une correspondance biunivoque doit exister entre les numéros personnels et les membres de la population. Pour que le registre puisse être tenu à jour, les citoyens doivent informer les responsables de tout changement. Les numéros personnels doivent également figurer dans les fichiers des divers organismes administratifs qui tiennent des dossiers afin de pouvoir servir à l'appariement de tous les dossiers à des fins statistiques.

La tenue d'un registre de population sert essentiellement à des fins administratives. Il s'agit d'une façon efficace d'organiser les nombreux rapports entre les pouvoirs publics, au niveau central et local, et le simple citoyen, pour ce qui a trait notamment aux impôts, à la sécurité sociale, aux services de santé offerts par l'état et à l'inscription sur les listes électorales. Pour être d'une utilité optimale, ces registres doivent servir à une vaste gamme d'activités administratives, de manière que les occasions de les mettre à jour et de les corriger soient fréquentes et que les citoyens s'habituent à donner leur numéro personnel.

La clé du système est le registre central de population dans lequel sont consignés des renseignements permettant d'identifier chaque personne (nom, date et lieu de naissance, date d'immigration, état matrimonial et éventuellement origine et citoyenneté) ainsi qu'un numéro de référence permanent. Dans la plupart des pays, le registre central de population contient les adresses les plus récentes, ce qui n'est toutefois pas le cas du registre français, **le Répertoire national d'identification des personnes physiques.** La fonction administrative première du registre central est de servir de point de référence pour les organismes administratifs qui peuvent y vérifier l'identité des personnes avec lesquelles ils ont des rapports et, s'il y a lieu, corriger ou inscrire les numéros de référence personnels dans leurs fichiers.

2.2 Autres registres importants

Pour réaliser un recensement de la population et du logement à partir de registres, on aussi recours à des registres fondés sur des unités autres que les personnes. Les plus importants sont les registres centraux des logements et les registres centraux des entreprises et des établissements (lieux de travail). Dans la mesure où le registre des logements attribue à chaque unité de logement (et pas seulement à l'immeuble ou à l'adresse) un code qui fait aussi partie de l'adresse inscrite dans le registre de population, les données sur les unités de logement contenues dans ce registre peuvent être reliées aux données qui se trouvent dans le registre de population et qui se rapportent aux occupants de ces unités. Autrement dit, les deux registres peuvent être appariés. On s'y prend de la même façon pour apparier le registre où sont inscrits le nom de l'employeur et le lieu de travail de chaque personne et le registre central des entreprises et des établissements et ainsi obtenir des données concernant la branche d'activité dans laquelle une personne travaille, ses déplacements pour aller travailler (navettage), etc.

3. LE RECENSEMENT DANS LES PAYS SCANDINAVES

Les quatre pays scandinaves possèdent des registres de population bien conçus du type décrit dans la section 2.1. Ils ont constitué, ou ont l'intention de le faire, des registres centraux des immeubles et des logements servant principalement à des fins administratives. Dans cette section, je vais décrire brièvement le genre de recensement effectué dans chaque pays et résumer l'orientation que cette activité est en train de prendre en Scandinavie.

3.1 Danemark

Le Danemark est le seul pays scandinave et, à ma connaissance, le seul pays d'Europe, qui a complètement renoncé à effectuer des recensements classiques pour n'en réaliser qu'à partir de registres. Ce changement s'est fait sur une période de plus de dix ans. Le registre central de population contenant des numéros de référence personnels a été créé en 1968 à des fins administratives et, en 1976, un recensement de la population (mais pas du logement) a été effectué à partir de ce registre. En 1977, on a créé un registre central des immeubles et des logements, toujours principalement dans un but administratif, et, en 1981, on a réalisé cette fois un recensement de la population **et** du logement, toujours à partir de registres. En 1979-1980, une mesure importante de plus a été prise, celle de demander aux employeurs d'ajouter un renseignement aux déclarations qu'ils envoient au fisc et dans lesquelles ils inscrivent les gains de chaque employé: ceux qui avaient plus d'un établissement devaient indiquer le lieu de travail de chaque employé. Ce renseignement supplémentaire a été exigé à des fins purement statistiques, et le bureau de la statistique a dû déployer des efforts considérables pour obtenir la collaboration des déclarants.

Toutefois, pour des questions de coût, il n'est pas possible de faire le genre d'analyse qu'on fait d'après les données de recensement à une fréquence le moindrement comparable: pour avoir des analyses semblables à celles qui ont suivi le recensement de 1981, il faudra attendre 1991, et encore là risquent-elles d'être plus restreintes.

Le passage à un recensement réalisé à partir de registres a été facilité par la réorganisation du bureau central de la statistique du Danemark en 1966. Le Danmarks Statistik s'est vu accorder une certaine indépendance vis-à-vis du gouvernement central, ce qui a peut-être rassuré le public relativement à la question de la confidentialité. Cet organisme a maintenant le droit d'exiger, et d'utiliser à des fins statistiques, des données recueillies par les pouvoirs publics à des fins administratives et celui de participer à la création de registres contenant ce genre de données.

Les problèmes auxquels Danmarks Statistik doit à présent faire face concernent principalement la qualité et l'actualité des données, qui sont toutes deux tributaires de l'efficacité de la procédure administrative. Ainsi, la lenteur avec laquelle les données des fichiers fiscaux ont été compilées (données sur la branche d'activité, la profession, le navettage et le revenu) a retardé l'analyse de ces données pour le recensement de 1981 jusqu'à l'été 1983. On s'attend par ailleurs que les statistiques sur la population active continueront d'avoir au moins un an de retard par rapport à l'année de référence à laquelle elles se rapportent. Il est particulièrement difficile d'obtenir des données fiables sur la profession, parce que ce sujet revêt peu d'intérêt du point de vue administratif; la principale source d'information est la déclaration d'impôt annuelle. Malgré ces problèmes, Danmarks Statistik considère que les recensements réalisés à partir de registres constituent une réalité bien ancrée au Danemark parce qu'ils permettent de réduire les coûts et d'alléger le fardeau de réponse imposé au public (Jensen 1983).

3.2 Finlande

Les recensements réalisés à partir de registres ont une longue histoire en Finlande. Au 17^{e} siècle, tous les membres d'une paroisse âgés de plus de 12 ans étaient inscrits dans les registres paroissiaux, et en 1749 on a compilé les données se rapportant à la population entière et on les a analysées selon l'âge, le sexe, l'état matrimonial et la classe sociale. Était-ce là un des premiers recensements jamais réalisé à partir de registres? Les recensements suivants ont été effectués de la même manière, mais en 1950 et en 1960 on a adopté la méthode classique qui consiste à recueillir des renseignements au moyen d'un questionnaire. Cependant, depuis le recensement de 1970, on tire des registres un éventail de plus en plus large de données. Les questions posées au recensement quinquennal de 1985 appartenaient uniquement au domaine économique: genre d'activité (le cas échéant) et situation professionnelle, employeur et lieu de travail, profession et nombre de mois travaillés pendant l'année précédente. Les données sur le logement ont été tirées du registre des immeubles et des logements créé à partir des données du recensement de 1980 et mis à jour à l'aide des renseignements fournis par les municipalités.

Le recensement de 1985 a été conçu de manière à coûter un peu moins que l'équivalent d'un dollar américain par personne, soit le **quart** du coût du recensement de 1980 en termes réels, tout en comportant la même gamme de variables. Parmi les facteurs qui ont permis de réaliser cet exploit, mentionnons: les questionnaires envoyés par la poste sur lesquels étaient préimprimées les données sur le lieu de travail (d'après les renseignements recueillis au recensement de 1980) et sur la profession (d'après les renseignements inscrits dans le registre central de population) que les recensés pouvaient corriger au besoin; le retour des questionnaires par la poste directement au bureau central, sans passer par une organisation locale; un seul rappel, et aucun suivi dans le cas des formules qui n'avaient pas été renvoyées ou qui étaient incomplètes (3.7%); et l'imputation des données manquantes en ayant recours, dans le mesure du possible, à

divers registres, dont les dossiers des régimes de retraite pour ce qui est des employés du secteur privé. Le taux de réponse au questionnaire était de 97.4% et, après l'imputation des données manquantes, le taux de couverture final était de 98.6%. Le faible coût du recensement est également attribuable au fait que le fardeau financier et administratif a été assumé en partie par les services chargés d'administrer les registres, dont la vérification annuelle sur le terrain des registres de population par l'envoi de formules à chaque ménage ou logement et la vérification quinquennale du registre des immeubles et des logements par l'envoi de formules aux propriétaires et aux occupants.

La comparaison des données tirées du recensement de 1980 et des données se rapportant aux variables économiques tirées des registres a donné des résultats considérés comme encourageants. C'est pour cette raison et aussi grâce aux méthodes d'imputation des caractéristiques économiques des non-répondants élaborées à l'occasion du recensement de 1985 qu'il y a de bonnes chances que le recensement de la Finlande soit entièrement réalisé à partir de registres en 1990. Pour combler la seule lacune des données tirées de registres, les employeurs possédant plus d'un établissement devront à l'avenir déclarer le lieu de travail de chaque employé (Laihonen et Myrskylä 1987; Heinonen et Laihonen 1987).

3.3 Norvège

Le recensement de la Norvège de 1980 a été réalisé dans une grande mesure à partir de registres. Ceux-ci ont fourni les données démographiques de base, celles sur le revenu et celles sur les études terminées (à l'exception des études faites à l'étranger). Pour compléter ces données, on a envoyé a chaque personnes âgées de 16 ans et plus un questionnaire qu'elles devaient retourner par la poste et qui comportait des questions de nature économique et des questions sur les études faites à l'étranger, sur le pays de naissance, sur la religion et sur le logement. Toutes les personnes faisant partie d'un même ménage devaient renvoyer leur formule, ainsi qu'une formule sur le logement, dans la même enveloppe, établissant ainsi la composition du ménage aux fins du recensement.

Il y a plusieurs raisons pour lesquelles il ne sera pas possible, en 1990, de faire un recensement réalisé entièrement à partir de registres. Premièrement, les données des registres relatives à certaines variables importantes du recensement ne sont pas compatibles avec les définitions statistiques retenues ou ne sont pas d'assez bonne qualité pour le recensement (c'est le cas notamment des données relatives à la branche d'activité) ou encore les registres ne contiennent pas de données se rapportant à certaines variables (par exemple la profession). Deuxièmement, il est peu probable que le registre des propriétés foncières, des adresses et des immeubles (le "GAB") qu'on a commencé à constituer en 1983 soit assez avancé en 1990 pour qu'on puisse en tirer des données sur le logement. Troisièmement, comme c'est l'adresse qui constitue le lien entre le GAB et le registre de population, il n'est pas possible de déterminer la composition du ménage ni d'associer caractéristiques du logement et caractéristiques personnelles lorsque plusieurs unités de logement portent la même adresse.

Pour le recensement de 1990, on ira encore chercher dans les registres les données démographiques de base, celles sur le revenu et celles sur les études terminées (autres que des personnes âges de 16 ans et plus (100% dans les municipalités de moins de 6,000 habitants), afin qu'elles soient compatibles avec les définitions statistiques retenues. Pour corriger les données des registres se rapportant à une sous-population, on se servira, d'une part, des données recueillies auprès d'un échantillon de cette sous-population et, d'autre part, des données recueillies auprès d'un échantillon d'une population dont l'effectif est plus grand, ce qui aura pour effet d'éliminer en partie le biais dont sont entachées les données des registres. Le sondage sera la seule source de données pour les sujets sur lesquels il n'existe rien dans les registres, c'est-à-dire la profession et probablement aussi le logement et la composition des ménages.

On estime le coût de cette approche, soit l'utilisation des registres et la tenue d'un sondage auprès d'un échantillon de 10% de la population visée, à 60% du coût d'un recensement comme celui de 1980. Le prix à payer sera, premièrement, la variabilité d'échantillonnage, qui sera à son plus fort pour les sujets sur lesquels il n'existe aucune donnée dans les registres, et, deuxièmement, un certain biais dans le cas des données contenues dans les registres mais dont la qualité n'est pas celle qu'on recherche pour un recensement (Johansen 1987).

3.4 Suède

En ce qui concerne le recensement de la Suède, la situation s'est inversée au cours des vingt dernières années: en 1970, la plupart des données provenaient des questionnaires et un petit nombre de registres; en 1985, c'était le contraire. Cette année-là, dans le questionnaire envoyé et retourné par la poste, on demandait à chaque personne âgée de 16 ans et plus (ou couple marié) d'indiquer uniquement: (1) si elle était active au cours d'une semaine précise et, le cas échéant, quel métier ou profession elle exerçait; (2) la composition du ménage, soit la liste des adultes habitant dans le même logement; (3) des renseignements sur le logement. Il a été possible d'omettre les questions sur le nom de l'entreprise pour laquelle la personne travaillait, le lieu de travail et la branche d'activité, questions qui avaient été posées au recensement précédent, parce qu'on a demandé aux employeurs d'ajouter un renseignement à leur déclaration d'impôt annuelle, en l'occurrence le lieu de travail de chaque employé. Par contre, les employeurs se sont opposés à indiquer également dans leur déclaration d'impôt le nombre d'heures travaillées, et ce sujet a donc été abandonné au recensement de 1985.

Après le recensement de 1980, on avait fait une étude portant sur les mesures qu'il faudrait prendre pour que le recensement de 1985 soit entièrement réalisé à partir de registres. Parmi ces mesures, il y avait les suivantes:

1) Utiliser les données sur la profession tirées des formules dans lesquelles les personnes occupées déclarent tout changement de revenu au bureau de l'assurance nationale.

2) Créer un registre de la composition des ménages qu'on mettrait à jour en recueillant des renseignements à l'occasion des déménagements.

3) Créer un registre des immeubles qui contiendrait des données sur les unités de logement et qui serait mis à jour par les municipalités.

4) Créer un registre des études terminées qui serait mis à jour à l'aide des renseignements fournis par les établissements d'enseignement sur les diplômes décernés.

Comme on l'a vu plus haut, on a conservé le questionnaire au recensement de 1985 principalement parce qu'on avait des doutes au sujet de la qualité des renseignements qui pourraient être tirés des registres relativement à la profession, à la composition des ménages et au logement. De tous les nouveaux registres proposés, seul celui sur les études terminées est en voie d'être constitué. Cependant, un comité étudie actuellement actuellement la possibilité d'indiquer l'unité de logement avec l'adresse dans les registres de population, ce qui est essentiel si l'on veut procéder au couplage des registres de population et des registres des logements.

Une commission parlementaire est en train de se pencher sur le recensement de 1985, particulièrement sur les aspects relatifs à la vie privée et à la confidentialité. Les conclusions auxquelles elle arrivera contribueront à déterminer la forme que prendra le recensement de 1990. Le rapport final de la commission est prévu pour 1988.

3.5 Les recensements en Scandinavie: résumé

L'évolution du recensement dans les quatre pays scandinaves se fait selon des voies différentes, mais il y a de nombreux points communs:

1) Tous prennent comme point de départ des registres de population exacts permettant de produire régulièrement des statistiques régionales fiables.

2) Tous souhaitent maximiser l'utilisation des renseignements contenus dans les autres registres et minimiser le fardeau de réponse imposé au public. Tous s'efforcent de limiter ou de réduire les coûts.

3) Tous reconnaissent que les renseignements contenus dans les registres, particulièrement les renseignements de nature économique, posent des problèmes en ce qui a trait aux définitions, à la qualité et aux périodes de référence. On est en train d'augmenter le nombre de renseignements que les employeurs doivent déclarer, notamment en ce qui concerne le lieu de travail de chaque employé et donc la branche d'activité. Cependant, le fait qu'on demande des renseignements supplémentaires à des fins purement statistiques est mal accueilli, ce qui risque d'avoir pour conséquence une baisse de la qualité des données. Les données sur la profession qu'on trouve dans les registres ne sont en général pas fiables, et il y a des sujets sur lesquels on ne trouve rien du tout, notamment les moyens de transport utilisés pour se rendre au travail.

4) Des registres des immeubles et des logements ont été créés, ou du moins sont à l'état de projet. Il est difficile de tenir certains registres à jour, que ce soit en ayant recours aux renseignements dont disposent les municipalités ou en recueillant les renseignements nécessaires directement auprès des propriétaires. Dans certains pays, il faudrait améliorer les registres en identifient chaque unité de logement de manière à pouvoir faire le lien avec les adresses inscrites dans les registres de population. Il y a un autre problème: comment obtenir des données sur la composition des ménages si, comme en Suède, le ménage n'est **pas** défini comme étant constitué de tous les occupants de l'unité familiale.

Les quatre pays semblent prêts à sacrifier en partie la qualité des résultats si cela permet de réduire les coûts et d'alléger le fardeau de réponse imposé à la population. Mais ils n'adoptent pas tous la même approche. C'est le Danemark qui, en abandonnant le questionnaire du recensement, est allé le plus loin. Comme la qualité des données provenant de certains registres était mise en doute, particulièrement en ce qui a trait aux questions de nature économique, la Finlande et la Suède ont conservé un questionnaire restreint pour le recensement de 1985 et elles ont combiné les réponses obtenues aux données démographiques et autres tirées des registres. Cependant, il est possible qu'en 1990 le recensement de la Finlande soit entièrement réalisé à partir de registres. En Norvège, où il n'y a pas de recensement quinquennal, il est prévu qu'au recensement de 1990 on conserve le questionnaire au moins pour les questions d'ordre économique, mais qu'on ne l'envoie qu'à un échantillon de 10% de la population, dans le but de réduire les coûts. Les données de nature économique qui se trouvent dans les registres pourront être corrigées d'après les résultats obtenus auprès des échantillons afin de devenir compatibles avec les définitions statistiques retenues. Johansson (1987) donne un compte rendu très utile de l'expérience suédoise relative à l'utilisation des registres comme source de données pour le recensement.

4. POSSIBILITÉ DE RÉALISER DES RECENSEMENTS À PARTIR DE REGISTRES DANS LES AUTRES PAYS

Les deux principaux facteurs qui ont poussé les pays scandinaves à réaliser leurs recensements à partir de registres, soit la nécessité de réduire les coûts et d'alléger le fardeau de réponse, ne sont pas pour autant négligés ailleurs. Ainsi, ils ont eu pour effet de freiner, et parfois de renverser, la tendance observée avant 1980 selon laquelle les questionnaires de recensement devenaient de plus en plus longs.

Un nouvel élément troublant, l'opposition du public, a perturbé le recensement dans deux pays. Aux Pays-Bas, le projet de réaliser un recensement en 1981 a été abandonné. En République fédérale d'Allemagne, le recensement prévu pour 1983 dû être reporté jusqu'en 1987 à cause des exigences plus strictes que la cour constitutionnelle a imposées en matière de confidentialité, et malgré cela tout le monde n'y a pas participé. Aucun pays n'est à l'abri de ce genre de contestation. Cependant, un recensement réalisé à partir de registres risque moins d'être saboté dans la mesure où il n'a pas besoin d'être complété par un questionnaire. C'est qu'un tel recensement ne crée pas une occasion (le jour du recensement) où tout le monde a un questionnaire à remplir et où les protestations d'une minorité peuvent s'étendre jusqu'à prendre la forme d'une opposition massive.

Si un recensement réalisé à partir de registres coûte tellement moins cher et réduit à la fois le fardeau de réponse et les risque de sabotage, pourquoi si peu de pays considèrent-ils qu'il s'agit d'un choix praticable? Pour trois raisons. Premièrement, dans certains domaines, particulièrement le domaine économique, les données administratives sont parfois de qualité inférieure aux données recueillies au moyen d'un questionnaire, tandis que dans d'autres domaines, il n'existe pas de données administratives. Les pays scandinaves sont conscients de ces problèmes, et c'est pourquoi certains utilisent encore un questionnaire et combinent les réponses ainsi recueillies avec les données tirées des registres (Section 3.5).

Deuxièmement, un grand nombre de pays ne possèdent pas un système de données au gendre de celui décrit à la section 2. Par exemple, certains pays comme la République fédérale d'Allemagne, la Grèce et l'Italie ont des registres de population locaux mais pas de registre central de population. Ou alors les registres de population ne sont pas à jour, et, dans certains pays comme l'Italie et l'Espagne, on compte même sur le dénombrement de la population effectué dans le cadre d'un recensement classique pour leur mise à jour. À part les pays scandinaves, il y a les pays du Benelux qui ont, ou auront probablement bientôt, l'infrastructure nécessaire à la tenue d'un recensement réalisé à partir de registres.

Le troisième obstacle à la réalisation d'un recensement à partir de registres découle du deuxième. Si les systèmes de données ont besoin d'être grandement améliorés, et particulièrement s'il faut étendre l'usage des numéros personnels et instaurer l'obligation de communiquer tout changement d'adresse, on peut s'attendre à l'opposition des politiciens et du public sous prétexte que cela constituerait une atteinte à la vie privée et risquerait d'entraîner une perte de liberté. On peut se demander également si le public accepterait de se plier à la discipline bureaucratique qu'exige un bon système de registres. En outre, même si l'infrastructure était mise en place, le couplage d'enregistrements aux fins du recensement ou à d'autres fins statistiques pourrait s'avérer une étape délicate. Ces questions sont importantes, mais elles débordent largement du domaine de la statistique. Elles concernent plutôt la politique et l'administration. Nous allons à présent les examiner par rapport à l'expérience du Royaume-Uni.

5. LES SYSTÈMES DE REGISTRES AU ROYAUME-UNI

Pour les recensements décennaux du Royaume-Uni, on utilise les méthodes classiques. Le recensement de 1981 a probablement été le recensement le plus réussi depuis la Deuxième Guerre mondiale, réussite à laquelle ont contribué l'utilisation d'un questionnaire abrégé et l'omission d'une question controversée sur l'origine ethnique. Trois facteurs se conjuguent donc pour éloigner la possibilité d'un recensement réalisé à partir de registres: la réussite du recensement de 1981, les doutes qui règnent au sujet de la gamme de données qu'on pourrait tirer des dossiers administratifs et de la qualité de ces données et enfin l'absence d'un registre de population permettant de coordonner les sytèmes de dossiers.

Pour leur part, les statisticiens sont conscients des avantages que présentent les registres de population, tant sur le plan administratif que statistique. Les deux initiatives prises à cet égard au cours des soixante-dix dernières années, initiatives qui ont échoué l'une et l'autre, sont décrites dans les sections 5.1 - 5.4. Dans le cadre d'une réforme controversée des taxes locales, le gouvernement, bien qu'opposé à l'idée d'un registre central de population, est en train d'introduire des registres locaux de population dont les utilisations seront limitées (Section 5.5).

5.1 L'inscription nationale pendant les deux guerres mondiales: le comité de l'inscription de 1918

En Angleterre, la réflexion sur les registres de population remonte à la Première Guerre mondiale, c'est-à-dire à plus de soixante-dix ans. La National Registration Act de 1915 (loi sur l'inscription nationale) obligeait chaque adulte à avoir sur lui un certificat d'inscription national et à signaler aux autorités tout changement d'adresse. C'est ce qui a amené Sir Bernard Mallet, directeur de l'état civil, à envisager l'établissement d'un système permanent qu'il a décrit dans son discours présidentiel lors de la rencontre de la Royal Statistical Society du mois de novembre 1916. Il savait cependant qu'on risquait de lui reprocher de vouloir prussianiser les institutions anglaises.

Les arguments mis de l'avant par Sir Bernard Mallet ont été développés dans le rapport du comité qu'il a présidé et qui avait été créé par le gouvernement en 1918. De nombreuses années plus tard, Sir Mallet en a présenté les conclusions dans le discours présidentiel qu'il a prononcé devant la Eugenics Society (Mallet 1929). Ce qu'il a dit alors demeure vrai aujourd'hui:

> "Nous avons trouvé qu'il existe, en Angleterre, un nombre très considérable de registres utilisés à diverses fins particulières, dont la tenue revient très cher et qui portent dans certains cas sur de grands segments de la population. Ces registres ont été constitués en vertu de différentes lois et ils sont tenus par diverses autorités, dans une variété de régions, dans des buts indépendants les uns des autres, sans qu'on ait rien prévu qui permette d'en coordonner le contenu." (Cette citation et les autres qui apparaissent dans le présent article sont traduites de l'anglais)

Le comité proposait la tenue, au niveau local, de registres permanents de la population, d'où seraient tirés les renseignements servant à l'établissement des cartes d'identité. Un registre central sous forme d'index aurait permis faire le lien entre les registres locaux et de veiller à éviter les échevauchements et à supprimer les inscriptions inutiles. Un tel système aurait aussi permis de coordonner la tenue des registres remplissant une fonction spéciale comme les listes électorales, les registres de la fréquentation scolaire, le recensement décennal ainsi que les registres des naissances, des mariages et des décès. Il est intéressant de noter que le comité proposait, il y a près de soixante-dix ans de cela, que le recensement de la population soit rattaché au registre de population.

Dans le discours qu'il a prononcé en 1929, Sir Bernard Mallet a posé les principes que tout bon système devrait respecter: premièrement, chaque individu devrait être identifié correctement afin qu'on puisse (a) le rendre responsable des obligations qu'il a envers la collectivité et (b) lui garantir ses droits en tant que citoyen, que ceux-ci prennent la forme de privilèges qu'il peut exercer ou d'avantages dont il peut bénéficier; deuxièmement, les registres devraient permettre d'obtenir des renseignements statistiques et en particulier des chiffres courants sur la population des petites régions. L'analyse faite par Sir Mallet et les propositions qui en ont découlé constitueraient encore aujourd'hui une solution valable à la situation dans laquelle le Royaume-Uni se trouve aujourd'hui, quoique certains aspects ne seraient pas acceptables de nos jours. Notamment:

> "Les nombreux registres et enquêtes officielles qui actuellement n'ont aucun lien entre eux devraient être réunis en un seul système de données permettant d'établir, pour chaque individu, un **dossier** contenant les renseignements que l'État veut posséder à son sujet". (Mallet 1929)

Au grand regret de Sir Bernard Mallet, les recommandations contenues dans le rapport de son comité n'ont pas été suivies et, lorsque les lois du temps de guerre ont été suspendues, on a cessé de tenir un registre national jusqu'à ce que la Deuxième Guerre mondiale éclate.

Pendant la Seconde Guerre mondiale et les quelques années qui ont suivi, un système intégral d'inscription de la population a été en vigueur en Angleterre. On a créé un registre national dont on s'est servi pour délivrer à chaque personne une carte d'identité portant son numéro d'identité et son adresse. Les registres locaux étaient coordonnés au moyen d'un registre central indiquant, pour chaque personne inscrite, son nom, sa date de naissance, son numéro d'identité et le code correspondant à la région où elle habitait. Les changements d'adresse devaient être signalés au service des registres locaux. Le registre national a survécu jusqu'en 1952 où, emporté par la vague de libéralisme qui se propageait depuis la fin de la guerre, on a abandonné les cartes d'identité et l'obligation de communiquer les changements d'adresse.

5.2 Le registre central du service de santé national

Le registre central créé en 1939 pendant la période d'inscription nationale a été conservé après 1952, mais il joue depuis un rôle plus limité dans le cadre de l'administration du NHS (National Health Service/service de santé national). Rebaptisé NHSCR (National Health Service Central Register/registre central du service de santé national), il comprend à présent tous les résidents de l'Angleterre à l'exception des 1 ou 2 % qui sont nés à l'étranger et qui n'ont jamais figuré sur la liste des patients d'un médecin travaillant pour le NHS. Cependant, le NHSCR ne remplit pas les mêmes fonctions que les registres centraux de population qu'on trouve dans les pays du nord de l'Europe parce qu'il ne sert pas de point de référence par rapport auquel les autres organismes peuvent vérifier l'identité des personnes. Les autres organismes ne peuvent pas non plus reprendre les numéros de référence personnels du NHSCR et les utiliser pour leurs dossiers. En fait, les numéros d'identité inscrits dans le NHSCR servent uniquement aux fins du service de santé national. Il n'est pas possible d'étendre l'utilisation du registre pour les raisons suivantes:

1) Une forte proportion des données destinées au NHSCR ne sont pas assorties du numéro d'identification personnel. Du fait que le nom et la date de naissance peuvent difficilement servir d'identificateur à eux seuls, certaines données ne peuvent pas être intégrées aux dossiers existants du NHSCR. C'est notamment ce qui arrive avec 1 à 2 % des avis de décès. C'est en grande partie pour cette raison

et parce qu'on arrive pas à éliminer tous les immigrants du registre que le nombre de personnes inscrites dans le registre est supérieur à ce qu'il devrait être. Les cas, dont la proportion est estimée actuellement à environ 5%, devraient bientôt être moins nombreux, lorsque le registre sera informatisé.

2) Les adresses figurent au complet dans les registres locaux et sont représentées par un code régional dans le NHSCR. Cependant, dans la plupart des cas, les changements d'adresse ne sont effectués que lorsqu'un personne s'inscrit auprès d'un nouveau médecin, ce qui peut survenir des années après qu'elle a déménagé.

5.3 L'éventail des registres au Royaume-Uni

Comme dans tout autre pays industrialisé, les autorités du Royaume-Uni tiennent une vaste gamme de registres contenant des renseignements personnels. Les principaux registres concernent les actes d'état civil (naissances, décès, mariages et divorces), l'immigration et la naturalisation, le service de santé national, la sécurité sociale (tant cotisants que bénéficiaires, comme les chômeurs, les retraités et les enfants), les impôts des particuliers, les passeports, les listes électorales, la possession d'une voiture et le permis de conduire. Mais ces registres sont tenus indépendamment les un des autres par des organismes distincts qui ont chacun leur système d'attribution de numéro personnel. Il y a une exception, et c'est l'entente conclue dans le cadre du système de retenue à la source pour la collecte conjointe des impôts sur le revenu et des cotisations versées par les employés à la sécurité sociale. Un seul numéro personnel est utilisé, le numéro d'assurance sociale. À part ça, les systèmes de dossiers ne sont coordonnés d'aucune façon; le contenu des dossiers n'est pas uniformisé et il n'y a pas un numéro personnel qui est utilisé de façon générale. Les renseignements relatifs à l'identité d'une personne, habituellement le nom et la date de naissance, peuvent varier d'un registre à un autre et même à l'intérieur d'un même registre. Par conséquent, il arrive que les même personnes figurent plus d'une fois dans les registres, et il n'est pas certain qu'on puisse combiner les renseignements que ces derniers contiennent à des fins statistiques. Si on pouvait le faire, ce serait très coûteux. Les renseignements concernant les adresses sont, quant à eux, encore moins cohérents. Enfin, il n'existe pas de mécanisme permettant de mettre à jour simultanément tous les dossiers visés par certains changements, notamment les changements d'adresse, le changement de nom après le mariage ou même les décès. Comme l'a si bien dit Sir John Boreham lorsqu'il dirigeait le GSS (Government Statistical Service/service statistique de l'État) : "les renseignements ne sont jamais réunis convenablement...tout le système est plutôt boiteux" (Boreham 1985).

5.4 Étude réalisée dans les années 60

Le système actuel de registres indépendants ne peut pas être administré de façon efficace. Et du point de vue statistique, il souffre d'un double handicap: les adresses ne sont pas à jour et on ne peut pas apparier les dossiers. Le GSS s'est donc mis à la recherche d'une solution vers la fin des années 60. Il a étudié la possibilité de remplacer les divers types de numéros personnels par un seul numéro qui serait consigné dans le registre central, lequel contiendrait éventuellement la dernière adresse des personnes inscrites (Penrice et coll. 1968). Toutefois, les ministres ont décidé que ces idées étaient politiquement inacceptables, et ils ont mis fin à l'étude (House of Lords 1969).

5.5 Les registres pour la nouvelle charge foncière collective

Il semble qu'un des plus grands obstacles à la création d'un registre de population en Angleterre soit sur le point, en 1987, d'être surmonté puisqu'on s'apprête à obliger les citoyens à déclarer tout changement d'adresse. On ne créera pas pour autant de registre de population à proprement parler. Le gouvernement s'y oppose absolument.

La nouvelle obligation de déclarer tout changement d'adresse, qui constitue une dérogation de taille à la tradition britannique en temps de paix, découle de la décision du gouvernement de changer la façon dont les taxes locales sont perçues. Par le passé, le montant de la taxe locale imposée aux occupants d'une propriété était déterminé par la valeur locative de la propriété. Cette taxe va être remplacée par la CC (**Community Charge**/charge foncière collective), qui est une taxe uniforme que devra payer toute personne âgée de 18 ans et plus habitant un logement. Pour administrer la nouvelle taxe, il faudra tenir un registre local des adresses indiquant le nom des personnes âgées de 18 ans et plus qui y habitent. Le responsable des inscriptions pourra faire enquête et adresser des demandes de renseignements aux autorités locales, aux commissions du logement et au bureau des élections, mais c'est à l'individu qu'incombera la responsabilité de l'informer des changements à apporter au registre. La loi visant à introduire ce nouveau système a déjà été adoptée en Écosse (United Kingdom 1987) où elle entrera en vigueur en 1989, et le gouvernement compte passer une loi semblable pour l'Angleterre et le pays de Galle pendant la session en cours du Parlement.

Cependant, les registres de la CC seront des instruments primitifs comparativement aux registres de population des pays scandinaves et des pays du Benelux, pour les raisons suivantes:

1) Les registres de la CC n'incluront pas tout le monde; en seront exclus notamment les moins de 18 ans et les résidents d'une pension ou d'un établissement spécialisé.

2) Les registres (où seront consignés le nom, la date de naissance et l'adresse des individus) seront tenus localement et la marche à suivre à leur égard ne sera pas pleinement uniformisée. Ainsi, il n'y aura pas de registre central qui imposerait une même façon de décrire l'identité de chaque personne inscrite et assurerait la coordination des registres locaux (afin de faciliter les transferts d'une autorité compétente à une autre, par exemple).

3) Bien que la loi écossaise ne prévoie pas explicitement l'utilisation d'un numéro de référence personnel dans les registres, il a été recommandé dernièrement dans un rapport que les autorités locales écossaises créent un nouveau numéro, et un algorithme pouvant servir à l'établissement de ce numéro au moyen du nom et de la date de naissance a été proposé (Chartered Institute of Public Finance and Accountancy 1987). Le ministre qui parraine la loi pour l'Angleterre et le pays de Galle a dit, pour sa part, que même si l'Écosse utilisait un numéro personnel, l'Angleterre et le pays de Galle n'auraient pas besoin d'en faire autant (Howard 1987).

4) La loi écossaise précise qui peut avoir accès à quelles parties du registre. À part les autorités locales qui y auront accès pour pouvoir administrer la CC, un particulier pourra examiner les renseignements qui le concernent, le public pourra examiner les listes d'adresses et le nom des personnes correspondant à ces adresses ("mais pas en vue de s'assurer si une personne habite bien à une certaine adresse") et le directeur général des élections aura accès aux registres pour les besoins liés à ses fonctions. Personne d'autre n'y aura accès.

Le rejet par le gouvernement d'un registre de population qui permettrait de coordonner les dossiers administratifs est justifiée dans le livre vert sur le projet de CC (Her Majesty's Government 1986). Les auteurs de ce document citent lecas de pays qui "ont fusionné leur divers registres et s'en servent centralement à plusieurs fins administratives différentes". Ils ajoutent: "La tradition britannique est différente. Les registres sont tenus séparément à des fins différentes par les organismes qui en ont besoin dans un but particulier...Il n'y aura pas registre national." La comparaison qui est faite entre la pratique observée dans les autres pays et celle proposée pour le Royaume-Uni est fautive, car, dans les autres pays, les organismes tiennent des registres **distincts** mais s'adressent

au service chargé du registre central pour identifier la personne à laquelle ils ont affaire. À mon avis, la déclaration "il n'y aura pas de registre national" découle d'un axiome politique et non d'une analyse rationnelle.

La création des registres de la CC représente peut-être une occasion manquée de constituer un registre de population efficace. Mais en fait, le projet de CC n'est pas le prétexte idéal pour cela. En effet, pour être efficace, le registre de population doit servir à plusieurs fins, et plus il y en a, mieux c'est. Il ne devrait pas avoir une seule fonction, surtout lorsque celle-ci est de permettre de prélever une taxe que beaucoup trouvent lourde et à laquelle un grand nombre de gens chercheront à se dérober. De plus, la CC est une mesure politique controversée parce qu'elle n'a pas le même effet sur les divers segments de la société: elle entraînera de façon générale un transfert des ressources des pauvres aux riches.

Il y a donc plusieurs raisons de mettre en doute l'efficacité opérationnelle des registres qui doivent être créés dans le cadre du projet de CC: le but unique et controversé des registres; le fait que toute la population n'y sera pas représentée (l'omission de certains groupes); l'absence de registre central permettant de coordonner les registres locaux; et enfin, le fait que ce sont apparemment le nom et la date de naissance qui serviront d'identificateur, plutôt qu'un numéro personnel permanent. Les autorités locales ont exprimé leurs réserves à l'égard de ce projet en décrivant les problèmes auxquels elles devront faire face au moment de constituer les registres (Rating and Valuation Association 1987). Il semble bien que le gouvernement se soit engagé à introduire une nouvelle loi fiscale sans avoir pensé à tous les aspects pratiques de son application.

Il y a un autre aspect du projet de CC qui est inquiétant, et c'est l'effet qu'il aura sur la réaction du public au recensement de la population de 1991. Parmi ceux qui voudront se dérober à la CC, un bon nombre essaieront aussi de se soustraire au recensement, malgré les efforts que feront les responsables de ce dernier pour les convaincre que les données recueillies ne seront pas communiquées à d'autres organismes. Par contre, si le questionnaire dit de façon trop explicite: "LES RENSEIGNEMENTS QUE VOUS DÉCLAREREZ NE SERONT PAS COMMUNIQUÉS AUX ORGANISMES CHARGÉS D'ADMINISTRER LES IMPÔTS, LA SÉCURITÉ SOCIALE, LES CHARGES COLLECTIVES,...", les responsables du recensement ne risquent-ils pas de donner l'impression qu'ils ne condamnent pas, ou même qu'ils encouragent, l'évasion et la fraude?

5.6 La situation au Royaume-Uni

Indépendamment de la CC, le climat actuel au Royaume-Uni est plutôt hostile à l'idée d'un registre de population. Mais on peut mentionner deux aspects positifs. Premièrement, la **Data Protection Act** de 1984 (loi sur la protection de la confidentialité des données) a introduit des mesures de protection touchant les renseignements personnels stockés dans les ordinateurs qui sont semblables à celles prévues par la Convention de 1981 du Conseil de l'Europe (Conseil de l'Europe 1981). En fait, le principal but visé par le gouvernement lorsqu'il a adopté la loi de 1984 était un but commercial: il s'agissait de montrer aux autres pays susceptibles d'envoyer leurs données au Royaume-Uni pour les faire traiter que ces données y seraient en sécurité. La protection de la vie privée était un but secondaire. Deuxièmement, le GSS, qui serait changé de voir à certains aspects du fonctionnement des registres de population, a toujours été irréprochable en ce qui a trait à la protection de la confidentialité des données; il a d'ailleurs publié un code de conduite à ce sujet (Government Statistical Service 1984). Le fait que le GSS soit décentralisé a également contribué à établir la réputation d'intégrité que cet organisme s'est méritée, l'échange de données, même à a des fins statistiques, étant empêché par des obstacles tant juridiques qu'administratifs. Il faudrait d'ailleurs retirer ces obstacles pour bénéficier des retombées statistiques qu'un registre de population pourrait avoir.

Pour ce qui est des aspects négatifs, la dépendance du GSS à l'égard du gouvernement central contraste avec l'autonomie relative dont jouissent notamment les organismes statistiques du Danemark et des Pays-Bas. Cette dépendance pourrait ébranler la confiance que le public accorde au GSS relativement au traitement des données. L'impression que peut donner le GSS d'être la marionnette du gouvernement central a été renforcée par les conclusions du rapport Rayner au début des années 80. À la suite de ce rapport, on a demandé au GSS d'accorder plus d'importance aux besoins du gouvernement central, et ce aux dépens des autorités locales, des entreprises, du milieu universitaire et du grand public.

Au Royaume-Uni, le principal obstacle à l'instauration d'un registre de population est la résistance traditionnelle du public à toute mesure gouvernementale perçue comme étant autoritaire ou bureaucratique. On peut s'attendre que le lobby de la protection de la vie privée mène l'opposition contre toute nouvelle obligation, pour le public, de déclarer des renseignements, tout ajout au nombre de données à caractère personnel que détient déjà l'État et tout projet de couplage de données. L'opposition ne tient pas compte des coûts et des injustices qu'entraîne la gestion inefficace des données. Elle ne tient pas compte non plus des freins que les lois sur la protection de la confidentialité des données et la liberté de l'information peuvent mettre, si elles sont appliquées, à la mauvaise utilisation des données à caractère personnel, ou alors elle sous-estime ces freins. La perception que le public a du gouvernement en place a renforcé ces dernières années les craintes qu'il éprouve à l'idée de fournir davantage de renseignements personnels à l'État: le gouvernement britannique est perçu comme ayant l'obsession du secret et comme cherchant à garder tout le pouvoir entre ses mains. Ainsi, non seulement n'y a-t-il pas de loi sur la liberté de l'information au Royaume-Uni, mais en plus tous les renseignements concernant l'État sont en principe protégés par une loi d'application générale, la **Official Secrets Act** de 1911 (loi sur les secrets officiels). Peter Hennessy, écrivain et personnalité des médias, affirme que les gouvernements britanniques ont, de tous les gouvernements occidentaux, adopté les mesures les plus rigoureuses pour protéger le secret des dossiers administratifs (Hennessy 1987). D'ailleurs, à cause d'événements récents, la nécessité pour les services secrets de justifier leurs actions a fait l'objet dediscussions visant à mieux définir cette responsabilité. S'exprimant sur la l'ensemble des activités gouvernementales, William Plowden, directeur général du Royal Institute of Public Admnistration, a dit qu'un gouvernement britannique contemporain, appuyé par une bonne majorité à la Chambre des communes, peu menacé par des commissions parlementaires qui aboient mais ne mordent pas, solidement protégé par la loi sur les secrets officiels, est, de tous les pouvoirs exécutifs des pays industrialisés, un de ceux qui sont le moins tenus de rendre compte de leurs décisions (Plowden 1987).

Le public se méfie donc de tout nouveau projet visant à créer un registre de population. Et, ainsi que nous l'avons vu, le gouvernement actuel a exprimé son opposition à l'idée d'un registre complet : comme celui des États-Unis, il a montré qu'il était décidé à lutter contre toute forme d'ingérence de l'État dans la vie des citoyens. Un de ses principaux objectifs est de réduire la taille et l'influence du secteur public, et il accorde parfois plus d'importance à ce principe qu'à celui de la rentabilité. On voit donc que les préoccupations du public relativement à la protection de la vie privée, l'idéologie politique et le manque de ressources sont trois facteurs qui se conjuguent pour empêcher la création d'un registre complet, lequel permettrait pourtant de réaliser des économies considérables et de rendre la société plus équitable. À vrai dire, les faits concernant ces questions n'ont pas été présentés de façon équilibrée, et il n'y a pas eu de débat public à ce sujet, depuis une cinquantaine d'années.

6. L'INITIATIVE AUSTRALIENNE: LES CARTES D'IDENTITÉ

Je ne connais pas bien le tempérament australien ni la situation politique dans ce pays, mais je suppose que l'opposition à un gouvernement bureaucratique est aussi forte là-bas qu'au Royaume-Uni. Malgré cela, le gouvernement australien a présenté un projet de loi visant la délivrance d'une carte d'identité à tous les citoyens, l'AC (Australia Card/carte australienne). Les raisons sont purement administratives: il s'agit de réduire l'évasion fiscale, la fraude en matière de la sécurité sociale et l'immigration illégale. La carte australienne porterait le nom, la photo, la signature et le numéro (numéro de référence personnel propre à cette carte d'identité de chaque personne), mais non son adresse. La carte australienne serait rattachée à un registre qui contiendrait les adresses et les dates de naissance, et auquel n'auraient accès que certains ministères ou services publics.

L'Australia Card Bill de 1986 (projet de loi concernant la carte australienne) a été approuvé par la Chambre des députés mais non par le Sénat, dont le parti au pouvoir ne détient pas la majorité des sièges. Apparemment, c'est en partie à cause de ce rejet que l'élection de 1987 a eu lieu. Après la victoire du parti au pouvoir, le Parlement devait être saisi du projet de loi de nouveau, mais, celui-ci aurait été retiré a cause de sérieuse imperfection légale. Je pense qu'il est quand même utile de décrire ici les dispositions qu'il contenait.

Le registre de l'AC serait un registre central de population. Cependant, il serait moins complet que ceux des pays scandinaves pour deux raisons principales:

1) En vertu du projet de loi, les citoyens ne seraient pas obligés de signaler leurs changements d'adresse aux responsables du registre. Si j'ai bien compris, on espérait que la plupart des changements d'adresse seraient communiqués à au moins un des organismes publics liés au projet, lequel en ferait part au service chargé de la mise à jour du registre de l'AC.

2) Le projet de carte australienne ne serait pas aussi polyvalent que plusieurs des registres de population européens. À cause des inquiétudes du public relativement à la protection de la confidentialité des renseignements personnels et au couplage incontrôlé des données, seuls les organismes publics qui s'occupent des impôts, de la sécurité sociale et de l'assurance-maladie auraient accès au registre de l'AC, et encore, uniquement pour vérifier les identités.

Le projet de loi précisait dans quelles situations on pouvait demander à quelqu'un de montrer sa carte; ce serait notamment à l'occasion d'un large éventail d'opérations financières et au moment de commencer un nouvel emploi, de se faire soigner à l'hôpital et de demander des prestations ou des services offerts par l'assurance-maladie ou la sécurité sociale. Il serait illégal de demander à quelqu'un de montrer sa carte dans d'autres circonstances.

Comme mesure supplémentaire de protection des renseignements personnels, le projet de loi prévoyait la création d'un organisme chargé de la protection de la confidentialité des données. Toutefois, le gouvernement est d'avis qu'il faut trouver le juste milieu entre la protection de la vie privée et les pertes que la fraude fiscale fait subir à l'État. Il estime que le projet d'AC coûterait $0.8 milliard en dix ans, mais que cette dépense serait plusieurs fois compensée par les montants de $4.1 milliards et de $1.4 milliard que le fisc et la sécurité sociale pourraient récupérer respectivement, permettant à l'État de réaliser pendant cette période des économies nettes de l'ordre de $4.7 milliards (Australian House of Representatives 1986).

Les remarques formulées par le ministre de la Santé au Parlement sont révélatrices des buts poursuivis par les ministres et de l'engagement politique sans équivoque pris par le gouvernement:

"Je tiens à saisir le Parlement aujourd'hui... d'une réforme qui s'est longuement faite attendre et qui vise à apporter justice et équité à tous les Australiens."
"Il ne fait aucun doute que la carte australienne permettra de contrôler la fraude fiscale; il ne fait aucun doute qu'elle contribuera à préserver l'intégrité de notre régime de sécurité sociale; il ne fait aucun doute qu'elle s'avérera une arme efficace contre l'immigration illégale; il ne fait aucun doute qu'en permettant de remonter la piste empruntée par l'argent, elle sera un instrument précieux de lutte contre la criminalité des entreprises et le crime organisé."
"Il faut protéger les citoyens contre l'intrusion du gouvernement dans leur vie privée, c'est un principe absolu. Mais il faut également protéger les citoyens contre ceux qui se cachent cyniquement derrière le droit à la confidentialité des renseignements personnels pour assumer une fausse identité et frauder la collectivité."
"Notre pays établira un système d'identification avant la fin du siècle, c'est inévitable."

Bien que le projet de loi concernant la carte australienne ait été retiré, ce n'est pas encore la fin de cette histoire, car le gouvernement continue à chercher d'autres moyens qui lui permettront de supprimer toute possibilité de fraude à l'impôt et à la sécurité sociale.

6.1 Cartes d'identité

Le projet australien met principalement l'accent sur la carte d'identité comme moyen de vérifier l'identité des gens et non sur le registre ou sur le numéro personnel. Dans certains pays d'Europe, la délivrance d'une carte d'identité est liée à la tenue d'un registre de population, le système en place en Belgique étant l'un des plus perfectionnés. Il est évident que la carte d'identité constitue une mesure supplémentaire de sécurité en autant qu'elle n'est ni contrefaite ni volée. Dans certains pays, comme la France, il n'y a aucun lien entre la carte d'identité et le registre de population.

Dans les pays où l'on n'a jamais utilisé de cartes d'identité en temps de paix, celles-ci sont perçues comme un symbole des régimes autoritaires et comme une atteinte aux libertés civiques. Voilà peut-être une des raisons pour lesquelles le projet australien a suscité une telle opposition de la part du public. Cependant, on peut profiter d'une bonne partie des avantages des registres de population sans avoir recours aux cartes d'identité à condition que les citoyens connaissent leur numéro personnel et le donnent aux autorités lorsqu'ils ont affaire à elles. C'est ce qui se passe au Danemark et en Suède où les cartes d'identité n'existent pas mais où les registres de population sont malgré tout efficaces tant du point de vue administratif que statistique.

Un pays comme le Royaume-Uni ne devrait pas avoir peur de remédier au manque de cohérence de ses dossiers administratifs sous prétexte qu'une critique mal informée confonde le remède nécessaire, en l'occurrence la création d'un registre de population, avec son complément qui, lui, est facultatif, soit les cartes d'identité.

7. CONCLUSION

La création d'un registre de population contenant des adresses à jour et des numéros de référence personnels qu'on retrouverait dans les fichiers administratifs reviendrait tout juste à mettre de l'ordre dans un système boiteux, car même dans le plus boiteux des systèmes, il faut bien que le citoyen s'identifie et informe les autorités compétentes de ses changements d'adresse. Il y a malgré tout des gens que l'idée d'un registre de population inquiète parce qu'ils y voient une menace à la liberté et à la vie privée et parce qu'ils craignent qu'un gouvernement autoritaire ou tyrannique n'abuse du pouvoir accru

dont l'État pourrait ainsi disposer. Pourtant il existe des antidotes auxquels il faudrait avoir recours, notamment des règles efficaces permettant de protéger la confidentialité des données et des lois sur la liberté de l'information.

En revanche, un système de dossiers bien coordonné aurait des avantages politiques qu'on a trop tendance à oublier. Placés par ordre d'importance, les deux premiers seraient, à mon avis:

1) Un frein à la fraude, au crime et à l'immigration illégale.

2) Une société plus juste dont les devoirs seraient mieux partagés et dont les privilèges iraient uniquement à ceux qui y ont droit. Autrement dit, la liberté ne devrait pas signifier le loisir de frauder le reste de la collectivité.

Plus loin, j'ajouterais:

3) Les économies financières que l'État pourrait réaliser. Si les dossiers étaient plus précis, les frais administratifs seraient inférieurs, il serait possible de prélever davantage de taxes et le montant des prestations versées à tort serait réduit, comme le montrent les chiffres australiens (Section 6).

4) Le gouvernement aurait plus de choix quand aux moyens à sa disposition pour mettre en oeuvre ses politiques. Si, par exemple, il existait déjà un registre de population au Royaume-Uni, le gouvernement n'aurait pas à en créer un **spécial** pour son projet de charge collective et il pourrait surveiller l'entrée des immigrants en ajoutant aux contrôles exercés aux aéroports et aux ports de mer un contrôle à partir de l'adresse du domicile.

5) Il y aurait d'autres avantages découlant d'une meilleure vérification des identités. L'ancien registraire général citait comme exemple la possibilité de mieux s'assurer si deux personnes ont le droit de se marier. Le nombre de numéros de référence à donner serait inférieur, et peut-être aussi le nombre de cartes en plastique à avoir sur soi.

6) Les statistiques seraient meilleures, mais, à cet égard, je donnerai des précisions au paragraphe suivant le tableau 1.

Il y a plusieurs réponses à donner aux personnes qui pensent tout de suite Grand Frère et état policier lorsqu'on leur parle de registres de population, et les arguments que je viens d'énumérer en sont une. En fait, ces personnes n'auraient peut-être pas tort s'il n'existait aucune mesure de protection et si les registres tombaient effectivement dans de mauvaises mains. Mais ces registres peuvent aussi représenter la voie privilégiée vers une société plus équitable. La question qui se pose est la suivante: quel genre de société voulons-nous? Une société qui encourage la fraude, l'évasion fiscale et le crime, ou du moins ferme les yeux sur ces délits? Les ministres australiens ont donné comme exemple un homme qui a été reconnu coupable d'avoir touché, tous les quinze jours, plus de cinquante chèques d'assurance-chômage (Australian House of Representatives 1986). Au Royaume-Uni, un avocat, membre du Parlement, vient d'être condamné à la prison (avec droit d'appel) parce qu'il avait enfreint les règlements en souscrivant plusieurs fois à des actions, changeant chaque fois de nom, d'adresse et de compte en banque; l'argument invoqué par la défense était qu'il s'agissait d'une pratique courante.

Une autre façon de répondre à l'accusation de totalitarisme consiste à regarder l'utilisation qui est faite des registres de population dans les autres pays. Dans le tableau 1, quinze pays, soit tous les pays de l'Europe occidentale à l'exception de l'Autriche et de la Suisse, sont répartis en quatre groupes selon le genre de système de registres qu'ils emploient. Les six pays du groupe A possèdent le système le plus efficace: leurs dossiers administratifs sont coordonnés au moyen de registres de

population. Les quatre pays du groupe B sont dans une situation intermédiaire. Dans les trois pays du groupe C, il existe des registres de population uniquement au niveau local et leur qualité n'est pas toujours bonne. Enfin, l'Irlande et le Royaume-Uni, soit le groupe D, viennent en dernier, avec le système le moins perfectionné. Si le Royaume-Uni choisissait une ligne de conduite rationnelle et réaliste et adoptait le système en vigueur dans le groupe A, il ne se trouverait pas en compagnie de pays totalitaires.

Je dois à présent nuancer ce que j'ai dit dans la section 6 lorsque j'ai indiqué que les statistiques sont meilleures quant on met en place un système de registres bien coordonnés. La conséquence **directe** d'une telle action est certainement la production de meilleures statistiques, notamment la production de statistiques régionales fiables à intervalles réguliers. Mais si, comme conséquence **indirecte**, les pressions visant à remplacer le recensement classique par un recensement réalisé entièrement à partir de registres se mettent à augmenter de façon irrésistible, les avantages sont alors accompagnés d'inconvénients. À la diminution des coûts, du fardeau de réponse imposé au public et des risques de sabotage s'oppose la détérioration possible de la portée et de la qualité des résultats de recensement, notamment dans le domaine économique et dans celui du logement. Il y a alors le danger que les dossiers administratifs rendent de moins en moins bien la complexité et l'évolution des modes de vie actuels, c'est-à-dire ce qu'un recensement classique essaye de mettre en évidence, par exemple l'augmentation du travail à temps partiel et du travail autonome, la croissance du nombre de résidences secondaires et le relâchement des liens familiaux et des liens à l'intérieur des ménages. C'est là que l'expérience des pays scandinaves (section 3) se révèle utile.

Les statisticiens ne sous-estimeront probablement pas l'utilité d'avoir de meilleures statistiques. Cependant, les considérations d'ordre politique et administratif ont plus de poids dans le débat qui entoure les registres de population. Ce sont donc les décideurs, les politiciens et le public qui doivent participer à ce débat. Au Royaume-Uni, celui-ci devrait porter sur le bien-fondé, voire la faisabilité, d'un registre de population qui remplirait une seule fonction, en l'occurrence le prélèvement de la CC, et qui serait totalement indépendant des autres registres existants, par opposition à un registre de population polyvalent qui, par définition, présenterait de nombreux avantages.

Je crois néanmoins qu'il convient de soulever cette question auprès de statisticiens, et ce pour trois raisons. Premièrement, les statisticiens comprennent à la fois les problèmes techniques et les questions d'intérêt plus général, de sorte qu'ils peuvent indiquer la voie à suivre. Ainsi, au Royaume-Uni, les deux initiatives antérieures concernant les registres de population se sont situées dans un contexte dont la dimension statistique n'était pas exclue (section 5). Deuxièmement, il se peut qu'on donne aux organismes statistiques la responsabilité d'administrer le principal mécanisme de coordination, notamment le registre central de population; c'est le cas de L'INSEE en France et du SSB en Norvège. Troisièmement, les statisticiens gagneraient à disposer de données fiables.

J'espère que les statisticiens feront connaître leur opinion. La question des registres est pleinement d'actualité, particulièrement dans les pays aussi "sous-développés" à cet égard que le Royaume-Uni et l'Australie. Les statisticiens au service de l'État devraient réfléchir à ce commentaire sur l'éthique professionnelle formulé à l'intention du US Bureau of the Census; il a été écrit dans un contexte différent à l'occasion de la table ronde sur la méthodologie des recensements décennaux qui a lieu en 1984 (Citro et Cohen 1985), mais il est encore très pertinent:

> "Nous reconnaissons que le climat actuel n'est pas favorable à l'introduction de nouveaux programmes, mais nous pensons que les statisticiens ont la responsabilité de décrire les faits et de recommander les mesures qu'ils jugent raisonnables."

Tableau 1

Caractéristiques des registres de population utilisés dans 15 pays, indiquées au moyen d'un "x". Pour plus de détails, voir Redfern 1987.

	Registres locaux de population	Registre central de population permettant de coordonner les dossiers administratifs	Numéros de référence personnels
A Système complet de registres de population			
Belgique	x	x	x
Danemark	x	x	x
Finlande	x	x	x
Luxembourg	x	x	x
Norvège	x	x	x
Suède	x	x	x
B Groupe intermédiaire			
France	.	x	x
Pays-Bas	x	.	x
Portugal	.	x	x
Espagne	x	(x)	x
C Registres locaux de population seulement			
Rép. fédérale d'Allemagne	x	.	.
Grèce	x	.	.
Italie	x	.	.
D Aucun registre de population			
Irlande	.	.	.
Royaume-Uni	.	.	.
Nombre de pays possédant cette caractéristique	11	8+	10

REMERCIEMENTS

Pour tous les renseignements qui ont servi à la préparation de cette communication, je tiens à remercier les bureaux statistiques de l'Australie, de la Finlande et de la Norvège et, bien entendu, les pays ayant collaboré à l'étude que j'ai réalisée pour la CEE. Quant aux erreurs et lacunes, j'en assume l'entière responsabilité.

BIBLIOGRAPHIE

Australian House of Representatives (1986). The Honorable Neal Blewett MP, à la discussion en deuxième lecture sur l'Australian Card Bill, 1986.

Boreham, J. (1985). Cité dans How Whitehall plays the Numbers Game, *The Times*, Londres, 30 juillet 1985.

Chartered Institute of Public Finance and Accountancy (1987). Preparation of a specification of user requirements for the system of community charge in Scotland. Document non publié CIPFA Services, Londres.

Citro, C.F., et Cohen, M.L. (eds.) (1985). The Bicentennial Census: New Directions for Methodology in 1990. Washington, D.C.: National Academy Press.

Conseil de l'Europe (1981). Convention pour la protection des personnes à l'égard du traitement automatique des données à caractère personnel.

Government Statistical Service (1984). The Government Statistical Service code of practice on the handling of data obtained from statistical inquiries. Cmnd 9270, Her Majesty's Stationary Office.

Heinonen, R., et Laihonen, A. (1987). Some new solutions and methods for census data production: Finnish experiences from the 1985 census. Communication présentée au séminaire de la CEE/CES sur les aspects informatisés des recensements de la population et du logement, Belgrade.

Hennessy, P. (1987). *The Independent*, Londres, le 1 avril 1987.

Her Majesty's Government (1986). Paying for local government. Cmnd 9714, Her Majesty's Stationary Office.

House of Lords (1969). The Lord Chancellor, Lord Gardiner, dans Hansard, 3 décembre 1969.

Howard, M. (1987). Michael Howard, ministre d'État à l'Environnement, à la BBC, Radio 2 le 30 septembre 1987.

Jensen, P. (1983). Towards a register-based statistical system -some Danish experience. *Statistical Journal of the United Nations Economic Commission for Europe*, 1, 341-365.

Johansen, S. (1987). Input dubious — output OK. Communication présentée à la conférence démographie européenne, Jyväskylä, Finlande.

Johansson, S. (1987). Statistics based on administrative records as a substitute or a valid alternative to a population census. Communication présentée à la rencontre de l'IIS, Tokyo.

Laihonen, A., et Myrskylä, P. (1987). Use of registers and administrative records in population censuses in Finland. Communication présentée à la Conférence démographique européenne, Jyväskylä, Finlande.

Mallet, B. (1917). The organization of registration in its bearing on vital statistics. *Journal of the Royal Statistical Society*, Partie 1, 80, 1-24.

Mallet, B. (1929). Reform of vital statistics. *Eugenics Review*, 21, 87-94.

Penrice, G., Redfern, P., Evans, D., Whitehead, F.E., Bishop, H.E., et Rudoe, W. (1968). Discussion de Papers on social and medical statistics, *Journal of the Royal Statistical Society*, Ser. A, 131, 26-33.

Plowden, W. (1987). The battles of ideology that ill serve the public. Dans *The Independent*, Londres, 24 juin 1987.

Rating and Valuation Association (1987). Community charge, poll tax: the facts. Document non publié, Rating and Valuation Association, Londres.

Redfern, P. (1987). A study of the future of the census of population: alternative approaches. Eurostat Theme 3 Series C, bureau statistique des communautés européennes, Luxembourg.

United Kingdom (1987). Abolition of Domestic Rates etc. (Scotland) Act, 1987.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

LA PROTECTION DES DONNÉES FISCALES

H.J. LAGASSÉ[1]

RÉSUMÉ

Pour pouvoir utiliser les données fiscales à des fins autres que l'administration du régime fiscal, il faut arriver à composer avec des objectifs contradictoires. Le premier objectif de Revenu Canada, Impôt est de veiller à ce que les données fournies par les déclarants demeurent confidentielles et ne servent qu'à des fins permises par la loi. Le cadre juridique dans lequel évolue Revenu Canada, Impôt ainsi que les obligations et les exigences de ce ministère limitent considérablement l'utilisation de telles données à des fins statistiques.

1. INTRODUCTION

Le régime fiscal canadien repose sur le principe de l'"autocotisation", selon lequel les particuliers et les entreprises fournissent à Revenu Canada, Impôt les renseignements qui les concernent et qui concernent leur situation financière. Ce régime fonctionne dans la mesure où les déclarants sont très nombreux à accepter de fournir ces renseignements de leur plein gré. Pour cela, il faut qu'ils soient convaincus que les renseignements personnels et financiers fournis seront gardés secrets et ne serviront qu'à des fins permises par la loi. Diverses mesures de nature législative, administrative et technique ont été adoptées pour garantir le respect de ces conditions.

En revanche, si l'on veut augmenter l'efficacité et l'efficience et alléger le fardeau de réponse imposé aux particuliers et aux entreprises, il est alors souhaitable de pouvoir communiquer les données fiscales à certains organismes gouvernementaux dans des conditions prescrites. Certaines lois, dont celle de l'impôt sur le revenu, permettent à Revenu Canada, Impôt de le faire dans certains cas.

Il faut donc maintenir l'équilibre entre des valeurs et des buts qui s'opposent, et c'est là un défi permanent non seulement pour les fonctionnaires de Revenu Canada, Impôt, mais aussi pour ceux des autres ministères et organismes fédéraux susceptibles de recevoir des données fiscales. Ces derniers sont également liés par les dispositions de la Loi de l'impôt sur le revenu concernant la confidentialité, et ils ne peuvent utiliser et communiquer les données fiscales que dans les conditions permises par cette loi, sans quoi ils risquent de subir les sanctions qu'elle prévoit.

Étant donné que la raison d'être du régime fiscal n'est pas, ni en premier, ni même en deuxième lieu, de fournir des données aux ministères et organismes publics autres que Revenu Canada, Impôt, les fonctionnaires de ce ministère et d'ailleurs doivent être sensibles aux avantages et coûts réels, pour Revenu Canada, Impôt, de la diffusion des

[1] H.J. Lagassé, Directeur Général Systèmes, Revenu Canada Impôt, 875, chemin Heron Pièce 5140, Ottawa, (Ontario). K1A 0L8.

données fiscales à des fins autres que l'administration du régime fiscal. Les fonctionnaires de Revenu Canada, Impôt, doivent également s'assurer que les autres utilisateurs autorisés connaissent et acceptent leurs obligations et leur responsabilité en ce qui a trait à ces coûts; il y a, d'une part, les coûts monétaires et, d'autre part, le prix, intangible mais néanmoins réel, qu'il faudrait payer si le public n'avait plus la certitude que les renseignements personnels et financiers demeurent confidentiels.

Dans cette communication, je vais passer en revue les principales dispositions législatives et administratives qui concernent la protection[2] données fiscales et la communication de ces données à certains ministères et organismes fédéraux à des fins précises **autres que** l'administration du régime fiscal. J'aborderai toutes les utilisations prévues, puisque ce ne sont pas seulement les utilisations statistiques qui soulèvent des considérations d'ordre organisationnel et administratif. Cependant, l'utilisation des données fiscales à des fins statistiques pose quelques problèmes particuliers du point de vue du fisc, dont je parlerai brièvement.

2. CADRE JURIDIQUE

Il y a des lois fédérales qui permettent à certains ministères ou organismes fédéraux d'obtenir et d'utiliser à des fins précises des données fiscales qui sont révélatrices de l'identité des personnes ou entreprises auxquelles elles se rapportent. Les fins en question varient de l'administration, l'application ou l'évaluation de certaines lois ou de programmes gouvernementaux à la compilation de statistiques, en passant par l'analyse de la politique fiscale. Par exemple, la Loi sur la statistique permet à Statistique Canada d'avoir accès aux données fiscales à des fins statistiques, tandis que la Loi de l'impôt sur le revenu permet aux fonctionnaires de Revenu Canada, Impôt de communiquer des données fiscales aux fonctionnaires du ministère des Finances qui s'en servent pour analyser la politique fiscale. Par contre, certaines lois comme la Loi de l'impôt sur le revenu et celle sur l'accès à l'information contiennent des dispositions limitant la communication et l'utilisation des données fiscales. Dans certains cas, il faut consulter et la Loi de l'impôt sur le revenu et les autres lois pour déterminer si un ministère ou organisme fédéral peut recevoir et utiliser des données fiscales à certaines fins en particulier.

Dans la présente section, je donnerai un aperçu des principales dispositions légales concernant l'utilisation et la protection des données fiscales. Il est utile de commencer par la Loi de l'impôt sur le revenu, dont je résumerai en langage courant les dispositions pertinentes.

L'article 241 de la Loi de l'impôt sur le revenu remplit plusieurs fonctions. Tout d'abord, il pose le principe fondamental de la confidentialité en interdisant aux "fonctionnaires" et "personnes autorisées" de communiquer ou permettre que soit communiqué un renseignement obtenu par le ministre ou en son nom **sauf comme l'autorise ce même article ailleurs.** La Loi de l'impôt sur le revenu définit un "fonctionnaire" comme étant toute personne employée à une fonction de responsabilité ou occupant un tel poste au service de Sa Majesté, ou toute personne précédemment ainsi employé ou ayant précédemment occupé un tel poste. L'expression "personne autorisée" revêt un sens semblable dans le contexte de l'administration de cette loi.

Étant donné cette définition générale du mot "fonctionnaire", il ne serait pas prudent de s'arrêter là puisqu'on pourrait très bien imaginer des fonctionnaires communiquant des

[2] Quand on parle de protection, on pense à la protection des renseignements personnels, à la question de la confidentialité et à celle de la sécurité.

données fiscales à d'autres fonctionnaires ou personnes autorisées, ce qui aurait pour effet de compromettre rapidement le principe de la confidentialité. J'ai l'impression que les personnes qui ont rédigé la Loi de l'impôt sur le revenu étaient conscientes de ce risque. Cette loi contient donc des dispositions selon lesquelles **toute personne** qui contrevient au principe fondamental de la confidentialité ou qui communique un renseignement pour qu'il serve **à une fin autre que celle pour laquelle il a été recueilli** comme une infraction. Autrement dit, si un fonctionnaire du Ministère recueille des données fiscales à une fin précise autorisée par la Loi, il **n'a pas** le droit de communiquer ces données à qui que ce soit à moins que cette personne n'en fasse un usage conforme à celui qui a été autorisé. On appelle cette disposition importante le principe de l'usage compatible

Une fois établis le principe fondamental de la confidentialité et le principe de l'usage compatible, l'article 241 de la Loi de l'impôt sur le revenu permet aux "fonctionnaires" ou "personnes autorisées" de communiquer des données fiscales dans certains cas **exceptionnels.**

Voici quelques exemples de cas où la Loi de l'impôt sur le revenu permet à un fonctionnaire ou à une personne autorisée de communiquer des données fiscales:

- à un fonctionnaire du ministère des Finances, uniquement aux fins d'évaluer et de formuler la politique fiscale;
- à un fonctionnaire du ministère du Revenu national, Douanes et Accises, uniquement aux fins de l'application et de l'exécution de certaines lois administrées par ce ministère, notamment la Loi sur les douanes et la Loi sur l'accise;
- à un fonctionnaire de la Commission de l'emploi et de l'immigration du Canada ou du ministère de l'Emploi et de l'Immigration, uniquement aux fins de l'application, de l'évaluation ou de l'exécution de la Loi sur l'assurance-chômage ou d'un programme d'emploi prescrit.

Les statisticiens au service de l'État sauront que l'article 241 de la Loi de l'impôt sur le revenu contient également une disposition permettant de communiquer certaines données fiscales uniquement dans le but de permettre à certains ministères ou agences de recueillir des données statistiques à des fins d'analyse et de recherche. Les données en question sont limitées; il s'agit du nom, de l'adresse, de la profession ou du genre d'entreprise d'un contribuable.

L'article 241 contient également une disposition générale permettant de communiquer des données fiscales à toute personne "qui y a par ailleurs légalement droit". Cette disposition prend toute son importance lorsqu'on la considère en même temps que d'autres textes de loi comme la Loi sur la cession du droit au remboursement en matière d'impôt, la Loi sur le vérificateur général et la Loi sur la statistique. En bref, ces lois permettent aux fonctionnaires des organismes visés d'avoir accès aux données fiscales à des fins propres à leur organisme respectif; c'est donc d'eux qu'il s'agit dans l'article 241 de la Loi de l'impôt sur le revenu lorsqu'on parle des personnes "qui y a par ailleurs légalement droit".

Comme on vient de le voir, c'est principalement dans l'article 241 de la Loi de l'impôt sur le revenu, et parfois dans d'autres lois, qu'on traite des cas où la diffusion des données fiscales à l'extérieur de Revenu Canada, Impôt est permise. Il est donc facile d'oublier deux autres articles de la Loi de l'impôt sur le revenu qui sont également pertinents. Selon l'article 230, le ministre du Revenu national doit transmettre au directeur général des élections un rapport énonçant le montant total des contributions versées à chaque parti politique fédéral enregistré et le montant total des contributions versées à chaque candidat à l'élection d'un ou plusieurs députés à la Chambre des communes. Ce rapport est un document public. Enfin, selon l'article 149.1, le ministère du Revenu national doit

communiquer les renseignements contenus dans la déclaration publique de renseignements produite par les organismes de charité enregistrés. Le ministre peut rendre publics d'autres renseignements concernant les organismes de charité, notamment leurs nom, adresse et numéro d'enregistrement.

L'aperçu que je viens de présenter donne une idée des objectifs poursuivis en plus de la protection des données fiscales. Ces objectifs comprennent l'administration efficace et efficiente de certains programmes sociaux et économiques, la sécurité nationale, la collecte efficiente de statistiques, la formulation et l'évaluation de la politique fiscale et enfin, pour ce qui a trait aux organismes de charité et aux contributions versées à un parti politique, l'information du public. Il est possible qu'il y en ait d'autres qui m'ont échappé.

Il existe deux autres lois qui entrent également dans le cadre juridique de la protection des données fiscales et qu'il ne faut pas oublier, ce sont la Loi sur l'accès à l'information et la Loi sur la protection des renseignements personnels. La Loi sur l'accès à l'information accorde à tout particulier le droit d'avoir accès aux documents d'un ministère ou organisme fédéral, tandis que la Loi sur la protection des renseignements personnels accorde à tout particulier le droit d'avoir accès aux renseignements personnels le concernant.

En ce qui concerne Revenu Canada, Impôt, la divulgation de renseignements en vertu de la Loi sur l'accès à l'information est déterminée par les dispositions relatives à la confidentialité contenues dans l'article 241 de la Loi de l'impôt sur le revenu. À moins que cet article ne permette la divulgation de renseignements personnels et financiers, la confidentialité des renseignements de cette nature contenus dans les fichiers fiscaux est garantie en tout temps. Voici, à titre d'illustration, quelques exemples de demandes de renseignements sur les déclarants présentées en vertu de la Loi sur l'accès à l'information qui ont été refusées.

- Une liste des sociétés qui ont reçu des crédits d'impôt à la recherche scientifique.
- Une liste des sociétés qui offrent à leurs employés un régime de participation différée aux bénéfices.
- Une liste des noms les plus courants.
- Une liste des personnes qui ont déclaré plus de 1000$ de revenu d'intérêts.
- Le nombre de compagnies d'assurance-vie et la valeur totale que représentent les nouvelles cotisations établies à leur égard.

Je ne suis pas un expert en la matière, mais j'ai l'impression que les lois sur l'accès à l'information et sur la protection des renseignements personnels ont pour effet de limiter la diffusion des données fiscales et de renforcer les dispositions de l'article 241 de la Loi de l'impôt sur le revenu se rapportant à la confidentialité. Autrement dit, la Loi sur l'accès à l'information ne retire pas la protection des données fiscales garantie aux termes de la Loi de l'impôt sur le revenu, tandis que la Loi sur la protection des renseignements personnels renforce, à mon avis, les dispositions de la Loi de l'impôt sur le revenu qui permettent à une personne ou à son agent d'avoir accès à des renseignements personnels la concernant que le ministère possède.

Pour conclure ce survol du cadre juridique qui entoure la protection des données fiscales, je dois mentionner deux autres protections. Premièrement, les activités de la plupart des ministères et organismes fédéraux (sinon tous) qui peuvent obtenir des données fiscales à des fins spéciales sont régies par des lois contenant des dispositions dont l'esprit est plus ou moins semblable à celui de l'article 241 de la Loi de l'impôt sur le revenu. Par exemple, la Loi sur la statistique contient une disposition concernant le respect de la confidentialité des renseignements recueillis et prévoit des sanctions si cette confidentialité n'est pas respectée. Il en est de même de la Loi sur le vérificateur

général. Deuxièmement, tous les fonctionnaires doivent prêter un serment de discrétion auquel ils sont liés. Ces dispositions légales, qui s'ajoutent à celles de la Loi de l'impôt sur le revenu concernant la confidentialité, le principe de l'usage compatible et les sanctions imposées dans chaque cas ainsi qu'aux exceptions prévues par la Loi sur l'accès à l'information constituent, à mon avis, des mesures de protection supplémentaires qu'il ne faut pas négliger.

3. QUESTIONS D'ORDRE ORGANISATIONNEL ET POLITIQUE EN MATIÈRE DE SÉCURITÉ

Sécurité

Comme je l'ai indiqué dans mon introduction, le régime fiscal canadien fondé sur l'autocotisation fonctionne dans la mesure où les déclarants sont très nombreux à fournir de leur plein gré les renseignements qu'on leur demande. Pour cela, il faut que les déclarants aient l'assurance que seules les personnes autorisées ont accès à ces renseignements délicats et que ces personnes les utilisent uniquement aux fins permises par la loi.

Comme je l'ai également souligné, s'il est vrai que la loi contient des dispositions visant la protection des données fiscales, il y a des limites au pouvoir des lois. La politique et les pratiques organisationnelles adoptées à cet égard sont au moins aussi importantes.

À Revenu Canada, Impôt, nous avons toujours pris très au sérieux la responsabilité d'assurer que seules les personnes autorisées ont accès aux données fiscales. Nous nous attendons que les personnes autorisées dans les ministères et organismes fédéraux qui reçoivent ces données prennent aussi cette responsabilité au sérieux. C'est une question sur laquelle nous insistons.

La politique du Ministère en matière de sécurité a pour objectif de protéger l'information, les employés et les biens et de pouvoir donner aux déclarants l'assurance que les renseignements qu'ils fournissent seront gardés secrets et serviront uniquement aux fins permises par la loi. C'est pourquoi le Ministère possède depuis longtemps une politique officielle sur la sécurité ainsi que des mécanismes favorisant son application et qu'il offre un programme de sensibilisation à la sécurité destinée à ses employés.

Au mois de juin 1986, le Conseil du Trésor a rendu publique la politique du gouvernement du Canada concernant la sécurité. Cette politique a pour objectif de prescrire un système de sécurité pour le gouvernement du Canada qui permettra de protéger efficacement les renseignements classifiés et autres biens d'une importance capitale pour le pays de toute divulgation, destruction, modification, interruption ou de tout retrait non autorisés. Elle a aussi pour but de prescrire des moyens de sauvegarder d'autres renseignements de nature délicate et d'autres biens de grande valeur comme les données fiscales.

La politique concernant la sécurité exige le respect des normes de sécurité **matérielle** établies par la Gendarmerie royale du Canada (GRC), des normes de sécurité relatives au **personnel** établies par le Secrétariat du Conseil du Trésor et des normes de sécurité relatives aux **technologies de l'information** établies par la GRC et le Centre de la sécurité des télécommunications. En bref, ces normes définissent les exigences obligatoires minimales auxquelles il faut satisfaire en vue de protéger l'information et les biens de l'administration publique fédérale et sur lesquelles il faut se fonder pour effectuer les enquêtes de sécurité et la vérification de la fiabilité des employés.

En novembre 1986, Revenu Canada, Impôt a commencé à suivre les recommandations relatives au personnel que contient la nouvelle politique en matière de sécurité. Il s'agit

de s'assurer que tous les employés du Ministère qui ont accès à des renseignements de nature délicate dans le cadre de leur travail répondent aux normes de fiabilité, de fidélité et de loyauté.

Le Ministère est en train d'achever l'examen systématique des renseignements en sa possession, examen qui vise à déterminer le niveau de protection nécessaire dans chaque cas, et d'étudier les normes de sécurité matérielle et de sécurité de l'information en vue de commencer à les appliquer au début de l'année 1988. Entre temps, les mesures de sécurité existantes sont appliquées avec soin et les améliorations possibles, apportées là où elles s'imposent. Ainsi, de concert avec Statistique Canada, nous avons pris des mesures dernièrement pour augmenter la protection des données fiscales en transit entre les deux organismes à Ottawa.

Au fur et à mesure que les ministères et organismes autorisés à recevoir des données fiscales commenceront à appliquer la nouvelle politique concernant la sécurité du gouvernement et les normes dont elle est assortie, la protection dont bénéficient les données fiscales en général devrait augmenter, ce dont nous nous réjouissons à Revenu Canada. Par ailleurs, le fait que ces ministères et organismes ont adopté les mêmes normes de sécurité devrait simplifier la tâche qui nous incombe à tous de protéger les données fiscales.

De même qu'il y a une limite au pouvoir des lois d'assurer la protection des données fiscales, la politique et les lignes de conduite en matière de sécurité ont leurs limites aussi. Il existe d'autres lignes directrices et pratiques à Revenu Canada, Impôt qui visent à assurer la confidentialité des renseignements et leur utilisation aux fin permises par la loi seulement. Ces lignes directrices et pratiques exigent de plus en plus la participation des autres ministères et organismes.

4. CONTRÔLE DE LA COMMUNICATION DES DONNÉES FISCALES

Revenu Canada, Impôt a toujours eu recours à plusieurs moyens pour contrôler la communication des données fiscales.

C'est le ministre des Finances qui a le pouvoir de proposer que des modifications soient apportées à la Loi de l'impôt sur le revenu. Les obligations de Revenu Canada, Impôt en tant qu'autorité compétente en matière de fiscalité font cependant que ce ministère s'intéresse beaucoup aux conséquences administratives des modifications proposées. Les fonctionnaires de Revenu Canada, Impôt participent donc habituellement à l'élaboration des modifications proposées lorsque celles-ci touchent la communication et la protection des données fiscales, qu'elles concernent la Loi de l'impôt sur le revenu ou toute autre loi.

J'espère ne pas avoir donné l'impression que nous prenons ces propositions à la légère ni porter les personnes au service des autres ministères à croire qu'il est facile d'incorporer les modifications proposées dans les lois. En fait, c'est plutôt le contraire. Les propositions visant à modifier les lois qui ont une incidence sur la protection des données fiscales sont examinées minutieusement par les fonctionnaires de Revenu Canada, Impôt, comme elles le sont aussi par les membres du Parlement. À Revenu Canada, Impôt, nous tenons à déterminer, dans leur ensemble, les coûts et les avantages de ces propositions, et ce pour l'État, pour la collectivité et pour le Ministère. Nous tenons également à mesurer les conséquences que ces propositions auront sur les priorités et exigences opérationnelles du Ministère. Et nous informons le ministre de nos conclusions.

Les demandes émanant de ministères ou organismes qui ont le droit de recevoir des données fiscales, comme toutes les autres demandes d'ailleurs, font depuis longtemps l'objet d'un examen attentif. Premièrement, on s'assure que l'organisme demandeur a bien droit aux données voulues et que l'usage qu'il veut en faire est celui qui lui est permis. Deuxièmement, en tenant toujours compte des obligations et exigences de Revenu Canada

Impôt en tant qu'autorité compétente en matière de fiscalité, on détermine la disponibilité des données, la possibilité de les produire ainsi que le coût de production éventuel et enfin le délai raisonnable dans lequel elles peuvent être fournies.

Les ministères et organismes publics se sont déjà vu refuser des données fiscales auxquelles ils n'avaient pas droit en vertu de la loi. Ils se sont vu refuser des données qu'ils voulaient utiliser à des fins non permises par la loi. Ils ont retiré ou modifié leur demande parce qu'on ne pouvait leur fournir les données telles que demandées, au moment requis ou à un coût acceptable. Dans d'autres cas enfin, ils ont accepté des données agrégées au lieu de données fiscales confidentielles, lorsque celles-ci convenaient à leurs besoins.

Malgré les économies ou les gains d'efficacité qui auraient pu être réalisés, Revenu Canada, Impôt n'a pas permis que ses bases de données informatisées soient reliées à des terminaux ou à des ordinateurs situés dans les ministères qui avaient le droit de recevoir ou d'utiliser des données fiscales. Nous n'avons pas l'intention de le faire à l'avenir non plus.

Nous avons également toujours pris soin de ne pas diffuser plus de données qu'il n'en fallait pour un usage permis en particulier.

Récemment, Revenu Canada, Impôt a revu et reformulé sa politique interne en ce qui a trait à la diffusion des données fiscales à des fins autres que l'administration du régime fiscal. Conformément à cette politique, nous sommes en train de conclure des ententes écrites avec les ministères et organismes publics qui ont le droit de recevoir de telles données. Nous avons pour but de mieux gérer la diffusion des données qui leur sont destinées et de les encourager à augmenter la protection accordée à ces données. Nous sommes convaincus que ces ententes seront d'une grande utilité pour les deux parties, et pas seulement pour Revenu Canada, Impôt.

En général, ces ententes déterminent les données qui seront fournies ainsi que les modalités relatives à leur diffusion et à leur protection. En outre, elles établissent clairement les voies de communication entre Revenu Canada, Impôt et les ministères et organismes visés pour toutes les questions relatives à la diffusion des données, à la mise à jour des ententes et à la sécurité.

Conformément aussi à la nouvelle politique de Revenu Canada, Impôt, la Direction des systèmes est en train de vérifier les documents produits par ordinateur pour s'assurer que les renseignements diffusés sont bien ceux dont la diffusion a été autorisée. Le Ministère tient également un registre de toutes les données envoyées à chaque ministère ou organisme.

5. UTILISATIONS STATISTIQUES

Jusqu'à maintenant, j'ai parlé de la protection des données fiscales du point de vue de Revenu Canada, Impôt comme source de données fournies à certains ministères et organismes fédéraux à des fins autorisées. Le Ministère produit également des statistiques fiscales ou données agrégées, dont certaines sont publiées, pour de nombreux clients tant à Revenu Canada, Impôt qu'ailleurs. Il s'efforce dans ce contexte aussi d'assurer la protection des données fiscales et d'amener ses clients à assumer le coût marginal que cette production entraîne et à accepter la responsabilité de l'usage qu'ils font des données en question.

Chaque fois que la production de statistiques pour un client ou la prestation de services statistiques exige au moins dix journées-personnes, nous rédigeons une description détaillée des paramètres du projet que nous faisons approuver par le client. En général, cette description précise les statistiques ou services statistiques devant être fournis, le

but et les objectifs poursuivis ainsi que les méthodes, les délais, les étapes-clés et les ressources nécessaires. Lorsqu'il signe cette description, le client accepte la responsabilité à la fois du produit et de l'interprétation et de l'utilisation qu'il fera des statistiques par la suite.

Lorsque des statistiques sont produites pour des clients ou autres utilisateurs qui n'ont pas le droit de recevoir des données fiscales révélatrices de l'identité des personnes ou entreprises auxquelles elles se rapportent, on applique une méthode statistique standard qui permet d'éviter la divulgation éventuelle par différence de renseignements confidentiels. En outre, périodiquement, nos statisticiens passent au peigne fin notre principale publication annuelle, **Statistique fiscale,** pour déterminer si, en dépit des précautions prises, il y aurait des changements à y apporter. Dans un domaine qui évolue autant que la politique fiscale et son application, il est à mon avis plus que souhaitable de prendre ce genre de précautions supplémentaires.

En terminant, j'aimerais aborder deux questions relatives à l'utilisation statistique des données fiscales. Ce sont des questions qui concernent principalement les utilisateurs qui ne sont pas à Revenu Canada, Impôt, mais elles nous touchent aussi dans une certaine mesure.

Premièrement, il y a la question de la définition de termes comme "revenu" et "conjoint". Ces termes ont un sens particulier dans le domaine du droit fiscal et de son administration. Ce sens peut évoluer avec le temps et convenir ou non au domaine dans lequel divers organismes compilent ou utilisent des statistiques fondées sur les données fiscales.

Deuxièmement, il y a la question de la constance avec laquelle on retrouve certains éléments de données dans les dossiers fiscaux et les conséquences que cela peut avoir pour la production de séries chronologiques homogènes. À cause des modifications apportées au lois fiscales et des changements qui en découlent dans les formules de déclarations d'impôt et dans les annexes, les renseignements fournis par les déclarants peuvent varier d'une année à une autre et même à l'intérieur d'une même année, et ils varient parfois beaucoup. Par exemple, le montant maximal des déductions pour dépenses relatives à un emploi est de 500$ depuis plusieurs années. Pour l'année d'imposition 1982, la déduction était de 3 % du revenu d'emploi jusqu'à concurrence du maximum. Depuis l'année d'imposition 1983, elle est de 20 %. Une première analyse de ces données pourrait révéler des écarts attribuables aux changements apportés à la politique fiscale mais non pas aux conditions socio-économiques. Le fait qu'on se serve de l'adresse du déclarant pour le codage des identificateurs géographiques est un autre exemple. Ce qui intéresse Revenu Canada, Impôt, c'est d'avoir une adresse permettant de communiquer avec le déclarant que ce soit pour lui envoyer un remboursement par la poste ou pour une autre raison d'ordre fiscal. On ne tient pas nécessairement à savoir si l'adresse est celle de la résidence principale, du lieu de travail ou d'un voisin ou s'il s'agit d'une case postale. Toutefois, cette question peut être importante pour les personnes qui étudient les séries chronologiques en vue de faire une analyse comparative des tendances observées, à un niveau infraprovincial, dans diverses régions géographiques.

Autrement dit, il y a des limites considérables propres à l'utilisation des données fiscales à des fins statistiques. D'après ce que j'ai pu voir, la plupart des utilisateurs de ces données comprennent et acceptent ces limites.

6. CONCLUSION

Dans l'administration publique et dans l'application des politiques d'intérêt général, il faut habituellement essayer de trouver un équilibre entre des buts qui s'opposent. Comme j'ai essayé de le montrer, l'utilisation des données fiscales à des fins non fiscales, statistiques ou autres, illustre bien ce propos.

Revenu Canada, Impôt est un ministère qui a des obligations en tant qu'autorité compétente en matière de fiscalité que nous nous efforçons de remplir avec efficience et efficacité. Nous tentons également de le faire de manière à respecter les droits des contribuables. Et ceux-ci ont notamment le droit de s'attendre que les renseignements financiers et personnels qu'ils fournissent soient gardés secrets et servent uniquement aux fins permises par la loi.

Comme je l'ai montré, ces fins sont nombreuses et variées. Les fonctionnaires des autres ministères et organismes publics qui se servent des données fiscales sont assujettis eux aussi aux dispositions de l'article 241 de la Loi de l'impôt sur le revenu qui ont trait à la confidentialité et à l'usage compatible. Nous sommes tous tenus de suivre la politique du gouvernement en matière de sécurité.

Le régime fiscal n'a pas pour raison d'être de fournir des données à des fins autres que celle de son administration. L'utilisation des données fiscales à des fins statistiques exige donc des compromis.

Nous avons réussi collectivement, l'expérience le montre je crois, à maintenir l'équilibre entre les buts souvent contradictoires que nous poursuivons, et ce, avec responsabilité et crédibilité. Pour continuer dans cette voie, il faudra faire preuve de compréhension, d'ardeur et d'esprit de collaboration. Je suis persuadé que nous y arriverons.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

UTILISATION STATISTIQUE DES DOSSIERS ADMINISTRATIFS AUX ÉTATS-UNIS: OÙ EN SOMMES-NOUS ET OÙ ALLONS-NOUS?

THOMAS B. JABINE et FRITZ SCHEUREN[1]

RÉSUMÉ

Compte tenu des ressources et de la technologie disponibles, le système statistique fédéral américain peut, si la chose est jugée souhaitable, donner une place beaucoup plus grande aux dossiers administratifs. Les facteurs importants qui déterminent une stratégie pour l'élaboration ultérieure de l'utilisation statistique des dossiers administratifs sont l'opinion des organismes statistiques, des services qui ont la responsabilité des dossiers administratifs et du public; la nécessité de faire face à des changements fréquents du champ d'observation, du contenu et de la structure des sources de dossiers administratifs; les dispositions légales et les énoncés de politique touchant la confidentialité, la divulgation et l'accès aux dossiers administratifs; et la nécessité de coordonner l'activité des organismes statistiques dans un système décentralisé tel que le nôtre. On examine ici les tendances et les changements récents survenus aux États-Unis touchant à chacun des facteurs susmentionnés ainsi que leurs effets possibles sur la recherche relative à l'utilisation des dossiers administratifs.

INTRODUCTION

La statistique, dans le sens étynologique du terme, signifie la présentation de données relatives à l'État. Les dossiers administratifs étaient la source principale de ces données. Depuis qu'on procède régulièrement à des recensements, les sociétés modernes ont commencé à avoir d'autres bonnes sources de données. Pourtant, il s'est toujours trouvé des personnes pour dire: "Nous avons tous ces dossiers administratifs (sur un sujet ou un autre). On peut sûrement en tirer plus".

Dans l'ensemble, ces gens ont raison mais, comme Tom Peters le dit si bien: "Les choses évidentes ne vont pas toujours de soi". (Peters et Waterman, 1982) -- et certainement pas aux yeux des responsables de ces dossiers, qui fonctionnent habituellement avec un budget serré et qui utilisent les dossiers à des fins différentes. Le statisticien apprend vite, s'il ne le sait pas déjà, que faire ce qui est évident, **en admettant qu'il réussisse à convaincre suffisamment de personnes que ce qu'il fait va de soi**, s'avère n'être jamais facile.

[1] Thomas B. Jabine, expert-conseil en statistique, 3231 rue Worthington, N.W., Washington, D.C., 20015, États-Unis; Fritz Scheuren, directeur, Statistics of Income Division, Internal Revenue Service TR:S, 1111 Constitution Avenue, N.W., Washington, D.C., 20024, États-Unis.

Une fois que vous décidez de faire une conférence comme celle-ci, qui réunit des gens **qui font ce qui est evident mais pas facile**, vous vous sentez obligés (du moins nous nous sentons obligés de le faire) de définir le domaine. Cela dit, nous avons du mal à dire au juste **ce que cherchent réellement** ceux d'entre nous qu'intéresse l'utilisation statistique des dossiers administratifs. Comme Gordon Brackstone (1987a, 1987b) vient de le dire, les sources de données dont nous nous servons couvrent à toutes fins pratiques tous les secteurs de la vie moderne. Les techniques que nous utilisons couvrent aussi une bonne partie des statistiques modernes, même s'il y a certaines techniques de base que nous avons tendance à utiliser plus souvent, comme l'appariement d'enregistrements (Internal Revenue Service, 1985; Howe et Spasoff, 1986) et l'estimation synthétique (voir par exemple Gonzalez et Hoza, 1978; Fay et Herriot, 1979; Hidiroglou et coll., 1984).

Il serait peut-être utile d'examiner qui travaille dans ce domaine. Martin Wilk (1985), il y a déjà quelques années, quand il était Statisticien en chef du Canada, a écrit un article sur les différences entre les statisticiens "cols bleus" et les statisticiens "cols blancs". Plusieurs d'entre vous ici se rappellent certainement ses définitions. Les statisticiens "cols bleus" sont essentiellement ceux qui compilent les données, tandis que les statisticiens "cols blancs" sont essentiellement des théoriciens.

Le domaine de l'utilisation statistique des dossiers administratifs a nécessairement et à juste titre toujours été le fief des statisticiens de type "col bleu" travaillant pour la plupart dans les administrations. Cela dit, nous ne voulons pas insister outre mesure sur ce point, mais depuis un certain temps, on cherche comment raffiner les techniques plutôt qu' inventer de nouvelles sous-disciplines de la science statistique. On trouvera un bon exemple de ce que nous voulons dire dans les travaux d'Howard Newcombe (Newcombe et coll., 1959; Newcombe, 1967) à qui revient le mérite d'avoir élaboré avec ses collègues la technique de l'appariement d'enregistrements bien avant qu'Yvan Fellegi et Alan Sunter (1969) n'appliquent la méthode des tests d'hypothèses Neyman-Pearson pour résoudre ce problème.

Vous conviendrez peut-être avec nous que le domaine de l'utilisation statistique des dossiers administratifs est dominé en grande partie par les statisticiens "cols bleus". Cela ne veut pas dire que les statisticiens de type "col blanc" n'ont pas de rôle à jouer dans ce domaine. En fait, comme on pouvait s'y attendre, on constate qu'il y a à cette conférence plus de communications de "cols blancs" qu'il y en a habituellement.

Le plus difficile dans tout cela, bien entendu, c'est qu'on n'a pas seulement des statisticiens "cols blancs" et des statisticiens "cols bleus"; il y a aussi d'autres professions. Comme ce que nous faisons a une réelle importance pour nos sociétés, il y a des dirigeants, des éthiciens et d'autres personnes encore qui s'inquiètent de la possibilité qu'on utilise mal ce que nous faisons (voir par exemple Flaherty, 1979; Cox et Burch, 1985; Gastwirth, 1986). Nous classerions ces gens dans la catégorie des "travailleurs de bureau" si ce n'était que vous pourriez estimer que nous traitons avec trop de désinvolture une partie très fondamentale de ce domaine.

Notre article porte sur ce qui se fait à l'heure actuelle aux États-Unis dans le domaine de l'utilisation statistique des dossiers administratifs. Ce sujet est de toute évidence désespérément vaste. En outre, notre analyse des travaux effectués au Canada dans ce domaine nous amène à penser que vous avez pris pas mal d'avance sur nous dans le secteur aussi bien des cols blancs que des cols bleus de ce domaine (et peut-être même aussi dans le secteur du travail de bureau), même si vous ne faites pas face aux mêmes obstacles que nous. Malgré tout, vous pourriez peut-être trouver intéressant de passer en revue, mais cette fois-ci du point de vue américain, une partie des travaux que Gordon Brackstone vient juste de passer en revue du point de vue canadien.

Nos propos font suite à l'analyse détaillée de l'utilisation statistique des dossiers administratifs aux États-Unis que nous avons entreprise il y a quatre ans. A l'époque, nous avions proposé six objectifs à atteindre dans ce domaine aux cours des prochaines années

ainsi que la stratégie pour y arriver (Internal Revenue Service, 1984; Jabine et Scheuren, 1985). John Leyes, de Statistique Canada, a présidé les séances en panel de l'American Statistical Association, où nous avons présenté un document sur ce sujet. Les autres participants représentaient divers organismes statistiques américains (Butz, 1985a; Carroll, 1985; Norwood, 1985; Waite, 1985).

Ce symposium sur l'utilisation statistique des données administratives est une excellente occasion pour nous de nous demander dans quelle mesure nous avons progressé vers ces six objectifs. C'est ce que nous faisons dans la première partie de notre exposé. Sans vouloir annoncer à l'avance les résultats de notre analyse, il faut admettre que nous avons moins progressé par rapport à ce que nous avions espéré il y a quatre ans. Pourquoi cela? Les stratégies proposées sont-elles inadéquates? Les obstacles à une utilisation accrue des dossiers administratifs sont-ils plus importants que ce que nous avions pensé?

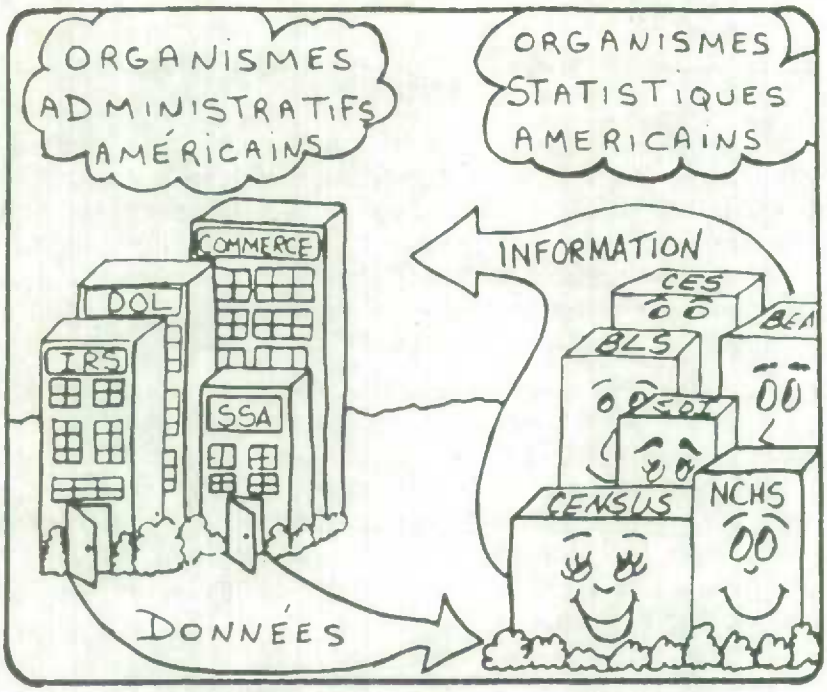
En cherchant à répondre à ces questions, nous avons identifié un certain nombre d'éléments qui, à notre avis, expliquent en grande partie dans quelle mesure on peut exploiter le plein potentiel des données administratives à des fins statistiques. Ces éléments sont examinés dans la deuxième partie du document, où nous accordons une importance particulière à ce qui s'est produit pendant les quatre années écoulées depuis que les six objectifs ont été proposés. Cette partie est suivie d'une troisième et dernière partie dans laquelle nous faisons de nouvelles prévisions sur la façon dont nous allons (ou pourrions) procéder dans l'avenir.

1. DANS QUELLE MESURE NOUS AVONS PROGRESSÉ VERS LES SIX OBJECTIFS

Avant de parler des objectifs précis que nous avons ambitieusement proposés, nous allons parler brièvement des différences et des similitudes existant entre la façon de s'organiser au Canada et la façon de s'organiser aux États-Unis pour produire des statistiques gouvernementales. La structure de base est la même dans les deux pays (voir figure 1). Il y a des organismes statistiques et des ministères administratifs. Bien que les organismes statistiques aient beaucoup de sources de données, incluant les données qu'ils recueillent directement, une de leurs principales sources est constituée par les données provenant des organismes administratifs. La figure 1 montre quelques organismes administratifs, chacun ayant son équivalent au Canada.

C'est du côté des organismes statistiques qu'il y a des différences. Aux États-Unis, il existe plusieurs organismes statistiques distincts (NCHS, Census, BEA, etc.) qui ici seraient centralisés à Statistique Canada. Mais le système canadien n'est pas lui non plus parfaitement centralisé, parce qu'il y a aussi en général des services statistiques dans les ministères administratifs au Canada. Un de ceux que nous connaissons très bien est l'équivalent canadien de la Statistics of Income (SOI) Division de l'IRS (la boîte avec des gros sourcis dans la figure), soit la Division des services statistiques de Revenu Canada Impôt. La Statistics of Income Division de l'IRS et la Division des services statistiques de Revenu Canada ont aussi à peu près le même mandat (comme on peut le constater en partie en comparant deux des communications présentées ici sur les statistiques tirées des déclarations d'impôt sur le revenu des sociétés, l'une par Fred Hostetter, Chris McCann et Brigitte Zirger, de Revenu Canada (1987), et l'autre par Hinkins, Jones et Scheuren (1987), de l'Internal Revenue Service. Voir aussi la communication présentée à cette conférence par John Czajka (1987).

Figure 1



NOTES: **Organismes administratifs**

DOL = Department of Labor;
SSA = Social Security Administration;
IRS = Internal Revenue Service.

Organismes statistiques

CES = Center for Education Statistics;
BEA = Bureau of Labor Statistics;
SOI = Statistics of Income Division, IRS;
NCHS = National Center for Health Statistics.

Comme au Canada, les dossiers administratifs sont un élément important des programmes statistiques fédéraux aux États-Unis. Les trois grands secteurs d'utilisation sont les suivants (voir figure 2):

— **Statistiques des programmes**

C'est un secteur bien structuré, et nous n'avions pas d'objectif bien précis à commenter explicitement ici, même si nous regrettons de ne pas parler plus de la nécessité d'améliorer les modèles de simulation de politiques (Revenu Canada Impôt, 1985). L'importante réforme fiscale (Tax Reform Act) de 1986 a poussé nos capacités à leur limite dans ce secteur et a fait ressortir (une fois de plus) les faiblesses auxquelles il faut remédier (Bristol, 1985). La communication présentée par Wolfson et coll. (1987) à ce symposium est comprise dans ce grand secteur.

— **Systèmes à usage général**

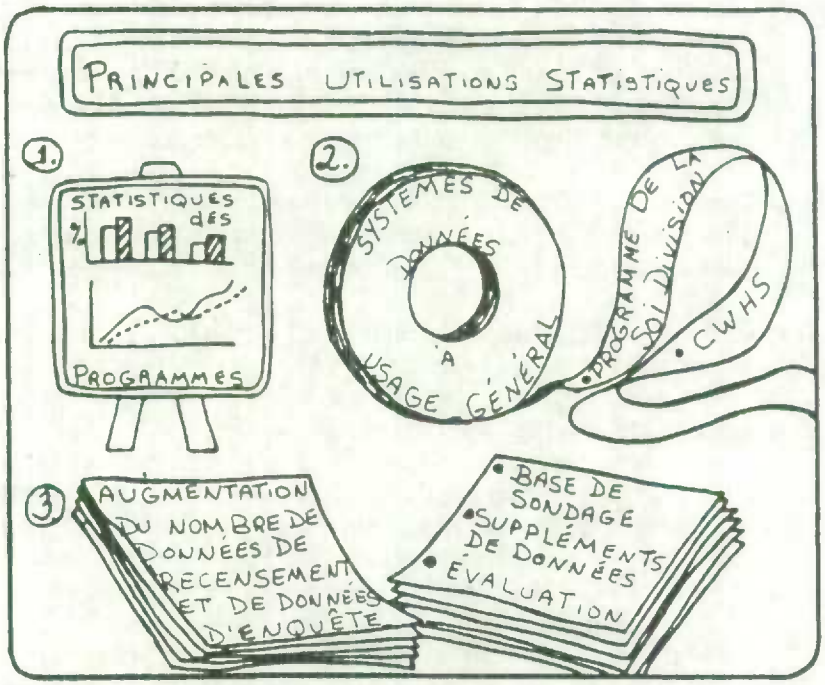
Par système à usage général, nous faisons allusion aux données statistiques fondées sur des dossiers administratifs servant à un usage général (et non pas construits spécialement pour des programmes précis). Entre autres exemples, notons le Continuous Work History Sample et une bonne partie de ce qui est fait avec les

dossiers fiscaux dans le cadre du Statistics of Income Program. Comme on le verra plus loin, nous avions deux objectifs dans ce secteur (Wilson, 1983; Buckler et Smith, 1980; Kilss, Scheuren et Buckler, 1980).

— **Amélioration des méthodes de collecte des données statistiques**

Cette catégorie comprend l'utilisation de dossiers administratifs comme bases de sondage pour obtenir plus de données d'enquête, etc. La plupart de nos objectifs et en fait la plupart des communications présentées au présent symposium concernent ce type d'utilisation.

Figure 2



Les six objectifs

Quand nous avons formulé les six objectifs à atteindre pour faire un plus grand usage des dossiers administratifs à des fins statistiques (voir figure 3), nous estimions (et nous estimons encore) que l'utilisation des dossiers administratifs dans les programmes statistiques allait continuer d'augmenter. En effet, les coûts de collecte directe des données augmentent toujours et il y a consensus général sur la nécessité de réduire le fardeau de réponse global imposé au grand public, à qui on demande de produire des renseignements à des fins tant administratives que statistiques.

Aussi estimions-nous que les organismes statistiques américains et le système statistique dans son ensemble devaient adopter une stratégie active cohérente pour développer de nouvelles utilisations statistiques des dossiers administratifs. Nous avons vu la nécessité d'une planification qui se ferait à l'échelle globale du système statistique et qui permettrait la meilleure utilisation possible, des systèmes de dossiers administratifs et d'avoir un certain contrôle sur les caractéristiques qui en déterminent l'utilité pour des applications statistiques. Le but des six objectifs était de diriger l'attention sur certaines applications qui, selon nous, étaient susceptibles de présenter le plus d'avantages, aussi bien pour les organismes statistiques que pour les utilisateurs de données en général.

Figure 3

Objectifs à atteindre concernant l'utilisation des dossiers administratifs: les 10 prochaines années (1984-1994)

Objectif 1: Explorer toutes les possibilités d'utilisation des principaux systèmes de dossiers administratifs dans la réalisation et l'évaluation des recensements décennaux de la population et la production des estimations courantes de la population.

Objectif 2: En prenant comme base le Continuous Work History Sample (CWHS), former un échantillon d'unités administratives et statistiques appariées (Linked Administrative Statistical Sample, LASS) en utilisant les dossiers administratifs provenant de plusieurs organismes pour construire un système de données longitudinales fondées sur l'individu qui servira à un large éventail d'utilisateurs.

Objectif 3: Donner la priorité aux applications statistiques des dossiers administratifs qui augmentent notre capacité de contrôler et d'analyser les effets à long terme de l'environnement sur la santé.

Objectif 4: Utiliser davantage les dossiers administratifs dans toutes les phases des enquêtes auprès des ménages.

Objectif 5: Créer et rendre opérationnel un "registre des entreprises" qui pourra servir de base de sondage aux recensements et aux enquêtes économiques ainsi qu'une source de codes géographiques et industriels pouvant être facilement utilisée par toutes les unités statistiques du gouvernement fédéral et des États qui auraient le droit de l'utiliser.

Objectif 6: Poursuivre les efforts destinés à rendre plus uniformes et plus compatibles les procédures utilisées pour définir et identifier les unités déclarantes à l'étape de la collecte des données aussi bien administratives que statistiques.

Source: Internal Revenue Service, 1984.

Dans cette partie, nous passons en revue les six objectifs un par un. Pour chaque objectif, nous décrivons d'abord l'état des activités se rapportant à l'objectif en question vers le milieu de 1984. Ensuite, nous examinons ce qui s'est produit par la suite et faisons une évaluation subjective de la mesure dans laquelle le chemin parcouru pour atteindre l'objectif a répondu à nos attentes.

1.1 Utilisation des dossiers administratifs dans la réalisation des recensements décennaux et la production des estimations courantes de la population

Objectif 1: Explorer toutes les possibilités d'utilisation des principaux systèmes de dossiers administratifs dans la réalisation et l'évaluation des recensements décennaux de la population et la production des estimations courantes de la population.

Situation en 1984

En 1982, Alvey et Scheuren ont proposé de considérer sérieusement la possibilité de recenser la population en se fondant sur des dossiers administratifs, principalement ceux de l'IRS et de la SSA, au lieu d'utiliser la méthode traditionnelle de dénombrement direct. Le Census Bureau a rejeté cette proposition pour le recensement de 1990. On a toutefois apporté à cette proposition une attention spéciale et le groupe de travail interne chargé de l'évaluer a recommandé que, dans le cadre des opérations du recensement de 1990, on procède à un test à grande échelle des procédures de recensement fondées sur des dossiers administratifs (Bureau of the Census, 1983). Childers et Hogan (1984) ont présenté les résultats d'une étude méthodologique dans laquelle on a apparié un échantillon de personnes de 18 à 64 ans qui ont produit une déclaration d'impôt et les unités du recensement de 1980. Un des principaux objectifs de l'étude était d'examiner la possibilité d'utiliser le Fichier principal des particuliers (Individual Master File) de l'IRS comme base de sondage pour procéder à un couplage avec les données du recensement dans le but d'estimer le taux de sous-dénombrement du recensement. Les auteurs ont conclu que cette méthode d'évaluation du taux de couverture était prometteuse et dit que "le Bureau entend poursuivre les recherches dans le domaine de l'utilisation d'autres sources que les ménages pour mesurer le taux de couverture." Toutefois, le Census Bureau a décidé de ne pas utiliser les dossiers de l'IRS dans les essais préliminaires qu'il devait effectuer en 1985 (voir cependant les utilisations ultérieures prévus dans la partie **Situation actuelle** ci-dessous).

Le rapport provisoire d'un groupe d'étude en panel sur la méthodologie des recensements décennaux (Panel on Decennial Census Methodology) mis sur pied par le Committee on National Statistics pour le compte du Census Bureau (voir National Research Council, 1984) ne contenait pas de recommandations définitives sur toutes les utilisations des dossiers administratifs dans le recensement de 1990. Les auteurs déconseillaient de procéder immédiatement à des essais sur le terrain de méthodes d'estimation, d'ajustement ou d'évaluation des données de recensement qui utiliseraient des listes multiples ou composites. Ils encourageaient toutefois l'utilisation des dossiers administratifs dans certaines études d'évaluation du contenu et peut-être aussi pour la collecte de certains types de données sur le logement.

Situation actuelle

Il se peut que les décisions finales au sujet de la question controversée de l'ajustement des chiffres de recensement influent beaucoup sur le rôle que les dossiers administratifs joueront dans le recensement de 1990. Le Département du Commerce a annoncé le 30 octobre qu'il n'avait pas l'intention d'ajuster les chiffres de population du recensement décennal de 1990 pour tenir compte d'un sous-dénombrement ou d'un surdénombrement prétendu de certains sous-groupes de la population. (**New York Times**, 1987). On a toutefois déposé au Congrès un projet de loi qui obligerait à procéder à des ajustements (**Wall Street Journal**, 1987). Advenant que le Census Bureau soit forcé d'apporter des ajustements officiels au niveau local, les données des principaux systèmes de dossiers administratifs pourraient être un élément important du processus d'ajustement. D'un autre côté, si l'idée d'ajuster les chiffres était rejetée, il semble probable qu'on aura moins recours aux dossiers administratifs pour améliorer le taux de couverture du recensement en 1990 qu'en 1980. Quelle que soit l'issue des discussions sur la question de l'**ajustement** du taux de couverture, on s'emploie actuellement à envisager de nouvelles applications des dossiers de l'IRS à l'**évaluation** du taux de couverture. La possibilité la plus probable est l'utilisation de listes de noms et d'adresses de contribuables pour vérifier l'exhaustivité de listes d'adresses créées pour une enquête indépendante d'évaluation du taux de couverture. Le Census Bureau prévoit inclure une composante utilisant des listes administratives dans l'enquête postcensitaire en vue de la répétition générale du recensement décennal qui aura lieu à St. Louis, Missouri, en 1988. Parmi les listes

administratives qu'on utilisera, il y aura les suivantes: IRS Individual Returns Extract avec les données correspondantes sur l'âge, le sexe et la race de la SSA, les fichiers sur les permis de conduire réguliers et temporaires des états, les fichiers de la Veterans Administration, et les fichiers de l'assurance-chômage. De toute façon, il est possible qu'une étude limitée, au moins au niveau de l'utilisation possible des dossiers administratifs dans le recensement, fasse partie du programme de recherche, d'évaluation et d'expérimentation de 1990 du Census Bureau (Research, Evaluation and Experimental Program, REX).

Même s'il n'est absolument pas question que le recensement de 1990 soit fait à partir de dossiers administratifs, certains événements récents pourraient améliorer les chances de cette approche pour l'an 2000. La Tax Reform Act de 1986 stipule qu'il faut indiquer les numéros de sécurité sociale (Social Security Numbers, SSN) de tous les contribuables et de toutes les personnes à charge de 5 ans et plus, de sorte qu'on pourra bientôt tirer des systèmes de dossiers de l'IRS et de la SSA les renseignements de base sur l'âge, le sexe et la race d'une proportion beaucoup plus grande de la population totale.

Ce fait nouveau a aussi des répercussions importantes sur les estimations courantes de la population. Le programme de partage général du revenu (General Revenue Sharing Program), qui avait besoin d'estimations courantes de la population et d'estimations courantes du revenu se rapportant à environ 39,000 unités gouvernementales locales, a été progressivement abandonné; néanmoins, les données de l'IRS et de la SSA continuent de jouer un rôle important dans le programme d'estimation de la population des petites régions et du revenu par habitant du Census Bureau. L'inclusion des numéros d'assurance sociale des personnes à charge dans les fichiers des données de l'IRS transmis au Census Bureau accroîtrait sensiblement la proportion de la population totale pour laquelle des estimations directes de la migration interne pourraient être faites grâce à un appariement des données dont dispose l'IRS pour différentes années.

Dans le cadre d'un project distinct, les dossiers de l'IRS et de la SSA ont été utilisés pour produire des données sur la population par région en fonction du code postal (ZIP CODE) pour 1985. Les estimations sont fondées sur l'appariement des données extraites du fichier principal des particuliers de l'IRS et des dossiers de la SSA qui indiquent l'âge, la race et le sexe d'un échantillon de 20% de toutes les personnes ayant un numéro d'assurance sociale. Les données relatives à la population d'origine hispanique ont été obtenues par codage de tous les noms de famille espagnols figurant dans le fichier de l'IRS (voir Passel et Word, 1980, 1987).

1.2 Amélioration du Continuous Work History Sample (CWHS)

Objectif 2. En prenant comme base le Continuous Work History Sample (CWHS), former un échantillon d'unités administratives et statistiques appariées (Linked Administrative Statistical Sample, LASS) en utilisant les dossiers administratifs provenant de plusieurs organismes pour construire un système de données longitudinales fondées sur l'individu qui servira à un large éventail d'utilisateurs.

Situation en 1984

Le CWHS, qui est tenu à jour par la Social Security Administration, est un système de fichiers statistiques contenant des données démographiques et longitudinales sur les gains pour des échantillons de personnes ayant un SSN. Jusqu'au milieu des années 1970, les fichiers de microdonnées du CWHS pouvaient être obtenus par les chercheurs à l'intérieur et à l'extérieur du gouvernement et étaient largement utilisés à des fins de recherche analytique sur le comportement du marché du travail, les tendances des gains sur la durée

de vie, la migration interne, les caractéristiques de la population active régionale et sur beaucoup d'autres sujets encore (U.S. Department of Health, Education and Welfare, 1978). Des plans ont été mis au point pour améliorer le système du CWHS en établissant des liens avec les données tirées d'autres systèmes portant sur la profession, le lieu de résidence actuel, les impôts sur le revenu, les prestations de sécurité sociale et la mortalité (Kilss, Scheuren et Buckler, 1980). Toutefois, la Tax Reform Act de 1986 a mis fin abruptement aux plans d'amélioration et à la diffusion à grande échelle des fichiers de microdonnées en imposant de nouvelles restrictions sur l'utilisation statistique des données sur les gains des employeurs et des particuliers recueillies au début par l'IRS et utilisées comme intrants dans le CWHS (Buckler et Smith, 1980; Duleep, 1986). En 1984, des négociations ont été engagées avec le Bureau of Economic Analysis (BEA) pour que les fichiers courants du CWHS puissent recommencer d'être communiqués de façon restreinte à cet organisme, qui était alors un gros utilisateur de ces données et jouait aussi un rôle important dans le traitement des fichiers de microdonnées en vue de leur distribution aux autres utilisateurs.

Situation actuelle

Les projets de communication des fichiers du CWHS au BEA ont été contrecarrés par le risque de divulgation et ne sont plus sérieusement considérés. Avec l'appui du National Cancer Institute (NCI), l'IRS et la SSA ont entrepris un projet pilote de création de fichiers qui relieraient les enregistrements du CWHS aux données de l'IRS sur les professions et les industries et aux renseignements sur la mortalité et les professions tirés des actes de décès retrouvés à la suite de recherches dans le National Death Index (Crabbe, Sailer et Kilss, 1983). Les données sur la cause du décès et la profession indiquée sur les actes de décès sont déjà liées à un sous-ensemble d'unités de l'échantillon des particuliers de la SOI Division conçu expressement pour chevaucher l'échantillon du CWHS. On prévoit ajouter au fichier plusieurs années de données sur les gains du CWHS pour la fin de 1988. La réalisation du projet dépend en grande partie de la signature d'ententes entre les organismes qui permettront de communiquer au NCI et aux autres utilisateurs une version accessible au grand public du fichier des microdonnées appariées. Malgré la complexité de ces accords, on a déjà fait beaucoup de progrès.

Cela fait dix ans qu'on a suspendu la diffusion générale des fichiers de microdonnées du CWHS. Même si l'étude pilote commanditée par le NCI a donné quelques résultats, quoique minimes, il est difficile d'être optimiste au sujet d'un retour éventuel aux anciennes politiques d'accès en vigueur avant l'adoption de la Tax Reform Act de 1986, encore moins d'une amélioration du système. La SSA, même si elle continue de conserver le CWHS pour usage interne (voir par exemple Kestembaum, 1985, 1986; Kestembaum et Diez, 1981), n'est pas en mesure de participer à l'amélioration et au développement de mécanismes d'accès requis pour répondre aux besoins d'un large éventail d'utilisateurs. On avait espéré que la communication des fichiers courants du CWHS au BEA pourrait permettre à cet organisme d'évaluer une fois de plus la qualité des codes industriels et géographiques développés dans le cadre du programme de déclarations des établissements de la SSA. Sans cette évaluation, il est impossible de déterminer si, comme certains le craignent, la qualité de ces éléments du système s'est détériorée depuis que les établissements ont commencé à déclarer les taux de rémunération annuelle en 1978. (Le projet de l'étude ERUMS, décrit dans la partie 1.6 un peu plus loin, fournit toutefois certains des résultats dont nous avons besoin pour faire cette évaluation.)

Il faudra peut-être se résigner à abandonner l'idée de revenir au "bon vieux temps" et commencer à songer à d'autres systèmes de données longitudinales à usages multiples fondés sur des dossiers administratifs. Une possibilité qui pourrait satisfaire à certaines des exigences, mais non à toutes, est l'échantillon des statistiques du revenu des particuliers de l'IRS (IRS Individual Statistics of Income (SOI) sample), qu'on s'emploie actuellement à remanier en profondeur. Le plan de remaniement renforcera la

composante longitudinale de ces échantillons annuels et, grâce à l'obligation d'indiquer dorénavant les SSN des personnes à charge, permettra d'obtenir des données sur les unités contribuables considérées non pas comme des particuliers mais comme des familles, et donc d'aller un pas plus loin que le CWHS, qui est strictement un échantillon d'individus (Jabine, 1987b; Hostetter, 1987). Tout optimisme au sujet de ces possibilités doit, malgré une tradition de vingt-cinq ans de fichiers accessibles au grand public, être tempéré par le fait qu'on sait que bon nombre des mêmes restrictions concernant la divulgation des renseignements influeront sur l'accès à ce système de données (Strudler, Oh et Scheuren, 1986). Il est évident qu'on ne peut pas s'attendre que l'IRS assure a lui seul toutes les ressources additionnelles nécessaires pour améliorer le système et pour le rendre plus utile et plus accessible à une population d'utilisateurs très variée. Enfin, une certaine forme d'accès général à des extraits **limités** du CWHS semble offrir une solution très prometteuse comme objectif réalisable à court terme.

1.3 Améliorer la capacité de déceler les effets à long terme de l'environnement sur la santé

Objectif 3: Donner la priorité aux applications statistiques des dossiers administratifs qui augmentent notre capacité de contrôler et d'analyser les effets à long terme de l'environnement sur la santé.

Situation en 1984

Le National Death Index (NDI) est devenu opérationnel en 1981; il contenait alors des renseignements tirés de tous les actes de décès enregistrés depuis 1979 et avait déjà servi dans plusieurs projets de recherche dans le secteur de la santé et de la médecine (Patterson et Bilgard, 1985). Dans notre analyse de cet objectif, nous avons fortement recommandé d'apparier les données sur la mortalité du NDI et les données d'une grande base de données statistiques comme le CWHS. Nous avons également attiré l'attention sur l'importance de continuer les travaux dans le domaine du codage des actes de décès en fonction de la profession indiquée (Crouse et coll., 1983) et de l'appariement des données des échantillons de la Current Population Survey (CPS) et des données sur la mortalité obtenues par le truchement du NDI (Rogot et al., 1983).

Situation actuelle

Comme on l'a vu dans la partie 1.2, l'utilité du couplage des données sur la mortalité tirées du NDI avec les données de l'ensemble du CWHS dépend des projets pilotes de la SSA, de l'IRS et du NCI dans ce domaine. En outre, le couplage avec l'éventuel échantillon longitudinal des statistiques du revenu (Statistics of Income, SOI), qui inclut des données sur les professions, semble possible.

D'autres projets dans ce domaine avancent bien. À la suite du projet conjoint du NCHS, du National Cancer Institute et du National Institute for Occupational Safety and Health concernant le codage des actes de décès selon la profession et la branche d'activité, 32 états et le District de Colombia codent maintenant ces éléments; de plus, le NCHS introduit les données codées pour 23 des régions dans sa base de données sur la mortalité.

On est en train de procéder à l'appariement des données de la CPS et des données du NDI, et les résultats de la première ronde, couvrant les décès enregistrés de 1979 à 1981, seront publiés bientôt. L'appariement des données se rapportant aux années 1982 à 1985 est déjà terminé et l'on prévoit faire d'autres appariements pour les années 1989 à 1991 (Johnson et coll., 1985; Rogot et coll., 1985).

Le NDI est une source de données inestimable pour l'étude des effets de l'environnement sur la mortalité. Dans les cohortes pour lesquelles on dispose de renseignements d'identification généralement justes, on peut déterminer avec precision et à peu de frais quelles sont les personnes décédées depuis 1979. Toutefois, s'il faut des données sur la cause du décès, la profession ou d'autres types, les résultats sont moins satisfaisants: la plupart des chercheurs doivent s'adresser aux bureaux des registres de l'état civil dans chaque État où les décès ont eu lieu. Pour des études nationales, ce processus peut être coûteux et fastidieux.

1.4 Usage accru des dossiers administratifs dans les enquêtes auprès des ménages

Objectif 4: Utiliser davantage les dossiers administratifs dans toutes les phases des enquêtes auprès des ménages.

Situation en 1984.

Des techniques d'appariement exact avaient déjà été utilisées dans beaucoup de cas pour apparier des données de dossiers administratifs et des données recueillies directement auprès de personnes ou de ménages dans des enquêtes. En plus de cette façon directe d'améliorer des données d'enquête, beaucoup d'autres façons d'utiliser les dossiers administratifs ont été ou pourraient être employées dans des enquêtes, par exemple pour la stratification des unités d'échantillonnage, la création de bases de sondage, l'imputation de valeurs à des données manquantes, l'estimation et l'évaluation. Un des obstacles rencontrés dans certains projets d'appariement de données d'enquête et de données administratives était le problème de la "réidentification", c'est-à-dire le problème posé par la possibilité que des services responsables de fichiers administratifs puissent, en appariant à partir de leurs propres dossiers, réidentifier des personnes pour lesquelles des enregistrements d'enquête et des dossiers administratifs appariés figureraient dans un fichier de microdonnées accessible au grand public.

Situation actuelle

Des appariements directs de données d'enquête et de données administratives se font encore. À noter en particulier la politique récente du NCHS de demander le SSN et d'autres identificateurs dans ses enquêtes courantes, comme dans la National Health Interview Survey, qu'il utilise ensuite dans des études de suivi nécessitant des appariements avec des dossiers administratifs comme les actes de décès et les fichiers Medicare (Scheuren, 1985). Cette politique a ajouté une nouvelle dimension aux données du NCHS -- soit leur utilisation dans des analyses épidémiologiques prospectives et non pas seulement pour décrire en coupe transversale l'état de santé de la population, l'utilisation des services de soins de santé et le comportement de la population en matière de santé.

On n'a toutefois pas encore trouvé de solution au problème de la réidentification, et le Census Bureau a maintenant comme politique explicite de ne communiquer aucun fichier de microdonnées accessible au grand public qui combine des données administratives et des données tirées d'enquêtes effectuées en vertu du Titre 13. En conséquence de cette politique, on ne pourra plus dorénavant copier d'anciennes publications de fichiers comme les fichiers produits à partir de l'Exact Match Study de 1973 ou de la Longitudinal Retirement History Survey. Il en résulte que toute personne ou organisme qui commande une enquête et veut apparier des données d'enquête et des données administratives et avoir accès aux dossiers résultants ne peut profiter ni du recensement décennal, ni de la Current Population Survey comme base de sondage.

On continue d'utiliser d'autres façons les dossiers administratifs dans les enquêtes-ménages, notamment à des fins d'évaluation. Par exemple, deux communications présentées à ce symposium (Moore et Marquis, 1987; Bowie et Kasprzyk, 1987) décrivent comment le Census Bureau utilise les dossiers administratifs dans la SIPP et dans d'autres enquêtes. Tippett (1987) a décrit comment les répartiteurs fiscaux et les sociétés d'utilité publique utilisent les dossiers administratifs pour évaluer les données recueillies dans l'American Housing Survey et les essais préliminaires du recensement de 1990. L'Energy Information Administration continue d'utiliser les dossiers des sociétés d'utilité publique et des fournisseurs en combustible pour compléter les données recueillies auprès des ménages dans le cadre ou Residential Energy Consumption Survey (Energy Information Administration, 1987). Il y a encore beaucoup d'autres façons d'utiliser les dossiers administratifs dans les enquêtes-ménages. Dans l'ensemble, l'augmentation du niveau d'activité dans ce domaine est encourageante.

1.5 Création d'un "registre des entreprises" partagé

Objectif 5: Créer et rendre opérationnel un "registre des entreprises" qui pourra servir de base de sondage aux recensements et aux enquêtes économiques ainsi qu'une source de codes géographiques et industriels pouvant être facilement utilisée par toutes les unités statistiques du gouvernement fédéral et des États qui auraient le droit de l'utiliser.

Situation en 1984

Au mois de novembre 1983, l'Administration a retiré son appui à un vaste projet de loi sur la confidentialité -- l'Enclave Bill -- qui contenait des dispositions prévoyant que la Standard Statistical Establishment List (SSEL) du Census Bureau pourrait être communiquée à des fins statistiques aux organismes des États et aux organismes fédéraux autorisés à l'obtenir. Cette mesure n'était que la dernière d'une longue série de tentatives infructueuses pour atteindre ce que beaucoup de gens estimaient, il y a déjà presque 50 ans, comme un objectif sensé: l'utilisation de listes d'entreprises partagées pour réduire les coûts et augmenter le degré de comparabilité des données d'enquêtes-établissements menées par les différents organismes de notre système statistique décentralisé. Nous estimions en 1984 qu'il était temps pour les différents groupes concernés de s'engager dans une recherche constructive de nouvelles solutions. En particulier, nous avons recommandé d'envisager d'autres solutions de rechange comme la construction d'une liste d'entreprises contenant des informations se rapportant uniquement aux unités auprès desquelles le Census Bureau ou tout autre organisme coopérateur (n'incluant **pas** l'IRS) a recueilli directement de l'information.

Situation actuelle.

Après que l'Administration eut retiré son appui au Enclave Bill, le Census Bureau a décidé de concentrer ses efforts sur des mesures législatives visant un objectif plus restreint: la diffusion de l'information contenue dans la SSEL aux organismes statistiques des États et fédéraux autorisés à l'obtenir. Toutefois, étant donné que la SSEL contient certains renseignements provenant des déclarations d'impôt, l'IRS s'est fortement opposé à ce qu'on modifie le Tax Code de manière à permettre la diffusion de l'information contenue dans la SSEL à d'autres organismes.

Au mois de mars 1986, l'Economic Policy Council a créé un groupe de travail chargé d'examiner la qualité des statistiques économiques (Working Group on the Quality of Economic Statistics). La qualité des listes d'entreprises utilisées dans les enquêtes

économiques a été l'un des cinq points auxquels le groupe de travail a donné la priorité. Dans le rapport qu'il a présenté en avril 1987 au Economic Council, le groupe de travail a recommandé d'adopter l'approche à deux volets suivante pour améliorer les listes d'entreprises (Economic Policy Council, 1987).

(1) Le Département du Commerce devrait présenter un projet de loi visant à permettre au Census Bureau de divulguer les renseignements d'identification et de classification des entreprises à un certain nombre d'organismes statistiques mentionnés explicitement. **Aucune donnée fiscale obtenue de l'IRS ne serait incluse dans ces divulgations.**

(2) En vertu du Paperwork Reduction Act, l'OMB devrait désigner le Bureau of Labor Statistics (BLS) et le National Agriculture Statistics Service (NASS) comme des "Organismes centraux de collecte" (Central Collection Agencies) pour certaines listes d'entreprises non agricoles et d'entreprises agricoles respectivement.

Le groupe de travail a recommandé que des ébauches du projet de loi et de directives administratives soient "complétées et coordonnées" pour le mois de juin 1987. Le BLS et le NASS ont tous les deux présenté une proposition préliminaire en vue d'assurer leur fonction d'organismes centraux de collecte, mais les désignations officielles n'ont pas encore été faites par l'OMB. Au mois de novembre 1987, le Census Bureau n'avait pas encore présenté son projet de loi.

Il est difficile de prédire l'issue de ces nouvelles démarches. Les résultats dépendront en grande partie de certains facteurs: la loi permettant au Census Bureau de partager la SSEL sera-t-elle adoptée?; le BLS et le NASS obtiendront-ils le financement supplémentaire dont ils ont besoin pour leurs listes d'entreprises?; et le BLS pourra-t-il conclure des ententes satisfaisantes avec les organismes de sécurité de l'emploi des États qui administrent le programme d'assurance-chômage (Unemployment Insurance Program) et qui sont la principale source des listes d'entreprises du BLS?. Quant au dernier point, il est encourageant de noter que le BLS a déjà obtenu d'environ 40 états la permission de partager des parties du fichier de la liste du programme de l'assurance-chômage avec le NASS afin de permettre à cet organisme de dresser des listes, particulièrement dans le secteur des services agricoles. Le vrai test consistera à déterminer si des ententes de partage des listes dans les deux sens peuvent être conclues entre le Census Bureau et le BLS ainsi qu'entre le Census Bureau et le NASS. On pourra savoir avec certitude dans quelle mesure ce nouveau système fonctionne au plus tôt en 1992 avec le recensement de l'agriculture et les recensements économiques.

1.6 Standardisation des unités déclarantes pour la collecte de données administratives et statistiques

Objectif 6: Poursuivre les efforts destinés à rendre plus uniformes et plus compatibles les procédures utilisées pour définir et identifier les unités déclarantes à l'étape de la collecte des données aussi bien administratives que statistiques.

Situation en 1984.

Cet objectif concerne principalement, quoique pas exclusivement, les entreprises ou les unités économiques déclarantes. Nous avons fait remarquer que les entreprises, en particulier celles qui emploient, doivent déclarer (un bon nombre) de données (par exemple sur l'emploi et les salaires) en fonction de toutes sortes d'exigences administratives qui se recoupent mais qui ne sont pas entièrement compatibles entre elles par rapport aux

identificateurs et à la définition des unités déclarantes de plus, tout cela est imposé par des organismes ne partageant qu'une partie seulement des renseignements les uns avec les autres. Sans une meilleure intégration de ces systèmes de déclaration administratifs, il sera difficile d'arriver à une uniformisation complète dans les programmes statistiques qui utilisent ces systèmes pour identifier et classer les unités déclarantes. (La communication présentée par Colledge, 1987b, à ce symposium couvre en partie les mêmes questions, mais dans le contexte canadien.)

En 1984, un projet d'étude d'appariement d'unités déclarantes d'établissements (Establishment Reporting Unit Match Study, ERUMS) a été mis sur pied sous l'égide de l'Administrative Records Subcommittee du Federal Committee on Statistical Methodology (Cartwirght et coll., 1983; Buckler, 1985). L'objectif de l'étude ERUMS était de faire une étude de concordance d'échantillons de dossiers du BLS et de dossiers de la SSA dans un État pour voir s'il y avait des différences de couverture et de contenu entre les systèmes des deux organismes et pour formuler des recommandations dans le but d'augmenter la comparabilité des deux systèmes.

Situation actuelle.

Le projet ERUMS a progressé lentement, mais un échantillon de dossiers appariés et de dossiers non appariés des deux systèmes peut maintenant être obtenu et fait actuellement l'objet d'une analyse détaillée. La lenteur des travaux réalisés dans le cadre de ce projet de recherche est attribuable en partie aux difficultés juridiques et administratives qu'il faut surmonter dans tout projet d'appariement de dossiers provenant d'organismes différents et en partie au fait que l'exécution du projet est confiée au personnel d'organismes dont les principales priorités sont ailleurs. Néanmoins, les participants estiment que l'expérience acquise et les résultats obtenus seront utiles, non seulement pour ce qu'ils montrent sur la relation entre les listes d'entreprises du BLS et de la SSA, mais aussi comme modèle pour d'autres études semblables. En particulier, les études de concordance des données de listes d'entreprises appartenant au BLS, au Census Bureau et au NASS pourraient fournir des renseignements précieux pouvant servir à guider la restructuration proposée des listes d'entreprises utilisées par les organismes statistiques fédéraux dans leurs enquêtes économiques (voir la partie 1.5 plus haut).

Les travaux d'intégration des systèmes de déclaration administratifs sous-jacents n'ont pas beaucoup progressé. Dans un sens, le fardeau imposé aux employeurs a même augmenté. En 1978, on a réduit le fardeau de réponse en demandant aux entreprises de déclarer les salaires gagnés par chacun de leurs employés aux fins du programme de sécurité sociale une fois par année et non plus à tous les trimestres. Mais, plus récemment, une loi fédérale a été adoptée qui oblige les employeurs de tous les États à déclarer les salaires gagnés par chacun de leurs employés une fois par trimestre en vertu du système d'assurance-chômage tandis qu'avant, dans certains États, les employeurs n'avaient à déclarer les salaires gagnés par chacun de leurs employés que dans le cas des employés qui avaient fait des demandes de prestations d'assurance-chômage.

L'inclusion des numéros d'identification des employeurs (Employer Identification Numbers, EIN) dans le fichier d'identification des employeurs du BLS a été un pas en avant. Cet ajout a placé le BLS dans une meilleure position pour remplir son rôle éventuel de fournisseur de listes d'entreprises désigné par l'OMB; par exemple, cela permettra au BLS d'apparier ses dossiers et des listes provenant d'autres sources, comme les listes Dun et Bradstreet, qui contiennent des EIN.

1.7 Résumé

La figure 4 résume ce que nous pensons des progrès accomplis (ou du manque de progrès) depuis 1984 en vue d'atteindre les six objectifs fixés à ce moment-là. Dès le

départ, il faut reconnaître la subjectivité de cet exercice. Les objectifs ont été définis dans les grandes lignes: les critères d'évaluation des réalisations demeurent implicites, n'ayant pas été définis de façon précise. D'autres observateurs pourraient facilement avoir un point de vue plus optimiste ou plus pessimiste que le nôtre.

Il devrait y avoir un point d'interrogation après le chiffre 2000 dans la figure 4 (objectif 1). C'est le commentaire le plus décourageant que nous puissions faire sur les progrès accomplis en vue d'un recensement à partir de dossiers administratifs (Administrative Record Census, ARC). Ce qui se fera ailleurs dans les prochaines années pourrait faire la différence, et plus particulièrement ce qui se fera au Canada. Le U.S. Census Bureau a toutefois le mérite de ne pas rejeter cette possibilité, mais nous estimons (et nos collègues du Census Bureau pourraient ne pas être d'accord) que le Bureau a concentré ses efforts dans d'autres domaines. Il a surtout été préoccupé par la question de l'ajustement des données sur la population du recensement décennal de 1990 pour tenir compte des cas possibles de sous-dénombrement ou de surdénombrement. La décision de ne pas ajuster les chiffres obtenus en 1990 exclut pratiquement toute possibilité d'un ARC pour l'an 2000 puisque, si l'on avait décidé d'ajuster les données, les dossiers administratifs auraient pu être la ressource clé pour obtenir les chiffres locaux.

Les chances à court terme d'atteindre l'objectif 2, c'est-à-dire d'augmenter le CWHS pour qu'il devienne un système accessible de données longitudinales à usages multiples, semblent très minces. Des succès limités semblent encore possibles: toutefois, nous avons suggéré à regret que l'attention soit maintenant tournée vers d'autres systèmes de données administratives qui n'ont pas toutes les caractéristiques intéressantes du CWHS mais offrent de meilleures possibilités de développement futur. À l'autre extrême du spectre, on retrouve l'objectif 4, qui était de faire un plus grand usage des dossiers administratifs dans les enquêtes-ménages. Nous voyons que des progrès importants ont été réalisés dans ce secteur, le seul nuage à l'horizon étant le problème de la réidentification, qui empêche de diffuser des fichiers de microdonnées contenant des données d'enquête et des données administratives appariées.

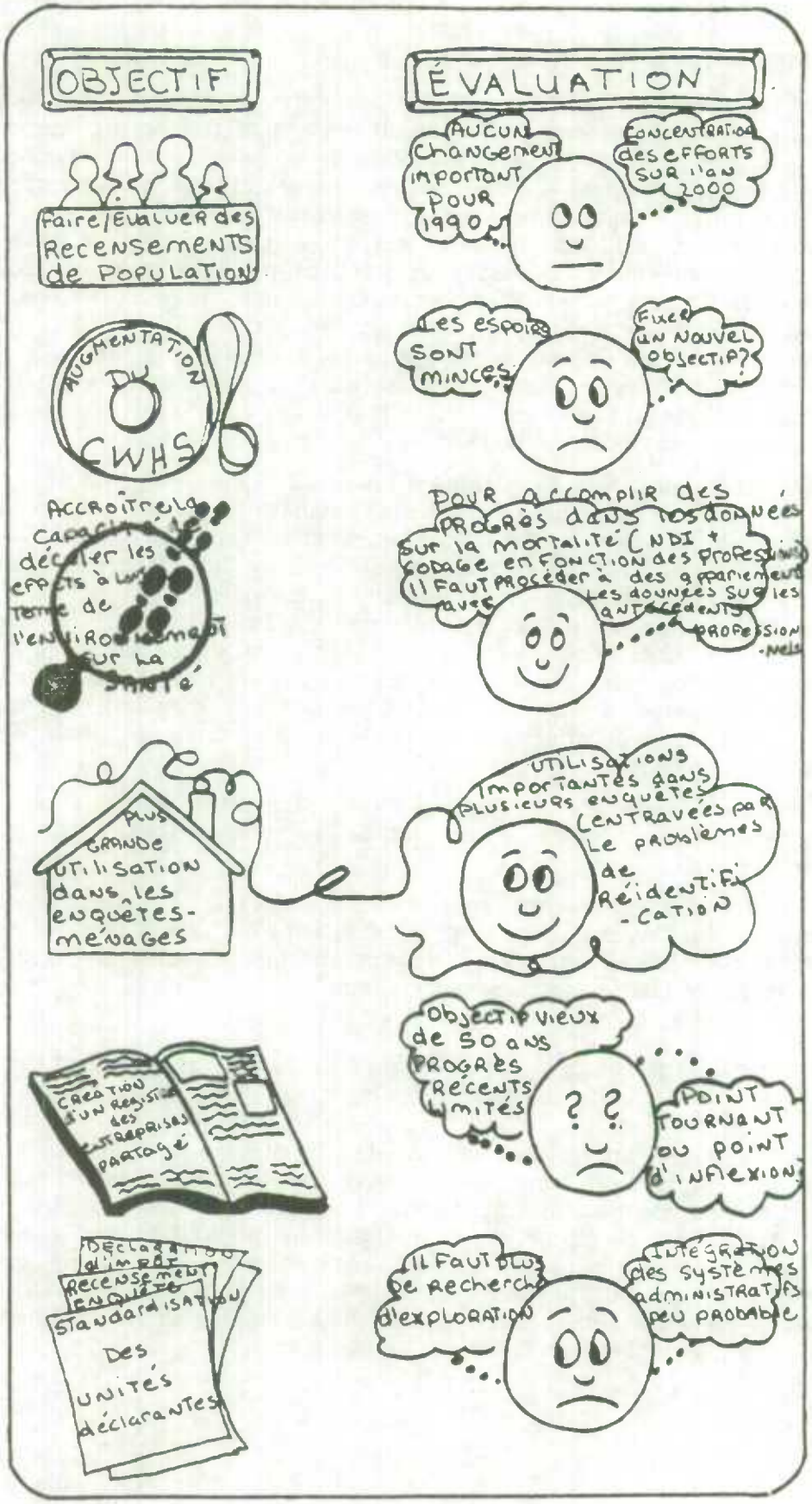
Pour les autres objectifs, la situation n'est sans doute pas plus mauvaise qu'en 1984, mais les progrès réalisés depuis ne sont rien de rejouissant. La situation globale n'est pas encourageante.

Quels sont les facteurs qui empêchent d'utiliser plus efficacement les dossiers administratifs à des fins statistiques; et peut-on faire quelque chose pour changer la façon dont ils fonctionnent? Ce sont les questions auxquelles nous allons tenter de répondre dans les deux autres parties du présent document.

2. DÉTERMINANTES DES PROGRÈS DANS L'UTILISATION STATISTIQUE DES DOSSIERS ADMINISTRATIFS

Dans cette partie, nous examinons un certain nombre de facteurs qui, à notre avis, jouent un rôle particulièrement déterminant dans l'application rentable des dossiers administratifs aux programmes statistiques. Le facteur de succès le plus important est un consensus général sur les objectifs et ce consensus général n'a pas encore été obtenu. Tout le monde n'a pas la même vision de l'avenir et, à l'heure actuelle, on ne connaît pas de moyen pour mettre tout le monde d'accord ou pour nous guider dans cette voie. Outre ce problème général, il conviendrait de mentionner quatre autres facteurs qui pourraient aussi avoir leurs équivalents au Canada ou ailleurs dans le monde.

Figure 4 Progrès accomplis en vue de atteindre les objectifs: 1984-1987



- **Changements dans les systèmes de dossiers administratifs**

  Quelle est leur importance et quels mécanismes peut-on développer pour s'y adapter?

- **Lois et politiques régissant l'accès et la divulgation**

  Ces lois et politiques influent sur l'accès des organismes statistiques aux données administratives, sur le droit d'apparier des données administratives et des données de recensement ou d'enquête et sur la mesure dans laquelle des données et des microdonnées agrégées, quelles soient en tout ou en partie de sources administratives, peuvent être communiquées à des fins statistiques.

- **Attitudes et perceptions à l'égard de l'utilisation statistique des données administratives**

  Il est essentiel de tenir compte des attitudes et des perceptions de tous les groupes concernés, à savoir les services responsables des systèmes de données administratives, les organismes statistiques qui utilisent les données et le public.

- **Coordination des activités au sein du système statistique américain**

  Les trois éléments énumérés précédemment sont susceptibles d'être importants pour tout les pays qui utilisent des données administratives à des fins statistiques. Toutefois, celui-ci revêt une importance particulière aux États-Unis principalement a cause de son système statistique décentralisé.

Dans le reste de cette partie, nous allons examiner chaque facteur plus en détail, en nous concentrant surtout sur ce qui s'est produit depuis le milieu de 1984.

2.1 Changements dans les systèmes de dossiers administratifs

La couverture, le contenu et la qualité des systèmes de dossiers administratifs et leur accessibilité font l'objet de changements fréquents, selon les exigences des programmes et les ressources disponibles. Quand les systèmes grossissent, les possibilités d'utilisation statistique augmentent. Par contre, quand les systèmes sont réduits, se détériorent ou deviennent moins accessibles, il peut en résulter une réduction des utilisations statistiques. Certains auteurs ont exprimé leur inquiétude au sujet de cette possibilité d'effets négatifs, notamment Butz (1984) dans ses commentaires sur nos six objectifs.

> *Quand on utilise des dossiers administratifs à la place de données statistiques au lieu de s'en servir pour les compléter, de sérieux problèmes peuvent survenir et surviennent effectivement de temps en temps. Comme substituts, les dossiers administratifs supplantent les données statistiques. Cela peut laisser un organisme statistique dangereusement vulnérable à des changements et à des échecs hors de son contrôle et exposer les utilisateurs à perdre des données importantes.*

Les résultats obtenus jusqu'ici révèlent que les utilisateurs statistiques des dossiers administratifs doivent être prêts à investir beaucoup de ressources pour avoir un contrôle sur les changements qui pourraient être apportés aux systèmes s'ils veulent s'éviter des surprises désagréables et profiter des nouvelles possibilités. Il faut absolument développer des canaux de communication efficaces entre les services responsables des données et les utilisateurs statistiques, et les utilisateurs doivent activement chercher à établir et à entretenir ces liens. S'ils y parviennent, il est même possible qu'ils puissent exercer une influence quelconque sur les types de changements qui pourraient se produire.

Les changements survenus dans les systèmes de dossiers administratifs depuis le milieu de 1984 et qui présentent le plus d'intérêt pour les statisticiens sont ceux qui résultent de

l'adoption de la Tax Reform Act de 1986. Comme nous l'avons dit dans la partie 1, les changements que cette loi a entraînés dans la couverture et le contenu des fichiers administratifs de l'IRS et de la SSA ont des effets importants sur le programme de la statistique du revenu de l'IRS (Statistics of Income (SOI) program) et sur plusieurs façons dont le Census Bureau utilise les dossiers de l'IRS et de la SSA (Jabine, 1987b). Par exemple, la Statistics of Income Division de l'IRS prévoit remanier en profondeur son échantillon des particuliers d'où elle tire ses statistiques du revenu des particuliers. Dans deux communications distinctes présentées à ce symposium, Gates (1987) et Hanczaryk et Jonas (1987) font ressortir clairement que le Census Bureau a déployé beaucoup d'efforts pour évaluer l'effet des changements résultant de la réforme fiscale sur son programme. (Ces documents illustrent certains des éléments de la stratégie active que doivent adopter les organismes statistiques pour se tenir au courant des changements apportés au contenu et à la qualité des systèmes de dossiers administratifs.)

Un autre changement important a été le développement graduel des fichiers du programme Medicare jusqu'au point où ils sont maintenant plus accessibles à des fins statistiques non pas seulement comme outil pour savoir qui fait partie de la population de 65 ans et plus, mais aussi comme moyen d'améliorer les données des enquêtes sur la santé effectuées auprès des personnes de ce groupe d'âge. La Heath Care Financing Administration a entrepris récemment plusieurs études analytiques fondées sur des appariements des dossiers Medicare et des dossiers administratifs d'autres organismes fédéraux et des États. L'utilisation et les coûts des services Medicare la dernière année de vie ont été étudiées par cause du décès en appariant d'une part les dossiers Medicare et d'autre part le National Death Index et le fichier sur la mortalité du NCHS (Riley et coll., 1987). L'utilisation des services Medicare par les travailleurs invalides prestataires de la sécurité sociale a été étudiée en appariant les dossiers Medicare et les fichiers administratifs du programme d'invalidité de la SSA (Bye et coll., 1987). L'incidence des services de soins d'infirmière à domicile sur les dépenses totales des régimes Medicare et Medicaid a été étudiée en appariant les fichiers administratifs Medicare et Medicaid se rapportant à quatre États (McMillan et coll., 1987).

Jabine (1984) a souligné la nécessité de déployer des efforts plus systématiques pour obtenir l'information sur les changements qu'on est en train d'apporter ou qu'on pourrait éventuellement apporter aux systèmes de dossiers administratifs et pour diffuser cette information aux utilisateurs d'un bout à l'autre du système statistique américain. Il nous fait plaisir d'indiquer qu'en s'appuyant sur les résultats des travaux préliminaires effectués par Crane et Kleweno (1985) le Census Bureau a relevé ce défi et construit un système d'information à partir de dossiers administratifs (Administrative Records Information System, ARIS), comme le mentionne Gates dans la communication qu'il a présentée à ce symposium.

Ces résultats sont encourageants: nous estimons que les organismes statistiques sont plus au courant aujourd'hui des possibilités et des problèmes créés par les changements survenant dans les systèmes de dossiers administratifs et réussissent mieux à exercer un contrôle sur ces changements.

2.2 Lois et politiques régissant l'accès et la divulgation

Les utilisations statistiques des dossiers administratifs dans notre système statistique décentralisé dépendent beaucoup des lois, règlements et politiques régissant la divulgation de dossiers identifiables. Comme nous l'avons indiqué dans la partie 1, la divulgation des dossiers administratifs de l'IRS et de la SSA, qui offrent le plus de possibilités d'utilisation à des fins statistiques, est très étroitement contrôlée par des lois et des règlements. Pour pouvoir utiliser les données administratives contrôlées par les États, comme les registres de l'état civil et les dossiers conservés pour le régime d'assurance-chômage, il faut se

conformer à un large éventail de lois propres à chaque État et par lesquelles les États régissent l'accès à leurs dossiers.

Les mêmes lois qui régissent le transfert de dossiers identifiables entre les organismes stipulent également qu'il est strictement interdit, dans presque tous les cas, de divulguer tout renseignement identifiable à partir d'agrégats de données publiés ou de fichiers de microdonnées accessibles au grand public. Cette exigence est toujours formulée en termes absolues -- aucune divulgation -- encore que la plupart des statisticiens admettront qu'aucune donnée le moindrement importante ne peut être publiée sans un certain risque de divulgation de renseignements identifiables (voir par exemple Paass, 1985; Cox et coll., 1985; Duncan et Lambert, 1987). Il n'est pas facile de quantifier ce risque, mais on estime généralement qu'il augmente de plus en plus parce qu'il est plus facile d'obtenir des données se rapportant à des personnes identifiables et que les techniques d'appariement de dossiers sont de plus en plus sophistiquées.

Une autre question importante est la suivante: que doit-on dire à ceux qui fournissent les renseignements sur les façons dont on entend ou dont on pourrait entendre utiliser cette information à des fins statistiques? Ceux qui produisent des déclarations d'impôt doivent-ils être mis au courant de toutes les façons dont on entend utiliser à des fins statistiques les renseignements qu'ils fournissent? Que doit-on dire aux répondants d'enquête sur la façon dont on entend apparier leurs réponses aux renseignements les concernant qui existent déjà dans des systèmes de dossiers administratifs?

Ces questions étaient déjà importantes en 1984 (American Statistical Association, 1977) et on leur a accordé de plus en plus d'attention depuis ce temps (voir, par exemple, Gastwirth, 1986; Jabine, 1986; Scheuren, 1986). Il convient de mentionner ici quelques résultats.

- Comme nous l'avons dit dans la partie 1.5, on s'efforce actuellement de créer des lois pour favoriser le partage de listes d'entreprises en supposant qu'aucune donnée obtenue directement de l'IRS par le Census Bureau ne sera retransmise à d'autres organismes.
- En 1985, le National Agriculture Statistics Service (NASS) a réussi à faire adopter une loi protégeant de façon rigoureuse la confidentialité des données qu'il recueille dans ses enquêtes, mais qui n'interdit toutefois pas de communiquer ces renseignements à d'autres organismes à des fins statistiques.
- En 1986, après de longues négociations, le National Cancer Institute a mis sur pied un mécanisme pour obtenir de la Social Security Administration des SSN à transmettre ensuite à l'IRS (par l'intermédiaire du National Institute for Occupational Safety and Health) en vue de se faire communiquer les adresses courantes des personnes suivies dans des études de suivi épidémiologiques. L'élément clé de cette entente était l'engagement formel de l'IRS qu'il utiliserait les données fournies par le NIOSH seulement pour avoir accès aux renseignements sur les adresses courantes, c'est-à-dire qu'il ne les utiliserait pas pour vérifier d'une façon quelconque si les gens se conforment aux règlements en vigueur.
- Comme on le mentionne dans la partie 1.4, le Census Bureau a adopté une politique plus restrictive concernant la diffusion des fichiers de micro-données qui sont accessibles au public et qui contiennent des données de dossiers administratifs appariées à des données de recensement ou d'enquête recueillies en vertu du titre 13.
- Il semble qu'on soit de plus en plus nombreux, autant chez les producteurs que chez les utilisateurs de données, à convenir qu'il faut d'autres mécanismes que les fichiers à grande diffusion pour permettre aux chercheurs d'avoir accès aux fichiers de microdonnées dont la nature en interdit le libre accès. Deux approches sont mentionnées: des arrangements de diffusion limitée qui prévoient des pénalités sévères pour tout utilisateur surpris à divulguer des données identifiables (et une

compensation pour les personnes lésées par la divulgation des renseignements les concernant) et l'accès interactif aux fichiers de données des producteurs des données, avec examen préliminaire des produits pour empêcher la divulgation des données identifiables.

- Le Committee on National Statistics et le Social Science Research Council ont travaillé conjointement à chercher des solutions aux problèmes de la confidentialité des données et de l'accès aux données. Une conférence sur l'accès aux données publiques (Conference on Access to Public Data) a eu lieu en novembre 1985 (Pearson, 1986), tandis qu'en septembre 1987 un atelier était organisé pour discuter des questions d'accès et de confidentialité relativement à une proposition d'étude de suivi qui serait fondée sur la Longitudinal Retirement History Survey de la SSA. Une vaste étude en panel sur la confidentialité et l'accès aux données (Confidentiality and Data Access) sera entreprise au milieu de 1988. Même si l'étude en panel ne se limitera pas uniquement aux statistiques fondées sur les dossiers administratifs, les questions étudiées auront un lien étroit avec les utilisations statistiques des dossiers administratifs.

Certains des événements qui se sont déjà produits ont contribué de manière effective à faciliter l'utilisation statistique des dossiers administratifs, mais ils représentent tout au plus des progrès mineurs. Si l'on regarde ce qui s'est passé par rapport à chacun des six objectifs discutés dans la partie 1 du présent document, il faut conclure que, sauf pour l'objectif 1 (utilisation des dossiers administratifs dans la réalisation des recensements décennaux et la production des estimations courantes de la population), il y a des restrictions légales qui empêchent d'atteindre certains des résultats souhaités dans chaque cas. Les lois peuvent, bien entendu, être modifiées, mais la possibilité que de nouvelles lois soient adoptées dépend beaucoup des attitudes et des perceptions des détenteurs des données, des utilisateurs des données et du grand public. Ces attitudes et perceptions sont étudiées dans la prochaine partie.

2.3 Attitudes et perceptions à l'égard de l'utilisation statistique des dossiers administratifs

Essentiellement, qu'on se dirige ou non vers une utilisation statistique accrue des dossiers administratifs dépend des attitudes institutionnelles des organismes comme l'IRS et la SSA qui ont charge des principaux systèmes de dossiers administratifs et des organismes comme le Census Bureau qui sont les principaux utilisateurs statistiques des dossiers administratifs.

Depuis l'adoption de la Tax Reform Act de 1986 (et même avant), l'IRS s'est toujours opposé à toute extension des utilisations non fiscales des données des déclarations d'impôt, qu'il s'agisse d'utilisations statistiques ou non. L'IRS justifiait sa position en alléguant que toute extension des utilisations des données fiscales pouvait rendre les gens moins disposées à remplir toutes les obligations de la loi fiscale. Cette préoccupation conduisit directement à l'adoption de la méthode de rechange consistant à établir des listes d'entreprises, qui est une des solutions actuellement envisagées et qui exclurait l'utilisation directe des données de l'IRS. (Voir la partie 1.5.) En fait, la position personnelle et professionnelle du deuxième auteur du présent document est qu'il n'est pas souhaitable d'augmenter par des lois l'accès aux dossiers fiscaux. D'autres solutions doivent être envisagées avant celle-là.

La position du Census Bureau sur la question des utilisations statistiques des données de l'IRS est ambivalente. Comme le signale Gates dans sa communication présentée à ce symposium, le Census Bureau utilise beaucoup les dossiers administratifs de l'IRS pour faire des recensements et des enquêtes économiques, pour produire des estimations courantes de la population et pour évaluer le taux de couverture du recensement de la

population (Census of Population). Ces utilisations ont été entièrement décrites dans des rencontres et rapports techniques. D'un autre côté, le Census Bureau a jusqu'à récemment hésité à informer les répondants de la possibilité que l'on procède à des appariements des données d'enquête et des renseignements sur leur compte obtenus de l'IRS. Même si les dossiers identifiables circulent dans un sens seulement, c'est-à-dire de l'IRS au Census Bureau, on s'inquiète de ce que la perception d'un lien entre les deux organismes dans l'esprit du public puisse diminuer la capacité du Census Bureau d'obtenir des taux de réponse élevés dans ses recensements et ses enquêtes. Le personnel du Census Bureau invoque l'opposition de plus en grande manifestée à l'endroit des recensements en Allemagne et dans les Pays-Bas et attache beaucoup d'importance au soin d'éviter d'entreprendre tout projet qui pourrait provoquer les mêmes réactions au pays (Butz, 1984, 1985b).

Ainsi, les attitudes des gardiens des dossiers administratifs et des responsables des organismes qui les utilisent à des fins statistiques dépendent au moins en partie de leurs opinions quant à la façon dont le public pourrait réagir à des utilisations statistiques des dossiers administratifs qui seraient largement connues du public. On peut s'y prendre de deux façons pour savoir ce que le grand public pense des utilisations statistiques des dossiers administratifs: en analysant ce qui se dit dans les médias ou en interrogeant directement la population dans des sondages d'opinion.

Depuis 1984, le meilleur exemple de réaction du public dont aient fait état les médias est peut-être le tollé de protestations qu'avait suscité en Suède une étude longitudinale, à laquelle on a dû mettre fin; la base de données de cette étude, tenue à jour au moyen de la collecte de données, servait à un projet de recherche en sciences sociales appelé Metropolit. La base de données appariait des données longitudinales d'enquêtes et de plusieurs sources de dossiers administratifs se rapportant à toutes les personnes nées dans la région métropolitaine de Stockholm en 1953 et y vivant au moment où la base de données a commencé dix ans plus tard. L'affaire est trop compliquée pour être racontée ici en détail, mais en gros voici ce qui est arrivé: quand une bonne partie des personnes incluses dans la base de données se sont rendu compte exactement de son contenu et de son ampleur, elles ont réagi si vivement qu'il a fallu supprimer certains identificateurs de la base de données, ce qui a eu pour effet d'empêcher de continuer les appariements.

Un fait assez important dans le contexte du présent document, c'est qu'au moment où la controverse a éclaté, les taux de réponse à l'enquête sur la population active en Suède (Swedish Labor force Survey) ont baissé de 5 points de pourcentage environ et ne sont pas revenus depuis à leur niveau précédent (Dalenius, 1986).

Il est difficile de recourir à des enquêtes pour connaître directement les attitudes du public à l'égard des utilisations statistiques des dossiers administratifs parce que la question n'intéresse pas beaucoup de monde et que la plupart des gens n'ont qu'une très vague idée de ce qui se fait. Les questions portant sur les utilisations non fiscales des données de l'IRS qui ont été incluses dans les enquêtes commanditées par l'IRS en 1984 et 1986 ont produit des résultats assez incohérents, peut-être parce que le peu d'importance attribué à ce sujet par la plupart des répondants a fait que leurs réponses ont été influencées par le libellé des questions dans les deux enquêtes (Gonzalez et Scheuren, 1985; Scheuren, 1986).

Pour son enquête de 1987 sur les attitudes des contribuables (Taxpayer Attitudes Survey), l'IRS a décidé de modifier sa façon d'aborder le sujet. Les questions portant sur ce sujet (questions 97 à 99) se succèdent dans l'ordre suivant (voir l'annexe jointe aux remarques faites par Tom Jabine, 1987a, en qualité de panéliste à la séance de clôture du symposium).

— Une note préliminaire amène le sujet du partage des données entre les divers organismes gouvernementaux. Un exemple concret est donné.

- La Q. 97, qui utilise une échelle graduée d'un à cinq, demande aux répondants ce qu'ils pensent de six facteurs précis qui pourraient inciter les gens à être favorables ou défavorables au partage des données. En fait, il y a cinq facteurs seulement: comme test de cohérence interne, la Q. 97f mentionne les mêmes facteurs que la Q. 97c, mais de façon inversée.
- La Q. 98, qui vient après que les répondants ont pris connaissance des facteurs mentionnés à la question précédente, demande maintenant aux répondants d'indiquer leur point de vue global sur le partage des données. La Q. 98b permet aux répondants de justifier la réponse choisie à la Q. 98a.
- La Q. 99 demande aux répondants ce qu'ils pensent de quatre utilisations autres que fiscales des données de l'IRS. Deux des utilisations (a et d) sont statistiques et deux ne sont pas statistiques. Dans chaque cas, l'organisme qui obtient les données et l'objectif du partage sont indiqués. L'ordre dans lequel les utilisations étaient présentées aux répondants variait.

Les taux de réponse globaux pondérés à ces questions sont également affichés dans la même annexe (Jabine, 1987a). La population cible se composait de personnes qui produisent normalement des déclarations d'impôt fédérales; il est en outre précisé qu'advenant le cas où deux personnes ou plus figurent sur une même déclaration, elles seront représentées par le contribuable considéré comme étant celui qui connaît le mieux la déclaration. Voici les résultats les plus intéressants.

- La très grande majorité des répondants veulent savoir quels organismes ont des renseignements sur eux et pourquoi ces organismes veulent ces renseignements.
- Des majorités plus petites s'accordent à reconnaître que le partage des données réduirait le fardeau de réponse imposé au public et les dépenses que devrait engager le gouvernement pour obtenir les renseignements dont il a besoin.
- Dans l'ensemble, il y a un peu plus de répondants opposés que de répondants favorables au partage des données (41 contre 38 pour cent). Le nombre de ceux qui s'opposent violemment dépasse de beaucoup le nombre de ceux qui sont très favorables.
- Les réactions au partage des données qui sont motivées par des raisons précises varient relativement peu. La communication de données au Département de la Justice à des fins d'enquêtes criminelles est la raison avec laquelle le plus de répondants sont d'accord, tandis que la communication de données aux gouvernements des États pour améliorer leurs méthodes de collecte de données est la raison avec laquelle le plus de répondants ne sont pas d'accord. Les utilisations statistiques des données par le Census Bureau et le Département du Commerce se classent entre les deux utilisations non statistiques.
- Face à des cas précis de partage des données plutôt qu'au concept général, moins de répondants sont indécis et les réactions sont un peu plus favorables.

Les données de ce genre sont très utiles pour mieux comprendre les attitudes du public à l'égard du partage des données, et c'est pourquoi nous nous employons actuellement à les analyser plus en détail. Des fichiers de microdonnées tirées de l'enquête seront mis à la disposition des chercheurs par l'intermédiaire de l'Inter-university Consortium for Political and Social Research. Nous devons toutefois faire remarquer que l'attitude du public à l'égard de ces sujets est très inconstante. Dans sa campagne pour les élections générales qui auront lieu cette année en Australie, le parti au pouvoir a dit aux électeurs qu'il voulait introduire une carte d'identité nationale. Des sondages d'opinion ont révélé que plus de 60 pour cent des Australiens étaient favorables au projet. Toutefois, après les élections, une opposition très nette s'est élevée et les enquêtes ont montré que le

pourcentage des gens favorables au projet avait baissé de près de la moitié (**Washington Post, 1987**).

2.4 Coordination des activités au sein du système statistique américain

Comme notre système statistique décentralisé n'a, par définition, pas d'organe central fort de coordination, il est difficile pour les organismes statistiques d'exercer beaucoup d'influence sur la couverture et le contenu des systèmes de dossiers administratifs ou sur les conditions d'accès aux sources de données administratives à des fins statistiques. L'Office of Statistical Policy (OSP), qui relève de l'Office of Management and Budget, est officiellement chargé de coordonner les activités statistiques fédérales. Avec des ressources limitées, l'OSP porte son attention principalement sur le contrôle du fardeau de réponse, la mise à jour des systèmes de classification type, comme le Standard Metropolitan Statistical Areas, ainsi que sur la collaboration entre les organismes sur des questions techniques par l'intermédiaire du Federal Committee on Statistical Methodology (FCSM). Comme nous l'avons décrit dans le document que nous avons présenté en 1985, l'Administrative Records Subcommittee du FCSM a commandé l'exécution de plusieurs projets visant à échanger de l'information sur certains aspects techniques des utilisations statistiques des dossiers administratifs. Ces projets se poursuivent, mais avec moins de ressources, depuis 1984.

L'Office of Statistical Policy ne cherche toutefois pas à amener les organismes statistiques fédéraux à s'engager dans la planification à moyen ou à long terme du système statistique dans son ensemble. Comme nous l'avons dit dans le document que nous avons présenté en 1985:

> *Beaucoup d'organismes statistiques font un excellent travail de planification stratégique à long terme. Toutefois, ils le font dans le cadre de leurs propres fonctions, programmes et intérêts et non comme entités intégrantes d'un système statistique global conçu pour répondre aux besoins de renseignements du gouvernement et du public le plus efficacement possible. Par conséquent, nous pensons qu'il y a un préjugé en faveur des programmes de collecte directe des données sous le contrôle d'un seul organisme de préférence aux programmes qui recourent davantage aux dossiers administratifs, mais qui nécessitent le partage des données et une étroite collaboration entre deux ou plusieurs organismes. Ce phénomène explique la rareté relative des ressources consacrées à la recherche et au développement de nouvelles utilisations des dossiers administratifs et peut aussi expliquer l'échec des tentatives effectuées pour faire adopter les lois qu'il faut pour créer un registre des entreprises partagé et pour mettre sur pied d'autres projets nécessitant des échanges de données identifiables.*

La situation n'a pas tellement changé.

Quand un tel vide existe, divers groupes font ce qu'ils peuvent pour le combler. Nous allons maintenant énumérer certains projets d'organismes gouvernementaux et non gouvernementaux en vigueur en 1984 qui ont eu ou devraient avoir une certaine influence sur les utilisations statistiques des dossiers administratifs.

Secteur gouvernemental

- Au niveau des politiques, nous avons déjà mentionné le groupe de travail sur la qualité des statistiques économiques (Working Group on the Quality of Economic Statistics) de l'Economic Policy Council et ses recommandations sur le partage des listes d'entreprises (partie 1.5). Il convient de noter que les organismes concernés sont déjà

en retard sur le calendrier recommandé pour la réalisation des travaux. D'autres recommandations du groupe de travail, en particulier celles qui concernent l'amélioration des statistiques sur le commerce des marchandises, pourraient aussi avoir un effet sur d'importantes utilisations statistiques des dossiers administratifs, même si l'on ne sait pas toujours pas très bien si elles vont mener à une plus grande ou à une moins grande dépendance à l'égard des sources administratives.

- Nous avons aussi manifesté notre appui au projet qui a abouti récemment à l'établissement par le Census Bureau d'un système d'information à partir de dossiers administratifs (Administrative Records Information System). Un autre nouveau projet du Census Bureau, qui a facilité l'échange de renseignements au niveau technique, est sa série de conférences annuelles sur la recherche qui a commencé en 1985. Les comptes rendus de ces conférences sont d'une valeur inestimable (voir par exemple Bureau of the Census, 1987). Plusieurs documents présentés à ces conférences concernent les utilisations des dossiers administratifs. Par exemple, à la conférence de 1987, Tippett (1987) et Diemer (1987) ont discuté de l'utilisation des données de sources administratives comme moyen d'évaluer la qualité des données tirées de recensements et d'enquêtes sur le logement et Colledge (1987a) a décrit le rôle que les données fiscales pouvaient jouer dans un remaniement en profondeur des enquêtes-entreprises de Statistique Canada.

- L'IRS a continué de publier régulièrement les documents présentés aux réunions annuelles de l'American Statistical Association sur l'utilisation statistique des dossiers administratifs (voir par exemple Internal Revenue Service, 1987b).

- Depuis 1984, le Committee on National Statistics a continué, au moyen d'études en panel et d'autres projets parrainés principalement par des organismes fédéraux, d'étudier des questions de politique et des questions techniques concernant de près le système statistique américain. Dans bon nombre de ces projets, il était question d'utilisation statistique des dossiers administratifs. Dans la partie 1.1, nous avons cité le Panel on Decennial Census Methodology. Un des trois principaux points sur lesquels on a demandé au groupe de travail de se pencher était "L'utilisation des dossiers administratifs, incluant l'examen de la possibilité d'utiliser divers types de dossiers pour améliorer l'exactitude des chiffres du recensement et l'efficacité des opérations de recensement" (National Research Council, 1985). Deux nouvelles études en panel viennent juste de démarrer, une sur les statistiques du commerce extérieur et une autre sur la confidentialité et l'accès aux données. Dans ces deux études, les questions liées à l'utilisation statistique des dossiers administratifs devraient être un élément important de l'analyse.

Secteur non gouvernemental

- Le Council of Professional Associations on Federal Statistics (COPAFS) a continué d'être une tribune où les dirigeants des organismes pouvaient se rencontrer et rencontrer les représentants de sociétés professionnelles pour discuter de sujets d'actualité en matière de politique statistique. Le COPAFS a des liens étroits avec le Congrès et sert de bureau central où s'échange l'information sur les mesures prises par le Congrès concernant les organismes statistiques.

- En 1985, la National Association of Business Economists a formé un comité sur les statistiques (Statistics Committee) "pour servir de liaison entre les membres de la NABE, et les organismes statistiques fédéraux, locaux et ceux des États, ainsi que pour aider à assurer l'intégrité, la couverture, la précision et l'actualité des statistiques sur les entreprises" (National Association of Business Economists, 1987). Un des sujets de préoccupation du comité a été la qualité des listes d'entreprises utilisées dans les enquêtes économiques.

— L'American Economic Association a créé un comité, présidé par Thomas Juster, pour vérifier si les concepts, définitions et systèmes de classification utilisés dans nos statistiques économiques conviennent toujours dans le cadre des changements structurels survenus récemment dans nos institutions économiques et sociales.

Des projets comme ceux-là sont importants pour la santé de la communauté statistique fédérale américaine et méritent d'être appuyés et encouragés. Ils sont toutefois encore loin d'être suffisants pour amener les organismes statistiques fédéraux américains à adopter une stratégie globale cohérente concernant l'utilisation statistique des dossiers administratifs (et en fait sur beaucoup d'autres questions connexes).

3. AUTRE PRÉVISION

Évidemment, la réponse à la question "Où allons-nous?" dans le titre ne peut être que conjecturale. Une chose semble toutefois assez certaine: les progrès technologiques de l'ère de l'information continueront à un rythme rapide. Ces progrès ouvriront des perspectives intéressantes pour une utilisation statistique plus variée et plus efficace des dossiers administratifs.

Comment le système statistique américain peut-il profiter de ces nouvelles possibilités? Pour nous aider à réfléchir à cette question, il peut être utile de considérer l'utilisation statistique des dossiers administratifs comme un **système.** Tous les systèmes, qu'ils soient statistiques ou pas, comportent trois éléments de base: des intrants, des procédures appliquées aux intrants et des produits. Dans la figure 5, nous présentons ce que nous considérons comme quelques-uns des principaux éléments du système qui nous intéresse.

Intrants.

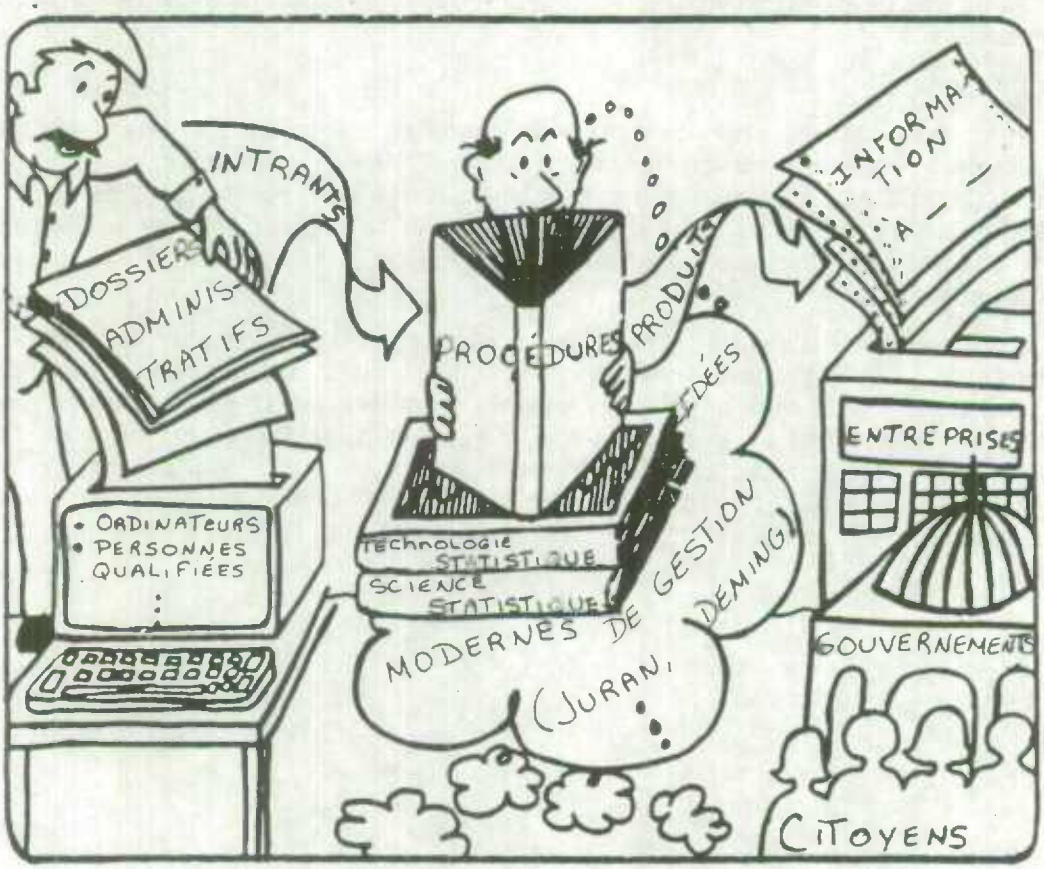
Dans nos analyses précédentes, nous avons concentré notre attention, peut-être trop, sur les dossiers administratifs comme intrants. Ils sont bien entendu la condition **sine qua non;** et le contrôle des changements au niveau de leur couverture, de leur contenu et de leur qualité et des conditions d'accès -- comme le Census Bureau le fait actuellemenmt systématiquement dans son Administrative Records Information System -- est essentiel à toute l'entreprise.

Il y a toutefois d'autres intrants qui sont tout aussi importants ou peut-être même plus importants. Les organismes statistiques doivent avoir des ordinateurs capables de traiter des bases de données très grosses et très complexes. Il est également essentiel qu'on augmente le nombre de personnes qualifiées ayant vraiment le souci de l'excellence. Il faudra que ces gens soient familiarisés avec les outils de pointe présentés dans la colonne "procédures" de la figure 5: méthodes statistiques, contrôle de la qualité, télécommunications et, éventuellement, intelligence artificielle (IA).

Figure 5

Utilisation statistique des dossiers administratifs considérée comme un système: quelques exemples

Intrants	Procédures	Produits
Dossiers administratifs Ordinateurs et périphériques Personnel de l'organisme	**Directes:** Conception de systèmes Traitement de l'information **Cadre:** Structure organisationnelle Stratégie **Outils:** Science statistique Contrôle et amélioration de la qualité Systèmes experts, IA, Télécommunications	**Produits:** Publications **Fichiers de données:** Données agrégées Microdonnées **Clients:** Gouvernements Entreprises Universités Particulier



Procédures

Un élément clé du traitement de l'information qui convertit des dossiers administratifs utilisés comme intrants en produits statistiques est l'**appariement de dossiers.** On avance à grands pas dans le développement de systèmes d'appariement de dossiers à usages multiples fondés sur des modèles (Internal Revenue Service, 1985; Jabine et Scheuren, 1986). Grâce aux premiers travaux de Fellegi et d'autres auteurs, la théorie sous-jacente semble bien comprise (Kirkendall, 1985); toutefois, pour pouvoir l'appliquer, il faut des compétences et une expérience qui manquent actuellement. Il faut poursuivre les recherches et continuer de partager les connaissances et de développer des systèmes d'appariement de dossiers faciles à utiliser (voir Internal Revenue Service, 1985, Recommandations, pp. 3-4). Ce symposium lui-même représente un effort considérable en vue d'atteindre cet objectif, comme en témoignent les nombreux documents présentés sur ce sujet.

Le cadre organisationnel dans lequel ces procédures sont appliquées n'est pas susceptible de changer beaucoup. On peut comparer notre système statistique décentralisé à un réseau informatique dispersé comportant plusieurs gros ordinateurs (et non pas seulement un très gros système) liés entre eux par des mini-ordinateurs et des PC pour former un réseau. Le problème, c'est que, pour poursuivre l'analogie, le degré de **connectivité** n'est pas suffisant. Il nous faut absolument un processus de planification stratégique à la grandeur du système pour nous aider à décider comment répartir les ressources (en particulier dans le domaine de la recherche et du développement) affectées aux divers programmes statistiques fédéraux. Des questions de sphères de compétence nous éloignent de notre mission principale de fonctionnaires. La diversité des points de vue est saine et nécessaire à tout progrès réel, mais il nous faut de meilleurs moyens pour résoudre nos divergences d'opinion. Certains problèmes, comme la question de la création de listes d'entreprises partagées (comme nous l'avons noté plus haut) remontent à 50 ans et n'ont pas encore été résolus. Il faudrait surmonter de telles lacunes.

Produits

Il faut accorder plus d'attention à l'aspect des produits du système. Il faut nous demander si nous sommes suffisamment "tributaires de la clientèle". À notre avis, il faut prendre des mesures pour que les programmes statistiques américains atteignent un plus grand nombre de gens, en particulier dans le domaine des dossiers administratifs. Nous sommes censés être à l'"Âge de l'information"; pourtant bon nombre des produits que nous produisons et le moment où nous les produisons n'ont pas changé suffisamment pour refléter adequatement ce phenomène. En particulier, à cause des contraintes de confidentialité, nous devons chercher d'autres moyens de produire des fichiers accessibles à tous les utilisateurs qui permettront aux utilisateurs d'avoir accès aux données dont ils ont besoin sans risquer de mettre en danger la confidentialité des renseignements personnels.

Commentaires finaux

Des améliorations pourront certainement être apportées dans des domaines précis. Par exemple, les programmes statistiques américains devraient profiter des améliorations importantes que tentent d'apporter certains organismes administratifs, en particulier l'Internal Revenue Service. À l'IRS, nous sommes très, très sérieux quand nous affirmons vouloir faire un meilleur travail pour la population des États-Unis (Scheuren, 1987). Cela veut dire qu'il faut améliorer non seulement les produits, mais aussi la technologie. Le recours à l'intelligence artificielle et à d'autres méthodes exigeant beaucoup d'informatique par les organismes administratifs, en particulier dans le domaine des systèmes experts (Beckman, 1987), devrait aider à construire l'infrastructure méthodologique nécessaire pour utiliser encore plus les dossiers administratifs à des fins statistiques. La révolution des télécommunications sera aussi un facteur qui forcera peut-

être l'IRS à prendres des mesures comme traiter sur une grande échelle des déclarations d'impôt produites par des moyens électroniques. (Pour les déclarations de 1988, il se peut que l'IRS doive traiter jusqu'à un demi-million de déclarations faites de cette façon.) On peut donc imaginer pour l'avenir un système "sans déclaration", avec toutes les améliorations au niveau de la qualité et des délais de production que cette innovation peut comporter (Wedick, 1986; Internal Revenue Service, 1987a).

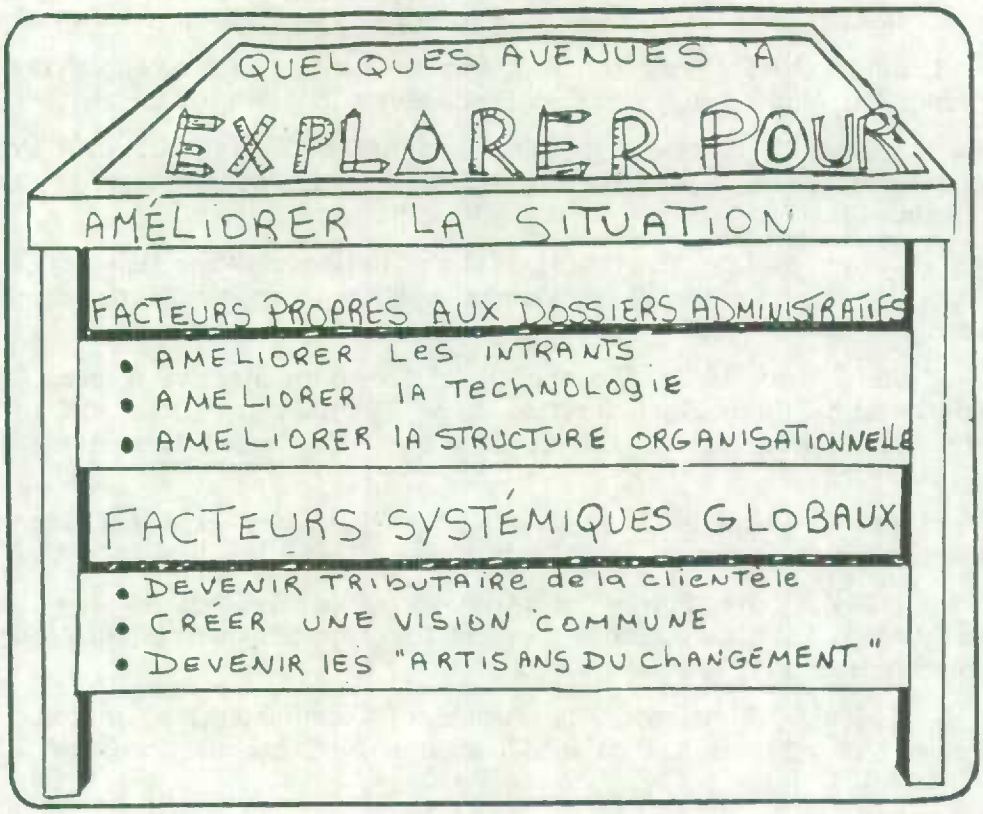
Autrement dit, sans réellement faire d'efforts, les organismes statistiques profiteront des forces opérantes des ministères administratifs. Pour tirer pleinement avantage de ces forces, il va toutefois falloir améliorer les structures organisationnelles servant à la production des statistiques gouvernementales (il faut en particulier améliorer la planification et la coordination des décisions prises en ce qui touche à la répartition des ressources consacrées à la R-D).

Enfin, nous aimerions signaler trois exigences systématiques globales qu'il faudrait combler pour réaliser des progrès dans ce domaine (voir la figure 6). En un certain sens, nous avons déjà parlé des deux premiers. Si les États-Unis veulent faire un meilleur usage stastistique de leurs dossiers administratifs, chacun d'entre nous, aussi bien les responsables de données comme la SOI Division de l'IRS que les employés des organismes statistiques comme nos amis du Census Bureau qui participent à ce symposium, devra répondre de ses actions à ses clients. Nous ne voulons pas seulement dire les clients en dernière analyse, c'est-à-dire le peuple américain, mais tous les clients intermédiaires aussi. En fait, nous sommes clients les uns des autres, comme on peut le voir dans la figure 1.

Comme l'a dit Peters, nous devons apprendre à écouter "naïvement" nos clients (Peters et Waterman, 1982). Naturellement, il y a beaucoup de sortes de clients, et être "tributaire de la clientèle" peut littéralement signifier devenir fou s'il n'y a pas de mécanisme pour résoudre les conflits. Peut-être que le meilleur mécanisme à mettre en place pour résoudre les conflits est de créer une vision commune des choses. Un processus de planification stratégique est un des outils qui ont déjà été mentionnés et qui pourraient être utiles ici. Mais la planification ne suffit pas. Pour développer une vision commune, il faut travailler fort à dégager les valeurs; cette volonté n'existe pas à l'heure actuelle. Chacun de nous semble déjà avoir assez de soucis comme cela sans devoir en plus essayer de résoudre des problèmes qui semblent durer depuis toujours. Nous n'avons pas perdu espoir, loin de là, mais à quoi servirait-il de donner des réponses toutes faites.

Un dernier point en terminant: nous avons deux façons d'envisager l'avenir, soit devenir les "artisans du changement" (Kanter, 1983), soit être à la remorque du changement. Il est inévitable que les dossiers administratifs servent de plus en plus à des fins statistiques. Qu'on soit d'accord ou pas avec ce phénomène, il progressera inévitablement, comme un glacier ou comme une avalanche. Pour reprendre une remarque faite au début: non seulement sommes-nous ceux **qui font ce qui est évident, mais pas facile,** mais nous sommes aussi ceux **qui travaillent dans un domaine appelé inévitablement à grandir en importance et en défis à relever.** Il y a un vieux proverbe chinois qui dit à peu près ceci: "Puissiez-vous vivre dans des temps intéressants". Eh bien! il ne fait aucun doute que dans ce domaine nous vivons dans des "temps intéressants". Que cela soit une malédiction ou non dépend au moins en partie de nous.

Figure 6



REMERCIEMENTS

Les auteurs tiennent à remercier Bettye Jamerson, Wendy Alvey et en particulier Beth Kilss pour leur aide à la rédaction du présent document. Ils remercient également Nancy Dutton, Dorothy Farmer et Darlene Reynolds pour avoir dactylographié les nombreuses versions préliminaires du présent document et de la communication présentée au symposium.

BIBLIOGRAPHIE

Alvey, W., et Scheuren, F. (1982). "Background for an Administrative Record Census", *Proceedings of the American Statistical Association, Social Statistics Section*, 137-146.

American Statistical Association (1977). "Report of Ad Hoc Committee on Privacy and Confidentiality", *The American Statistician*, Vol, 31, 59-78.

Beckman, T. (1987), "Trends in Selection and Development of Applications Using Artificial Intelligence Technology", document de travail non publié de l'IRS.

Bowie, C., et Kasprzyk, D. (1987). "L'utilisation de dossiers administratifs pour l'enquête sur le revenu et la participation aux programmes", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Statistique Canada, Ottawa, Ontario.

Brackstone, G. J. (1987a). "Utilisation des dossiers administratifs à des fins statistiques", *Techniques d'enquête*, vol. 13, n° 1, 35-51.

Brackstone, G. J. (1987b). "Utilisation statistique des données administratives: thèmes et défis", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Statistique Canada, Ottawa (Ontario).

Bristol, R. B., Jr. (1985). "Tax Modelling and the Policy Environment of the 1990's", *Multi-National Tax Modelling Symposium Proceedings*, Ottawa (Ontario), II-11-II-17.

Buckler, W. (1985). "Employer Reporting Unit Match Study (ERUMS): A Progress Report'" *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 434-437.

Buckler, W., et Smith, C. (1980). "The Continous Work History Sample (CWHS): Description and Contents", *Economic and Demographic Statistics*, Social Security Administration, 165-174.

Bureau of the Census (1983). "Feasability of an Administrative Record Census in 1990", rapport non publié du Subcommittee on an Administrative Record Census, Committee on the Use of Administrative Records in the 1990 Census, Département du Commerce des États-Unis.

Bureau of the Census (1987). *Proceedings of the Third Annual Research Conference*, Département du Commerce des États-Unis.

Butz, W. (1984). "The Future of Administrative Records in the Census Bureau's Demographic Activities", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 61-63.

Butz, W. (1985a). "Comment: The Future of Administrative Records in the Census Bureau's Demographic Activities", *Journal of Business and Economic Statistics*, vol. 3, n^{o} 4, 393-395.

Butz, W. (1985b). "Data Confidentiality and Public Perceptions: The Case of the European Censuses", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 90-97.

Bye, B., Riley, G., et Lubitz, J. (1987). "Medicare Utilization by Disabled-Worker Beneficiaries: A Longitudinal Analysis", document à paraitre dans le *Social Security Bulletin*, Vol. 50, no. 12, Decembre.

Carroll, J. (1985). "Comment: Uses of Administrative Records: A Social Security Point of View", *Journal of Business and Economic Statistics*, Vol. 3, no 4, 396-397.

Cartwright, D., Levine, B. et Buckler, W. (1983). "An Update on Establishment Reporting Issues: Practical Considerations", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 481-486.

Childers, D., et Hogan, H. (1984). "Matching IRS Records to Census Records: Some Problems and Results", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 301-306.

Colledge, M. (1987a). "The Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada", *Proceedings of the Census Bureau Third Annual Research Conference, Bureau of the Census*, 550-576.

Colledge, M. (1987b). "Utilisation des données administratives dans le projet de remaniement des enquêtes-entreprises", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Statistique Canada, Ottawa, (Ontario).

Cox, L., et Boruch, R. (1985). "Emerging Policy Issues in Record Linkage and Privacy", document présenté à la 45^{e} réunion de l'International Statistical Institute tenue à Amsterdam, Pays-Bas du 12 au 22 août 1985.

Cox, L., Johnson, B., McDonald, S., Nelson, D., et Vasquez, V. (1985). "Confidentiality Issues at the Census Bureau", *Proceedings of the Census Bureau First Annual Research Conference*, 199-218.

Crabbe, P., Sailer, P., et Kilss, B. (1983). "Occupation Data from Tax Returns: A Progress Report", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 312-317.

Crane, J., et Kleweno, D. (1985). "Project LINK-LINK: An Interactive Database of Administrative Record Linkage Studies", *Record Linkage Techniques* — 1985, Internal Revenue Service, 311-315.

Crouse, W., Schuster, L., Rosenberg, H., Kametani, D., et Sestito, J. (1983). "Using the Census Bureau's Occupation and Industry Coding System for Coding Death Certificates", Proceedings of the American Statistical Association, Section on Survey Research Methods, 306-311.

Czajka, J. (1987). "Une situation inversée: l'imputation de valeurs à des éléments d'information manquants lorsque les donneurs sont rares", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Ottawa, (Ontario).

Dalenius, T. (1986). "The 1986 Invasion of Privacy Debate in Sweden", rapport non publié.

Diemer, W. (1987). "Micro-Evaluation of the 1980 Census of Housing", *Proceedings of the Census Bureau Third Annual Research Conference*, Bureau of the Census, 437-476.

Duleep, H. (1986). "Incorporating Longitudinal Aspects into Mortality Research Using Social Security Administrative Record Data", *Journal of Economic and Social Measurement*, Vol. 14, 121-133.

Duncan, G., et Lambert, D. (1987). "The Risk of Disclosure for Microdata", *Proceedings of the Census Bureau Third Annual Research Conference*, 263-274.

Economic Policy Council (1987). "Report of the Working Group on the Quality of Economic Statistics, Département du Commerce des États-Unis, Washington (D.C.).

Energy Information Administration (1987). *Residential Energy Consumption Survey: Consumption and Expenditures, April 1984 Through March 1985*, Part 1: National Data, Appendix A, How the Survey was Conducted, Département de l'Énergie des États-Unis.

Fay, R. E., et Herriot, R. (1979). "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data", *Journal of the American Statistical Association*, Vol. 74, n° 366, 269-277.

Fellegi, I. P., et Sunter, A. B. (1969). "A Theory for Record Linkage", *Journal of the American Statistical Association*, Vol. 64, no 328, 1183-1210.

Flaherty, D. (1979). *Privacy and Government Data Banks: An International Perspective*, Mansell Publications, Londres, Royaume-Uni.

Gastwirth, J. (1986). "Ethical Issues in Access to and Linkage of Data Collected by Government Agencies", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 6-13.

Gates, G. (1987). "Évaluation de l'effet de la réforme fiscale sur les programmes du Census Bureau", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Statistique Canada, Ottawa, (Ontario).

Gonzalez, M., et Hoza, C. (1978). "Small-Area Estimation with Application to Unemployment and Housing Estimates", *Journal of the American Statistical Association*, Vol. 73, no 361, 7-15.

Gonzalez, M., et Scheuren, F. (1985). "Future Work by the Conference of European Statisticians on Population and Housing Censuses", communication présentée à la Trente-troisième séance plénière à la Conférence des Nations-Unies des statisticiens d'Europe.

Hanczaryk, P., et Jonas, J. (1987). "Traitement automatisé de l'assurance de la qualité des fichiers de dossiers administratifs", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Ottawa, (Ontario).

Hidiroglou, M. A., Morry, M., Dagum, E. B., Rao, J. N. K., et Särndal, C. E. (1984). "Evaluation of Alternative Small Area estmators Using Administrative Data", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 307-313.

Hinkins, S., Jones, H., et Scheuren, F. (1987). "Mise à jour des probabilités de sélection des déclarations d'impôt dans le cadre du programme de la statistique du revenu des sociétés", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Ottawa, (Ontario).

Hostetter, F., McCann, C., et Zirger, B. (1987). "Dossiers de l'impôt sur le revenu des sociétés utilisés à des fins d'analyse de la politique fiscale", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Ottawa, (Ontario).

Hostetter, S. (1987). "Measuring Income for Developing and Reviewing Individual Tax Law Changes: Exploration of Alternative Concepts", *Proceedings of the American Statistical Association, Section on Survey Research Methods*.

Howe, G. R., et Spasoff, R. A. (Eds.) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, University of Toronto Press.

Internal Revenue Service (1984). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. 2, Section VIII, "Summary and Prospects for the Future", Département du Trésor des États-Unis, 651-680.

Internal Revenue Service (1985). *Record Linkage Techniques 1985*, Proceedings of the Workshop on Exact Matching Methodologies, Département du Trésor des États-Unis.

Internal Revenue Service (1987a). "Report to Congress on the Return-Free Tax System", Département du Trésor des États-Unis.

Internal Revenue Service (1987b). *Statistics of Income and Related Administrative Record Research: 1986-1987*, Département du Trésor des États-Unis.

Jabine, T. (1984). "Proposal for an Administrative Records Monitoring System", dans *Statistical Uses of Administrative Records: Recent Research and Present Prospects* (Vol. 1). Internal Revenue Service, Département du Trésor des États-Unis, 37-38.

Jabine, T. (1986). "Selected Guidelines for Notification to Survey Participants", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1-5.

Jabine, T. (1987a). "Remarks at Panel Discussion", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Ottawa, (Ontario).

Jabine, T. (1987b). "Statistical Uses of Administrative Records in the United States: Some Recent Developments", document présenté à la Réunion annuelle de la Société statistique du Canada, Québec.

Jabine, T., et Scheuren, F. (1985). "Goals for Statistical Uses of Administrative Records: The Next Ten Years", *Journal of Business and Economic Statistics*, Vol. 3, no 4, 380-391.

Jabine, T., et Scheuren, F. (1986). "Record Linkages for Statistical Purposes: Methodological Issues", *Journal of Official Statistics*, Vol. 2, no 3, 255-277.

Johnson, N., Rogot, E., Glover, C., Sorlie, P., et McMillen, M. (1985). "General Mortality Among Selected Census Bureau Sample Cohorts 1979-1981", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 428-433.

Kanter, R. (1983). *The Change Masters*, Simon and Schuster.

Kestenbaum, B. (1986). "An Accounting of the 1919 Birth Cohort", *Proceedings of the American Statistical Association, Social Statistics Section*, 397-400.

Kestenbaum, B. (1985). "The Measurement of Early Retirement", *Journal of the American Statistical Association*, Vol. 80, no 389, 38-45.

Kestenbaum, B., et Diez, G. (1981). "Geographic Mobility of Older Workers", *Proceedings of the American Statistical Association, Social Statistics Section.*

Kilss, B., Scheuren, F., et Buckler, W. (1980). "Goals and Plans for a Linked Administrative Statistical Sample", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 450-455.

Kirkendall, N. (1985). "Weights in Computer Matching: Applications and An Information Theoretic Point of View", *Record Linkage Techniques – 1985*, Internal Revenue Service, 189-197.

McMillan, A., Gornick, M., Howell, E., Prihoda, R., Rabey, L., Russell, D., et Lubitz, J. (1987). "The Dually Entitled Medicare and Medicaid Elderly: Impact of Nursing Home Care on Costs", document préliminaire, Health Care Financing Administration.

Moore, J., et Marquis, K. (1987). "Utilisation des données de dossiers administratifs pour l'évaluation de la qualité des estimations d'enquêtes", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Statistique Canada, Ottawa, (Ontario).

National Association of Business Economists (1987). "Report of the Statistics Committee".

National Research Council (1984). "Planning the 1990 Census: Priorities for Research and Testing", Interim Report of the Panel on Decennial Census Methodology, Committee on National Statistics.

National Research Council (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, Report of the Panel on Decennial Census Methodology, Committee on National Statistics, Washington, D.C.: National Academy Press.

Newcombe, H. B. (1967). "Record Linkage: The Design of Efficient Systems for Linking Records into Individual and Family Histories", *American Journal of Human Genetics*, University of Chicago Press, Vol. 19, no 3, partie 1 (mai).

Newcombe, H. B., Kennedy, J. M., Axford, S. J., et James, A. P. (1959). "Automatic Linkage of Vital Records", *Science*, Vol. 130, no 3381, 954-959.

New York Times (1987). "U. S. Rejects Pleas to Adjust 1990 Census for Undercount", 31 octobre.

Norwood, J. (1985). "Comment: Administrative Statistics: A BLS Perspective", Journal of Business and Economic Statistics, Vol 3, no 4, 398-400.

Paass, G. (1985). "Disclosure Risk and Disclosure Avoidance for Microdata", document présenté à l'International Association for Social Service Information and Technology.

Passel, J., et Word, D. (1980). "Constructing the List of Spanish Surnames for the 1980 Census: an Application of Bayes Theorem", document présenté à la réunion annuelle de la Population Association of America, Denver, Colorado.

Passel, J., et Word, D. (1987). "Problems in Analyzing Race and Hispanic Origin Data from the 1980 Census: Solutions Based on Constructing Consistent Populations from Micro-level Data", document présenté à la réunion annuelle de la Population Association of America, Chicago, Illinois.

Patterson, J., et Bilgrad, R. (1985). "The National Death Index Experience: 1981-1985", dans *Record Linkage Techniques — 1985*, Internal Revenue Service, Département du Trésor des États-Unis, 245-254.

Pearson, R. (1986). "Researchers' Access to U. S. Federal Statistics", *Items*, Vol. 41, nos 1/2, 6-11.

Peters, T. J., et Waterman, R. H. Jr. (1982). *In Search of Excellence*, Warner Books, New York, New York.

Revenu Canada, Impôt (1986). *Multi-National Tax Modelling Symposium Proceedings* du symposium canado-américain sur la construction de modèles fiscaux tenu du 17 au 19 septembre 1985 au Mont Ste-Marie, Lac Ste-Marie (Québec).

Riley, G., Lubitz, J. Prohoda, R., et Robey, E. (1987). "The Use and Cost of Medicare Services by Cause of Death", *Inquiry*, Vol. 24, 233-244.

Rogot, E., Schwartz, S., O'Connor, K., et Olsen, C. (1983). "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 319-324.

Rogot, E., Sorlie, P., Johnson, N., Glover, C., et Makuc, D. (1985). "Mortality by Cause of Death Among Selected Census Bureau Sample Cohorts for 1979-1981", *Proceedings of the American Statistical Association, Sction on Survey Research Methods*, 444-449.

Scheuren, F. (1985). "Methodologic Issues in Linkage of Multiple Data Bases", *Record Linkage Techniques — 1985*, Internal Revenue Service, 155-178.

Scheuren, F. (1986). "Record Linkages for Statistical Purposes in the United States", *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, Ottawa, (Ontario). 198-210.

Scheuren, F. (1987). "Notes on IRS Quality Improvement Process", communication présentée à l'OMB Meeting of Departmental Productivity Officers, 4 novembre 1987.

Strudler, M., Oh, H. L., et Scheuren, F. (1986). "Protection of Taxpayer Confidentiality with Respect to the Tax Model", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 375-381.

Tippett, J. (1987). "Housing Data: The Quality of Selected Items", *Proceedings of the Census Bureau Third Annual Research Conference*, Bureau of the Census, 417-436.

U. S. Department of Health, Education and Welfare (1978). *Policy Analysis with Social Security Research Files*, compte rendu d'un atelier tenu en mars 1978 à Williamsburg, Virginie, Social Security Administration, Research Report No. 52, HEW Publication No. (SSA)79-11808.

Waite, C. A. (1985). "Comment: The Future of Administrative Records in the Economic Programs of the Census Bureau", *Journal of Business and Economic Statistics*, Vol. 3, no 4, 400-401.

Wall Street Journal (1987). "Backers of an Adjusted Census Won't Take No for an Answer", 3 novembre.

Washington Post (1987). "Australians in Uproar Over ID Cards", 29 octobre.

Wedick, J. L. Jr. (1986). "Electronic Filing at the IRS: The Goal is Global", *Journal of Accountancy*, 110-116.

Wilk, M. (1985). "Statisticiens et statisticiens", *Techniques d'enquête*, Vol. 11, no 2, 101-107.

Wilson, R. (1983). "Postal ZIP Code Area Statistics from Internal revenue Records", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 367-371.

Wolfson, M., Gribble, S., Bordt, M., Murphy, B., et Rowe, G. (1987). "La base de données de simulation de politique sociale: un exemple d'intégration de données d'enquêtes et de données administratives", communication présentée au Symposium international sur les utilisations statistiques des données administratives, Ottawa, (Ontario).

SESSION II: COMMUNICATIONS SOLLICITÉES

MÉTHODOLOGIE DU COUPLAGE DES ENREGISTREMENTS

Président: J.N.K. Rao, Carleton University

Statistique Canada Symposium
sur les utilisations statistiques
des données administratives
novembre 1987

UTILISATION DE GRANDES BASES DE DONNÉES POUR LA RECHERCHE EN CHIRURGIE

LESLIE L. ROOS et NORALOU P. ROOS[1]

RÉSUMÉ

Cette article a pour but d'analyser les caractéristiques courantes des bases de données sur les soins de santé en mettant l'accent sur les rapports entre ces caractéristiques et les questions que peuvent se poser les spécialistes. Bien qu'il existe des données sur les sorties d'hôpital pour la très grande majorité des provinces canadiennes et de nombreux endroits aux États-Unis, peu de séries de données contiennent des renseignements qui seraient d'une importance primordiale pour les plans de recherche. On pourrait vouloir connaître, par exemple, l'observation d'une population entière ou savoir s'il existe un numéro d'identification unique (ou une combinaison d'identificateurs) ou encore un fichier d'inscription qui indiquerait le début et la fin de l'observation. Dans cet article, nous définissons trois groupes de données selon le niveau de détail; nous voyons à quelles sortes de plans de recherche peuvent servir ces trois groupes de données.

Les données de la troisième catégorie (relevés des sorties d'hôpital seulement) peuvent servir à des analyses d'utilisation dans les divers marchés hospitaliers de même qu'à des études sur la durée de séjour et la mortalité hospitalière. Les données de la seconde catégorie (relevés des sorties d'hôpital plus identificateurs individuels cohérents) peuvent servir à des études à court terme sur les cas de réadmission et les complications après l'opération chirurgicale. Des études de ce genre sur l'assurance de la qualité et la limitation des coûts peuvent être réalisées dans des délais raisonnables pour alimenter les établissements de santé de rétro-information. Les données de la première catégorie (relevés des sorties d'hôpital, identificateurs individuels cohérents et fichiers d'inscription) permettent de faire des analyses longitudinales de qualité supérieure; on peut alors établir des comparaisons entre divers traitements et divers hôpitaux. Grâce aux données de la première catégorie, on peut aussi identifier les cas problèmes et faire des analyses portant sur une même personne dans les divers marchés hospitaliers. D'autres applications notables des bases de données existantes -particulièrement des données de la première catégorie - sont l'étude de divers aspects du comportement des médecins et la production d'information pour les analyses de décisions cliniques.

[1] Départements des sciences administratives et des sciences de la santé communautaire, Facultés d'administration et de médecine, Université du Manitoba, Canada, R3T 2N2.

Les données de la première catégorie sont extrêmement utiles pour faire ressortir les avantages et les faiblesses des séries de données moins complètes. Dans cet article, nous nous intéressons à la casuistique et mettons en contraste les données fondées sur les hôpitaux et celles fondées sur la population en signalant les questions qui font l'objet de la recherche à l'heure actuelle. En montrant dans quelles conditions on peut utiliser des renseignements moins complets sans pour autant sacrifier la précision de l'analyse, cet article incitera peut-être un plus grand nombre de chercheurs d'autres disciplines à produire des études utiles sur la question.

Nous abordons ensuite deux questions pratiques auxquelles se heurtent les spécialistes qui cherchent à utiliser les données administratives: l'applicabilité du jumelage d'enregistrements dans le but de produire des données plus détaillées et les caractéristiques des logiciels qui conviennent pour l'exploitation des bases de données existantes. Enfin, nous analysons des moyens d'améliorer les résultats d'opérations chirurgicales.

1. UTILISATION DE GRANDES BASES DE DONNÉES POUR LA RECHERCHE EN CHIRURGIE

On trouve à divers endroits des bases de données sur les soins de santé; ces bases, qui n'ont pas nécessairement la même qualité ni le même champ d'observation, sont exploitées par des groupes de recherche, des hôpitaux, des assureurs et des organismes gouvernementaux. Les données produites couramment par les régimes d'assurance-maladie d'Amérique du Nord, d'Europe, d'Australie et de Nouvelle Zélande présentent un intérêt particulier. En effet, ces régimes produisent un grand nombre de relevés de sortie originant des hôpitaux, et les scientifiques sont de plus en plus conscients des possibilités qu'offrent ces régimes au point de vue de la recherche. Ainsi, le National Institute of Medicine (par l'intermédiaire du Committee for Evaluating Medical Technologies in Clinical Use, 1985) a recommandé de renforcer les méthodes d'évaluation de la pratique médicale fondées sur les bases de données existantes. Feinleib [16] a souligné le nombre d'études qu'il est possible de réaliser lorsqu'on dispose de bases de données complètes de qualité supérieure. Les scientifiques, les administrateurs et les praticiens évoluent dans un milieu de mieux en mieux informé; si seulement ils savaient quoi faire des données dont ils disposent! Dans cet article, nous allons tâcher de voir comment on devrait construire les bases de données administratives de manière à faciliter la recherche et quelle est la meilleure façon d'analyser les données existantes.

Quelles seraient les caractéristiques d'une base de données idéale pour la recherche en chirurgie? Premièrement, l'observation d'une population entière permettrait d'analyser l'utilisation d'un point de vue épidémiologique en associant les services aux prestataires selon le lieu de résidence de ceux-ci, sans égard à l'endroit où les services ont été fournis. On pourrait ensuite comparer les quantités de ressources utilisées (par catégorie) pour diverses populations, les données de ce genre pouvant être corrigées en fonction de l'âge, du sexe et d'autres caractéristiques pour faciliter la comparaison.

Dans cette base de données idéale, chaque personne devrait être identifiée par un numéro unique ou une combinaison d'identificateurs. Dans ces conditions, il est facile de constituer un dossier d'utilisation pour chaque personne, sans égard à l'endroit où ont été dispensés les soins. Cette base de données devrait enregistrer tous les rapports qu'a chaque personne avec le système de santé (l'identificateur propre permettant de suivre le cheminement d'une personne en particulier). Les rapports que pourrait enregistrer cette

base de données seraient tous les soins hospitaliers (cliniques et externes), les services reçus dans des centres chirurgicaux autonomes ou des cabinets de médecin, l'admission dans une maison de soins, les soins reçus à domicile et l'utilisation de médicaments prescrits. Par conséquent, si une personne subit une intervention chirurgicale dans un établissement et qu'elle est réadmise par la suite dans un autre établissement, les deux "événements" seront consignés dans le système.

Enfin, il est très utile de savoir à quels moments commence et prend fin l'observation (statistique) d'une personne et pourquoi il en est ainsi. Un fichier d'inscriptions est nécessaire en ce sens; si la base de données n'indique pas qu'une personne quelconque a eu recours au système de santé, on peut se demander si cette personne habite le territoire visé et si effectivement elle n'a pas eu recours au système de santé ou encore si elle a quitté le territoire ou est décédée?

Une base de données de ce genre faciliterait non seulement les études comparatives d'utilisation dans divers marchés médicaux mais aussi les études longitudinales d'utilisation selon des groupes d'individus (définis de plusieurs façons différentes). Ce genre de suivi (discret) présente un certain nombre d'avantages par rapport au suivi effectué par des méthodes qui comportent plus d'opérations manuelles [40]. Il peut faire avancer considérablement la recherche visant à expliquer les résultats de diverses opérations chirurgicales. Les analyses longitudinales sont également indispensables pour suivre l'évolution de certaines affections qui conduisent ou non à l'hospitalisation.

Les données administratives conviennent particulièrement bien à l'étude des principales conséquences d'une dysfonction: décès, admission dans une maison de soins et hospitalisation. Il existe aussi des données sur d'autres conséquences de la dysfonction qui ont exigé le recours au système de santé; dans beaucoup de cas, on peut également évaluer la proportion de personnes qui n'ont jamais besoin de soins de santé.

Comme de nombreuses bases de données sont mises à jour pour des besoins administratifs, le chercheur peut vraisemblablement analyser ces bases, une fois qu'il en a obtenu l'accès, à un coût marginal relativement faible. Les coûts ayant été supportés principalement par l'organisme qui se sert des données à des fins administratives, le chercheur peut se livrer sans contrainte à une série d'analyses. On peut étendre la période étudiée (période où l'on subit des traitements ou d'autres événements) soit en ajoutant des données d'années plus récentes ou en remontant plus loin dans le temps. Cela facilite le suivi à long terme des personnes de même que les études d'événements rares; de façon plus générale, l'élargissement de la période étudiée accroît le nombre de cas et, partant, le pouvoir statistique de l'analyse.

Faiblesses de la base de données

Néanmoins, même une base de données "idéale" n'est pas exempte de faiblesses. En premier lieu, il est nécessaire d'évaluer la qualité des renseignements qui y ont été enregistrés. Plusieurs provinces canadiennes possèdent des bases de données qui ont beaucoup servi à la recherche [3,75]. Au Manitoba par exemple, les interventions chirurgicales pratiquées dans les hôpitaux et les actes médicaux facturables de moindre importance (ne serait-ce que des actes secondaires comme les prélèvements vaginaux) semblent être enregistrés avec soin dans le système des réclamations [42,47]. L'exactitude des données sur les diagnostics dépend aussi bien des médecins que des commis qui enregistrent les diagnostics. Le régime américain Medicare semble enregistrer les actes médicaux avec assez de précision, surtout si l'on ne tient pas compte de l'ordre dans lequel ces actes ont été exécutés. Depuis la mise sur pied du Prospective Payment System (PPS), il y a sans doute eu une amélioration de la qualité des données sur les soins médicaux mais les données sur les diagnostics ne sont peut-être pas aussi exactes que celles consignées dans les fichiers du Manitoba [12,74]. Une autre lacune du régime Medicare est qu'il ne recueille pas de données sur les soins externes fournis par les hôpitaux.

Les diagnostics consignés dans les registres des hôpitaux sont vraisemblablement plus précis que ceux qui apparaissent sur les relevés d'honoraires produits par les médecins à la suite d'une visite. Dans le réseau hospitalier du Manitoba, les diagnostics sont enregistrés avec un degré acceptable de précision et de spécificité, ce qui dénote la compétence professionnelle des auxiliaires aux archives médicales. Une comparaison de diagnostics inscrits dans les registres d'hôpitaux et de diagnostics figurant sur des relevés d'honoraires de médecins a révélé une correspondance de 95% pour les cas de maladie de la visicule biliaire et une correspondance de 89 à 92% (diagnostics similaires) pour les cas d'infarctus du myocarde aigu [41,42]. À l'heure actuelle, le système du Manitoba peut enregistrer jusqu'à 16 diagnostics pour chaque personne hospitalisée au moyen du système normalisé international ICD-9-CM. Les diagnostics concernant les patients ambulatoires sont utiles mais à un niveau beaucoup plus général. Une méthode qui s'est avérée fructueuse au Manitoba consistait à grouper les diagnostics figurant sur les relevés d'honoraires des médecins (par exemple les relevés produits à la suite de consultations pour des problèmes gynécologiques dans le cadre d'une étude sur des femmes subissant une hystérectomie, et ceux produits à la suite de consultations pour des problèmes de vésicule biliaire ou des douleurs abdominales dans le cadre d'une étude sur les consultations avant et après l'intervention chirurgicale pertinente) au lieu d'essayer de faire des distinctions poussées [11-12].

Les caractéristiques du système de codage utilisé soulèvent un autre problème. En effet, le système ICD-9-CM, dont l'usage est très répandu, ne distingue pas les interventions pratiquées sur le côté gauche de celles pratiquées sur le côté droit. Cela pose des difficultés lorsqu'il s'agit d'analyser les résultats de la chirurgie orthopédique; une seconde opération à la hanche ou au genou peut signifier que la personne a été réopérée ou qu'elle a subi une intervention pour l'autre hanche ou l'autre genou [53].

Quel que soit la quantité de renseignements que renferme une base de données, il peut arriver que l'on n'y trouve pas l'information voulue. On peut (ou non) obtenir des renseignements additionnels d'autres sources qui permettent l'appariement avec une base de données existante. En particulier, les bases de données administratives contiennent rarement les résultats de certains tests ou de radiographies qui ne sont pas facturables; les résultats de tests sont rarement consignés dans les bases de données. De même, les données sur les traitements médicaux (par exemple les médicaments utilisés) sont pratiquement inexistantes, ce qui complique la comparaison des solutions médicales et chirurgicales dans le traitement de nombreuses affections.

Enfin, le système qui produit les données est en relation avec la personne qui a besoin de soins et non avec le chercheur. Par conséquent, si une personne est malade et qu'elle ne recourt pas au système de santé, la base de données ne contiendra aucun enregistrement relatif à ce cas. La mesure dans laquelle cette lacune influera sur l'évaluation de l'état de la santé nationale dépendra de certaines caractéristiques du système (comme l'étendue du régime d'assurance) et des caractéristiques individuelles (tendance à recourir au système de santé). Étant donné l'universalité du régime, il y a un nombre relativement restreint de personnes malades qui n'ont aucun rapport avec le système de santé [30]. De fait, en quatre ans, 91% des adultes du Manitoba voient un médecin au moins une fois.

Les données sur les sorties d'hôpital se retrouvent vraisemblablement dans beaucoup d'administrations publiques. Les données ci-dessous semblent figurer couramment sur les formules de règlement des frais d'hospitalisation dans la très grande majorité des provinces canadiennes et à beaucoup d'endroits aux États-Unis.

Genre de données:

Date de naissance
Sexe
Lieu de résidence
Numéro d'identification (personne ou famille)

Autres données pour l'analyse:

Diagnostics de sortie (plusieurs)
Interventions pratiquées à l'hôpital (plusieurs)
Hôpital
Date d'admission
Date de sortie
Code de sortie (décès, transfert à un autre hôpital, retour à la maison, etc.)

Renseignements secondaires:

Numéro d'identification du médecin qui a autorisé l'admission
Numéro d'identification du médecin qui a pratiqué l'intervention

Comme nous l'avons souligné plus haut, la base de données peut ne pas contenir ces renseignements importants:

- Observation d'une population entière
- Numéro d'identification **unique** (ou combinaison d'identificateurs)
- Fichier d'inscription indiquant le début et la fin de la couverture

2. PLANS DE RECHERCHE

Compte tenu de la diversité des questions sur lesquelles s'interrogent les spécialistes et des caractéristiques des données recherchées, le tableau 1 expose le genre de renseignements nécessaires à diverses études sur la limitation des coûts, la qualité des soins et les résultats de traitement. Les plans de recherche vont du relativement simple au relativement complexe selon le genre de données qu'ils exigent. Dans ce tableau, les catégories de données sont définies selon l'usage qu'on peut en faire. Ainsi, la catégorie de données la plus complète (première catégorie) permet de faire toutes les analyses que rendent possibles les données des deux autres catégories, moins complètes, et aussi des travaux de recherche qui ne peuvent s'appuyer sur des séries de données moins complètes. Plus les exigences d'un plan de recherche sont minimes, plus grande sera la quantité de renseignements recueillis par les unités politiques (états, provinces, etc.) qui permettent ce genre de recherche.

Plans de recherche aux exigences élémentaires (données de la troisième catégorie)

Certains plans de recherche ont des exigences relativement limitées en ce qui a trait aux données. En l'absence d'un identificateur individuel cohérent et d'un fichier d'inscription, trois genres d'études transversales sont couramment effectuées. Ces études sont fondées sur les renseignements contenus dans les relevés de sortie des hôpitaux et ne visent pas à décrire le cheminement des personnes dans le système de santé. Elles impliquent normalement une grande quantité de données, ce qui permet en général d'établir des comparaisons entre de nombreux hôpitaux. Comme les études de mortalité et de durée de séjour décrites ci-dessous sont plus souvent fondées sur les établissements hospitaliers que sur la population, les totaux de contrôle pour la casuistique sont particulièrement importants dans les circonstances.

Tableau 1

Exigences relatives aux données et genres d'études nécessitant des données d'hôpitaux

Exigences relatives aux données	Genres d'études
Simples - 3ième catégorie Relevés de sortie des hôpitaux	Mortalité hospitalière (comparaison volume d'activité chirurgicale - résultats, analyse des hôpitaux) Durée de séjour Analyse par région
Intermédiaires - 2ième catégorie Relevés de sortie des hôpitau et identificateurs individuels cohérents	Analyse longitudinale Cas de réadmission à court terme; Comparaison volume d'activité chirurgicale - résultats, Analyse des hôpitaux, Assurance de la qualité et limitation des coûts
Élevées - 1ère catégorie Relevés de sortie des hôpitaux, identificateurs individuels cohérents et fichier d'inscription	Analyse longitudinale de première qualité, Analyses de résultats à court terme et à long terme; Identification des cas problèmes, Comparaison volume d'activité chirurgicale - résultats, Analyse des hôpitaux, Choix du traitement, Analyse par région par personne

Luft et ses collaborateurs [27-28,59], ont utilisé des données fournies par les hôpitaux qui cotisent à la Commission on Professional and Hospital Activities (CPHA) pour montrer que le volume d'activité chirurgicale dans les hôpitaux est une variable déterminante pour le taux de survie des patients qui ont subi une intervention chirurgicale complexe. L'étude de Luft révèle que pour de nombreux types d'intervention, le taux de survie est généralement plus élevé dans les hôpitaux où l'on pratique un plus grand nombre d'interventions. Bien que Sloan et coll [61] se servent de données fournies par 521 hôpitaux de la CPHA pour la période 1972-1981 pour soutenir qu'il n'existe pas de base statistique suffisante pour définir des normes minimales concernant le volume d'activité chirurgicale dans les hôpitaux, la majeure partie des renseignements dont nous disposons confirment le rapport qui existe entre cette variable et le taux de survie [26]. Ces analyses transversales se limitent le plus souvent à l'étude de la mortalité durant le séjour à l'hôpital qui accompagne l'intervention chirurgicale. De meilleures données permettraient de réaliser un plus grand nombre d'études sur le taux de survie à plus long terme, mais les recherches actuelles permettent de croire que pour un bon nombre d'actes chirurgicaux courants, les études qui se limitent à la mortalité hospitalière enregistrent une proportion suffisamment élevée des décès qui surviennent dans les trois mois suivant l'intervention chirurgicale pour permettre une analyse valable.

Un autre genre d'études utilise les renseignements contenus dans les relevés de sortie pour classer les patients selon le diagnostique. On suppose alors que les patients d'un groupe donné reçoivent tous les mêmes services et les mêmes soins hospitaliers [17]. On peut comparer les hôpitaux en fonction de chaque groupe pour savoir quels groupes de patients passent plus de temps à l'hôpital; on peut calculer un indice global de durée de séjour rajusté en fonction de la casuistique pour chaque hôpital. Aux États-Unis, ce genre

d'études est de plus en plus utilisé pour le suivi de l'utilisation des services hospitaliers en raison des lois prescrivant l'intégration des groupes formés selon le diagnostic dans le Prospective Payment System du régime Medicare. C'est une pratique courante chez les directeurs d'hôpitaux d'utiliser la durée de séjour comme substitut des coûts [63]. L'application de la méthode décrite ci-dessus a soulevé une vive controverse; on s'applique présentement à améliorer cette méthode adaptée à la casuistique [10,20,64].

Les analyses par région exigent elles aussi relativement peu de données, mais elles nécessitent une base démographique. Un marché hospitalier est défini comme une sous-unité géographique dont la majorité des habitants utilisent les services d'un hôpital plutôt qu'un autre ("règle de la majorité") [6]. La base démographique sert de dénominateur (c'est-à-dire que toutes les personnes qui demeurent dans une région sont dénombrées) et permet de rajuster les taux d'utilisation des services hospitaliers selon l'âge et le sexe, ce qui élimine une des principales sources de variation de ces taux liées aux patients [55]. On relève tous les cas d'utilisation des services de santé dans la région, quel que soit l'endroit de la prestation.

La combinaison de l'analyse de la durée de séjour et de l'analyse par région peut mettre en lumière la façon la plus efficace de contrôler l'utilisation des services de santé. Les traitements varieront vraisemblablement selon que la méthode DRG (qui met l'accent sur la durée de séjour) ou la méthode "des régions" (qui met l'accent sur le taux d'admission) explique le mieux la variation du nombre total de jours passés à l'hôpital [73]. Les analyses provenant du Manitoba font très nettement ressortir la différence. Tous les pontages coronariens pratiqués au Manitoba se font dans deux hôpitaux d'enseignement de Winnipeg. Les médecins qui pratiquent dans l'ouest du Manitoba envoient à Winnipeg relativement peu de patients devant subir une coronarographie puis un pontage. Le tableau 2 montre jusqu'à quel point les différences de taux d'utilisation entre les régions dépendent plus de la variation du nombre de pontages par habitant que de la variation de la durée de séjour pour les personnes qui subissent un pontage. La colonne à l'extrême droite donne le nombre total de jours d'utilisation des services de santé pour 10,000 adultes pour les sept régions administratives du Manitoba. Les chiffres de cette colonne sont le produit du taux d'admission rajusté directement (pour 10,000 adultes) par la durée de séjour moyenne pour les personnes qui subissent un pontage.

Tableau 2

Taux d'admission, durée de séjour et nombre total de jours d'hospitalisation par 10,000 adultes pour les personnes ayant subi un pontage coronarien, sept régions administratives du Manitoba (1979-1984)

Région	Taux d'admission, 10,000 adultes (corrigé directement)	Durée de séjour moyenne	Nombre total de jours par 10,000 adultes
Centrale	2.93	22.32	65.42
Est	3.73	21.47	80.22
Interlake	4.06	18.80	76.36
Nord	5.15	20.67	106.38
Parkland	3.28	19.47	63.83
Quest	2.40	20.47	49.05
Winnipeg	5.04	20.64	104.01

Les taux d'admission annuels moyens, la durée de séjour et le nombre total de jours par 10,000 adultes pour les pontages coronariens sont présentés dans ce tableau.

Du point de vue de la gestion de programmes, les deux mesures (réduction du nombre d'admissions et raccourcissement de la durée de séjour) peuvent être efficaces pour freiner les coûts. Elles ne sont d'ailleurs pas étrangères à la diminution récente du taux d'accroissement des coûts dans les hôpitaux américains [57]. Les analyses que nous venons de décrire ont un rôle particulièrement important puisque les taux d'admission, tant pour les cas de traitement thérapeutique pour que les cas de traitement chirurgical, varient selon les marchés hospitaliers. On observe même des différences notables à cet égard entre Boston et New Haven, où se trouvent d'importants centres d'enseignement affiliés à Harvard et à Yale respectivement [72]. Comme les fortes différences de taux de morbidité ne peuvent expliquer de façon satisfaisante les différences de taux d'hospitalisation rapportées dans les études sur les régions [39,55,71], on s'intéresse de plus en plus à l'origine de cette divergence [73]. La rétro-information a parfois permis de réduire des taux très élevés [70]. Ce genre d'information est également essentiel pour les régimes de paiement qui reposent sur la rémunération forfaitaire par personne et qui, de fait, sont fortement préconisés par le programme Medicare aux États-Unis [23].

Plans de recherche aux exigences moyennes (données de la seconde catégorie)

Les plans à exigences moyennes sont ceux qui conviennent quand la base de données renferme des identificateurs individuels cohérents mais qu'il est difficile d'avoir accès à des fichiers d'inscription. On peut classer les réclamations d'hôpitaux par date et par numéro d'identification afin de créer un dossier d'hospitalisation pour chaque personne. Comme il n'y a pas de fichier d'inscription, on ne peut connaître la période d'observation s'appliquant à une personne. Il y a donc lieu de s'intéresser davantage aux résultats à court terme au point de vue tant de la morbidité que de la mortalité. Les résultats à court terme pourraient être la réadmission à cause de complications particulières, la réadmission après une période déterminée (48 heures, 7 jours ou 6 semaines après la sortie de l'hôpital) ou le décès (lorsqu'il est indiqué sur un rapport ultérieur de l'hôpital). Dans beaucoup de cas, si la période donnée est relativement courte (jusqu'à 12 mois après l'intervention chirurgicale), on peut omettre l'absence de suivi (due à la migration ou, dans le cas des États-Unis, à un changement d'assureur) sans que cela fausse sensiblement les résultats. Aux États-Unis, cela semble se vérifier pour les bénéficiaires de Medicare mais non pour ceux de Medicaid [24,71]. Les plans de recherche qui ne sont pas tributaires d'un fichier d'inscription permettent d'analyser les résultats d'actes médicaux plus efficacement et plus rapidement que les plans qui prévoient le couplage de fichiers d'utilisation et d'inscription. Il peut s'agir là d'un avantage appréciable pour ce qui a trait à la transmission rapide de rétro-information aux comités chargés de surveiller la qualité des soins.

À des fins de contrôle, il serait utile de mettre au point un algorithme informatique qui ferait intervenir la variable "période écoulée depuis l'intervention chirurgicale" et la variable "diagnostic au moment de la réadmission" pour identifier les cas de réadmission attribuables à des complications postopératoires. Des études récentes sur des cas d'hystérectomie, de cholécystectomie et de prostatectomie ont montré qu'il était possible de mettre au point de tels algorithmes; les résultats obtenus avec ces algorithmes se rapprochaient sensiblement des conclusions de groupes de médecins [38]. Les analyses de résultats à court terme peuvent comporter l'impression d'un dossier complet pour permettre un examen plus poussé; un listage sur lequel figure le nom de personnes réadmises à l'hôpital par suite de certains diagnostics ou dans une période déterminée suivant le jour de l'intervention chirurgicale peut alléger sensiblement le fardeau administratif des comités d'assurance de la qualité. Ainsi, on peut tout d'abord recourir à l'ordinateur; ensuite, on peut consulter les dossiers des hôpitaux pour une analyse plus approfondie.

Les données fondées sur la population (par comparaison à celles fondées sur les hôpitaux) contribuent à améliorer ces plans intermédiaires. Si une personne a un numéro

d'identification qui lui a été attribué par un hôpital particulier, il ne sera pas possible de savoir si cette personne a été réadmise dans un autre hôpital. En revanche, si la même personne se voit attribuer un numéro d'identification en vertu d'un régime d'assurance qui s'applique à un territoire donné (comme certains régimes provinciaux au Canada et le programme Medicare aux États-Unis), il sera possible de recueillir des données sur cette personne auprès de tous les établissements où elle aura reçu des soins et qui seront reconnus par le régime. Cette formule est indispensable pour enregistrer la réadmission de personnes qui ont subi une intervention chirurgicale dans un hôpital urbain et qui vivent en région rurale; le plus souvent, ces personnes seront réadmises dans de petits hopitaux situés dans leur région. Au Manitoba par exemple, 45% des personnes réadmises à l'hôpital par suite de complications découlant d'une cholécystectomie et qui vivaient en région rurale ont été admises dans un autre hôpital que celui où elles avaient été opérées [38]. Ainsi, quand il n'y a pas de données pour l'ensemble d'un territoire, les analyses d'enregistrements sous-estiment systématiquement les taux de complication postopératoire.

Plans de recherche aux exigences élevées (données de la première catégorie)

Si nous disposons à la fois d'identificateurs individuels cohérents et d'un fichier d'inscription, il est possible d'estimer l'absence de suivi. Nous pouvons faire des analyses longitudinales qui répondent à des critères souhaitables pour une étude de première qualité sur les causes et les effets [21]. En exécutant un suivi pour chaque personne, il est plus facile de relever les nouveaux cas de maladie dans une population. Cet exercice est essentiel si l'on veut obtenir un groupe relativement homogène; on peut alors faire la distinction entre les nouveaux cas de maladie, d'une part, et une seconde opération chirurgicale ou la réapparition d'une affection, d'autre part. Des renseignements de ce genre peuvent améliorer sensiblement les études de résultats: comme un second pontage peut comporter plus de risques que le premier, il est très utile d'analyser les deux actes séparément. De même, on distinguera les cas de chirurgie de remplacement valvulaire après pontage des cas de chirurgie de remplacement valvulaire non précédée d'interventions majeures.

Dans un régime national d'assurance-maladie, les études de résultats à court terme qui sont fondées sur un fichier d'inscription seront à peine plus précises que celles qui ne reposent sur aucun fichier [45]. Si le régime n'est pas universel (du moins à l'intérieur d'un groupe d'âge donné, par exemple les personnes âgées), il faudra se reporter à un fichier d'inscription [24]. Le nombre de personnes qui quittent un régime d'assurance donné ou qui y deviennent admissibles peut être suffisamment élevé pour modifier considérablement les résultats.

Les résultats à long terme, comme une nouvelle opération, l'admission dans une maison de soins ou un décès, sont des sujets particulièrement indiqués pour des études qui utilisent des bases de données complètes. Comme il a déjà été souligné [56], les données tirées d'un régime d'assurance qui offre une protection complète à une population, indépendamment du lieu de prestation des services médicaux, peuvent révéler des problèmes qui seraient passés inaperçus dans une étude fondée sur les données d'un seul hôpital. Au fil des années, il devient de plus en plus probable que les personnes auront reçu des soins dans plus d'un hôpital; fonder une étude sur les données d'un seul hôpital risque donc d'accroître les taux d'erreur.

Les données tirées d'un réseau (et non d'un seul hôpital) facilitent grandement les études sur l'utilité et l'efficacité. Les études sur l'utilité, c'est-à-dire sur l'utilité en situation dite "optimale" (en règle générale, un hôpital d'enseignement), se retrouvent surtout dans des évaluations de technologies [5]. En revanche, on connaît peu de choses sur l'utilité de nombreux genres d'interventions [34]. L'absence relative d'études sérieuses sur les résultats semble confirmer l'indécision dans laquelle se trouvent les médecins

lorsqu'il s'agit de choisir le traitement approprié; c'est ce qui explique peut-être une bonne partie des variations observées entre les régions [69]. En outre, les pratiques des hôpitaux communautaires et les résultats de soins médicaux peuvent être différents de ceux dont font état les scientifiques dans les centres de recherches. Quant aux études d'efficacité -- -études exposant les résultats d'actes médicaux ou chirurgicaux à l'aide d'échantillons représentatifs d'hôpitaux et de médecins --, très peu ont été réalisées.

Des études récentes sur la prostatectomie peuvent servir de modèle pour les analyses longitudinales d'une intervention chirurgicale courante [54,74]. Des données du Maine et du Manitoba ont été combinées pour faire le suivi de personnes ayant subi une prostatectomie; la période de suivi pouvait aller jusqu'à huit ans. Cette combinaison de données a permis d'obtenir un plus grand nombre de cas, ce qui, en soi, peut être important parce que certains résultats, notamment les décès posopératoires, sont assez rares et que le nombre d'établissements est relativement faible. On a évalué à la fois la morbidité (par l'intermédiaire des révisions, des réadmissions dûes à des complications et ainsi de suite) et la mortalité postopératoire. De façon générale, les résultats défavorables (décès et complications non mortelles) étaient plus fréquents que ne l'avaient indiqué les ouvrages spécialisés. La base de données nous a permis de recenser des personnes qui ont été admises à un hôpital différent de celui où elles avaient été opérées; ce dénombrement additionnel de même que la période de suivi plus longue ont sans doute contribué à l'observation d'un plus grand nombre de résultats défavorables dans cette étude par rapport aux autres études (dont la plupart se rapportaient à des hôpitaux d'enseignement). La forte variabilité des résultats entre les divers hôpitaux a confirmé la nécessité de faire des études d'efficacité. Les risques relatifs corrigés de mortalité dans les trois mois suivant l'intervention chirurgicale variaient de 0.48 à 4.79 selon les hôpitaux. Dans un hôpital, le taux de mortalité était moitié moins élevé que dans l'hôpital de référence; dans un autre établissement, le taux était presque cinq fois plus élevé que le taux observé pour l'hôpital de référence. Pour un sous-groupe en particulier (hommes demeurant dans une maison de santé), le taux de mortalité était particulièrement élevé.

Les analyses longitudinales peuvent parfois servir à comparer les résultats de deux genres d'intervention ou les résultats de traitements thérapeutiques et de traitements chirurgicaux. Les études sur la prostatectomie, qui ont fait l'objet de discussions ci-dessus, ont permis d'observer des écarts statistiquement significatifs entre les résultats d'interventions sanglantes et les résultats d'interventions transurétrales [44]. Ces écarts sont d'ailleurs analysés plus en profondeur dans des études en cours. Par ailleurs, les données administratives se sont avérées particulièrement utiles pour regrouper un nombre suffisant de cas d'endocardite infectieuse à des fins d'analyse. Une analyse longitudinale a servi à étudier divers traitements pour cette affection (qui nécessite une hospitalisation suivant diagnostic) [1].

D'autres comparaisons de traitements ont porté sur des affections plus courantes. Au Manitoba par exemple, les analyses longitudinales sur l'amygdalectomie ont mis l'accent sur des variables comme l'âge, le sexe et le nombre d'épisodes de troubles respiratoires avant l'opération dans les groupes traités et les groupes non traités. Comme il était possible de comparer les membres d'une même fratrie (l'un ayant été opéré et l'autre non) à l'aide de la base de données, les variables de la famille ont aussi été prises en considération [37,43]. Des tests de sensibilité ont permis d'estimer l'exactitude de l'analyse des résultats de l'amygdalectomie. On a pu comparer les résultats postopératoires pour un sous-ensemble de patients du Manitoba -ceux répondant à l'essai randomisé de Pittsburgh des critères pour l'amygdalectomie, soit sept épisodes d'amygdalite dans les douze mois précédant l'intervention chirurgicale - avec les résultats postopératoires pour les patients visés par l'essai de Pittsburgh. Les deux groupes de patients répondant aux mêmes critères pour l'amygdalectomie, on a observé que les résultats du Manitoba (tirés des comparaisons longitudinales) se rapprochaient sensiblement des résultats de l'essai randomisé de Pittsburgh. Les deux groupes de

patients ont montré les mêmes signes de progrès à la suite de l'amygdalectomie; les personnes opérées ont eu jusqu'à la moitié moins d'épisodes de troubles respiratoires que les personnes qui n'ont pas été opérées.

3. QUESTIONS DE FOND

Un certain nombre de questions ressortent des séries de données et des plans de recherche. Premièrement, outre qu'elles mettent l'accent sur le patient, les bases de données administratives peuvent servir fructueusement à analyser divers aspects du comportement du médecin. Cet usage est important puisque les médecins décident finalement de la façon dont 90% de chaque dollar consacré aux soins de santé sera dépensé [14]. Bien que certaines études sur le comportement des médecins puissent être faites à l'aide de données transversales, la plupart reposeront sur les données de la première catégorie (niveau le plus complet). Les médecins s'établissent dans des régions ou les quittent de même qu'ils adhèrent à des régimes d'assurance ou s'en désengagent; un registre des médecins est donc une nécessité pour de nombreuses études.

Les bases de données existantes devraient avoir une utilité grandissante pour l'analyse de décisions cliniques, qui est un domaine en pleine expansion [35]. Abrams et coll., [1], soulignent que l'analyse de décisions a permis de clarifier la structure de certaines controverses administratives en mettant en balance les risques et les avantages afférents à diverses stratégies selon une méthode quantitative. Du fait qu'elles servent à estimer les probabilités se rattachant au choix de divers traitements, les méthodes d'analyse de décisions exigent une meilleure information sur les résultats. Pour beaucoup de traitements, même ceux qui sont en usage depuis plusieurs années, les données sont insuffisantes. L'analyse de décisions combinée aux données sur les réclamations semblerait pouvoir remplacer ou précéder les essais randomisés, surtout lorsque ceux-ci sont particulièrement difficiles à exécuter [1]. Des études de ce genre exigent des données longitudinales - données de la première catégorie ou, le cas échéant, de la seconde catégorie.

Les bases de données relativement complètes sont indispensables pour évaluer les limites des bases de données moins complètes. Les analyses de données de la première catégorie (la plus complète) peuvent constituer un "standard" pour d'autres données. Nous pouvons ainsi juger de l'utilité d'une étude particulière sur un territoire administratif où l'on dispose d'une base de données moins complète. Éventuellement, la transmission de rétro-information aux autorités peut amener des changements dans le processus de collecte des données. Dans la section sur la casuistique et celle où l'on compare les données fondées sur les hôpitaux et les données fondées sur la population, nous nous servons des bases de données relativement complètes pour mettre en lumière les problèmes qui se rattachent aux bases moins complètes.

Comportement des médecins

Les analyses par région et les analyse d'hôpitaux nous amènent à nous interroger sur les habitudes des médecins [13]. Si l'on dispose d'identificateurs pour les médecins (médecin ayant autorisé l'admission à l'hôpital, médecin ayant pratiqué l'opération chirurgicale), on peut appliquer les mêmes méthodes que pour l'analyse des hôpitaux. Le rapport entre le volume d'activité chirurgicale dans les hôpitaux et les résultats des interventions de même que le suivi des médecins sont des sujets de recherche et d'intérêt administratif qu'il est justifié d'approfondir à l'aide des bases de données administratives [26,39].

Il convient aussi de se pencher davantage sur la façon dont les médecins orientent leurs patients vers les établissements spécialisés. Dans plusieurs États américains et

provinces canadiennes, les habitants des régions rurales et des petites villes qui représentent des cas complexes sont dirigés automatiquement vers des centres de soins tertiaires urbains sont de leur lieu de résidence [25]. Nous avons vu plus haut que la variation du nombre d'actes chirurgicaux entre les régions était attribuable à la variation de la proportion des cas dirigés vers des centres de soins tertiaires [52]. Cette constatation venait appuyer l'hypothèse selon laquelle la concentration d'un petit nombre de chirurgiens-spécialistes à un endroit avantageusement situé peut contribuer à perpétuer les variations régionales; des données du Manitoba indiquent que les écarts entre les régions pourraient persister encore longtemps. Comme on a établi précédemment qu'il existait une relation inverse entre le nombre de patients subissant un traitement thérapeutique ou une opération chirurgicale dans un hôpital et le nombre de résultats défavorables [29], les mesures axées sur la centralisation doivent être conçues de manière à ne pas mettre en péril l'égalité d'accès aux soins.

L'attitude des médecins à l'égard de l'hospitalisation est un autre sujet qu'il est facile d'explorer à l'aide des bases de données existantes. Puisque seulement une faible proportion des cas d'hospitalisation répondent à un modèle fondé sur les besoins thérapeutiques [67], les spécialistes supposent habituellement que des différences notables dans les habitudes de pratique médicale sont la principale cause de variation des taux d'utilisation des services hospitaliers à des fins chirurgicales ou thérapeutiques dans les régions [18]. Bien que ces hypothèses (et celles concernant les besoins des patients) aient rarement été vérifiées directement, elles ont des conséquences importantes pour les responsables de l'élaboration des politiques en matière de santé. Si l'on pouvait satisfaire tous les besoins grâce à une médecine plus efficiente, on pourrait comprimer les dépenses de santé sans pour autant réduire les services [67].

La base de données complète du Manitoba a permis de construire un indice de l'attitude des médecins à l'égard de l'hospitalisation. À cette fin, il a fallu:

1- identifier les patients réguliers des médecins
2- calculer le taux d'hospitalisation prévu pour ces patients sur une période de deux ans en tenant compte des différences d'état de santé
3- comparer le taux calculé ci-dessus pour chaque médecin avec le taux d'hospitalisation observé (réel) [52].

Cet indice sert à évaluer l'effet des habitudes de pratique des médecins sur l'utilisation des services hospitaliers, à vérifier la stabilité des habitudes de pratique à long terme et à examiner les motifs et les résultats de l'hospitalisation pour les patients des médecins qui ont une propension à hospitaliser et de ceux qui n'ont pas cette propension. Cette méthode peut servir à analyser l'hospitalisation à des fins aussi bien thérapeutiques que chirurgicales.

Analyse de Décisions Cliniques

L'information contenue dans les bases de données existantes peut être très utile à l'analyse de décisions cliniques. Par exemple, une analyse des traitements possibles pour une endocardite infectieuse a permis de savoir comment utiliser les études rétrospectives des dossiers d'un seul hôpital avec une grande base de données sur les demandes de remboursement [1]. Les probabilités calculées à l'aide des deux sources de données étaient très semblables, compte tenu de la valeur relativement faible de N tant pour les données relatives à l'hôpital seul (16 cas) que pour les données provinciales sur les demandes de remboursement des hôpitaux (127 cas). L'exploitation de la base de données du Manitoba a permis, par une agrégation sur une période de cinq ans (1979-1984), de faire une analyse des taux de survie et de comparer divers traitements possibles.

Les spécialistes de l'analyse de décisions cliniques ont mis au point des tests de sensibilité perfectionnés. Etant donné que les membres des diverses "branches" d'une analyse de cohorte (c'est-à-dire ceux ou celles qui subissent une opération particulière dans des hospitaux différents, ceux ou celles qui reçoivent des traitements différents) ne sont pas répartis aléatoirement, les tests de sensibilité sont très utiles pour faire une distinction entre les résultats compte tenu de la possibilité d'une erreur de mesure.

L'analyse de décisions cliniques doit aussi comprendre les études dont il a été question précédemment et qui révèlent des différences notables entre les établissements pour ce qui a trait aux résultats des opérations chirurgicales corrigés pour la casuistique. Les probabilités de différents résultats varient manifestement selon les hôpitaux et les chirurgiens; nous disposons maintenant de ce genre de données pour environ une demi-douzaine d'opérations chirurgicales [26]. Les analyses de décisions pourraient comporter des arbres de décision pour les "traitements de qualité supérieure" et les "traitements de faible qualité". Malgré que les résultats de ces analyses pourraient être controversés, ils refléteraient avec justesse la réalité.

Correction pour la casuistique

Il est essentiel de tenir compte des différences qui peuvent exister entre les caractéristiques des patients quand on compare les traitements et les établissements. Il n'est pas facile de déterminer le genre de données qu'il faut utiliser dans les circonstances pour analyser la durée de séjour, les taux d'invalidité et de mortalité et les cas de réhospitalisation. Comme le soulignent Jencks et Dobson [22], les indices de gravité qui servent à prévoir les résultats ne sont pas les mêmes que ceux qui servent à prévoir les coûts. La recherche dans ce domaine arrive à point car on s'apprête à engager des fonds pour recueillir des données additionnelles. Ainsi, le Pennsylvania Health Care Cost Containment Council a demandé récemment que l'on ajoute des données (recueillies par MedisGroups) dans les relevés de sortie des hôpitaux de manière à mieux tenir compte du facteur de risque dans l'évaluation de la qualité [22]. Toutefois, la nécessité de données prospectives coûteuses pour la correction pour la casuistique doit être établie.

Si les données transversales peuvent fournir assez de renseignements pour les totaux de contrôle pour la casuistique, il sera assez facile de réaliser des études exhaustives sur la mortalité hospitalière après l'intervention chirurgicale. Les chercheurs doivent déterminer de façon empirique si les formules de réclamation renferment suffisamment de données de chaque catégorie:

a) Troisième catégorie - données transversales

b) Deuxième catégorie - données de niveau intermédiaire (longitudinales sans fichier d'inscription)

c) Première catégorie - données complètes (longitudinales avec fichier d'inscription) pour les totaux de contrôle de la casuistique. Des résultats différents pourraient bien exiger des totaux de contrôle de nature différente.

Comme nous l'avons souligné plus haut, des efforts considérables ont été faits - et sont faits encore aujourd'hui -pour améliorer au meilleur coût possible l'évaluation de la casuistique à l'aide de données transversales. On s'est servi de données sur l'âge, le sexe et la comorbidité pour tenir compte des différences entre les hôpitaux dans l'étude de Luft et coll. [28,59]; ces données avaient trait à la période d'hospitalisation suivant une intervention chirurgicale. Lorsqu'on utilise les données contenues dans les formules de réclamation pour les séjours hospitaliers, les groupes formés selon le diagnostic (DRG) permettent souvent de tenir compte de l'âge, du sexe et (parfois) de la comorbidité. Jencks et Dobson [22] estiment qu'aucun indice de gravité de la maladie ne peut, à l'heure actuelle, accroître de façon notable l'exactitude des paiements de Medicare s'il est utilisé

dans le but de compléter ou de remplacer le système des groupes formés selon le diagnostic.

Lorsqu'on dispose d'identificateurs individuels cohérents, on peut déterminer des covariables additionnelles à l'aide des données relatives à la période ayant précédé l'hospitalisation de référence (c'est-à-dire l'hospitalisation qui accompagne l'intervention chirurgicale). On est en train de mettre au point une méthode d'analyse de résultats qui utilise des covariables établies à l'aide de formules de réclamation antérieures [44,72]. Ces covariables ne risquent pas d'être entachées d'un biais comme peuvent l'être les covariables tirées de données transversales. Par exemple, dans une analyse transversale, un diagnostic d'"infarctus du myocarde" peut définir une affection préexistante ou un événement survenu après l'intervention chirurgicale mais durant la période d'hospitalisation qui a suivi [4].

Le fichier d'inscription permet de faire en sorte que les personnes dont l'observation a débuté juste avant l'intervention chirurgicale ne soient pas confondues avec celles dont l'observation remonte à beaucoup plus loin. Ce genre de fichier est souhaitable; sa nécessité dépend de la population à l'étude. Seulement 54 et 63% des bénéficiaires de Medicaid dans deux États américains (Georgie et Michigan respectivement) ont figuré au fichier six mois avant et six mois après leur admission à l'hôpital; ces personnes avaient été hospitalisées pour subir une des huit opérations chirurgicales prédéfinies [24]. Dans les cas où il existe un régime de santé universel, les résultats de nombreux tests de sensibilité montrent qu'un fichier d'inscription n'est pas indispensable quand la période pré-opératoire est brève (n'excédant pas deux ans) et que le taux de migration est relativement faible [45]. Par conséquent, les analyses longitudinales de suivi fondées sur des données de Medicare permettent de croire que l'on ne devrait pas rencontrer de difficultés majeures si l'on applique la correction pour la casuistique sans se reporter à un fichier d'inscription.

Il faut déterminer avec soin la façon d'élaborer des mesures de correction pour la casuistique. À cause du nombre relativement restreint de résultats défavorables, il est fort souhaitable d'avoir de grandes bases de données dans les circonstances. Certaines méthodes - comme le modèle "affection/traitement" proposé par Wennberg [68] - sont subordonnées à l'utilisation d'un certain nombre de variables dichotomiques indépendantes (par exemple présence ou absence de cancer antérieurement) comme variables explicatives distinctes; toutefois, pour la plupart des traitements thérapeutiques et chirurgicaux, on relève assez peu de cas d'affection spécifique (par exemple présence de cancer). Étant donné le faible nombre de cas de ce genre, il peut être essentiel de fondre quelques-unes des variables indépendantes en un indice de comorbidité ou de gravité de la maladie pour la correction pour la casuistique.

Mosteller et coll. [31] soulignent que la perspective d'un gain issu de l'évaluation est une raison d'élaborer des échelles d'évaluation. Les indices et les échelles sont le gage d'une fiabilité accrue et de l'utilisation d'une variable numérique au lieu d'une variable dichotomique; la taille d'échantillon requise peut donc être moindre. Charlson et coll. [7] ont élaboré et testé une méthode de classification des données de comorbidité pour estimer le risque de décès. Comme l'indice de Charlson a été construit à partir des dossiers médicaux des hôpitaux, on devrait aussi pouvoir le construire à l'aide des données transversales sur les réclamations ou des réclamations enregistrées dans la période préopératoire. Pour plusieurs interventions chirurgicales courantes, les réclamations d'hôpitaux permettent à elles seules une correction pour la casuistique presque aussi bonne que celle que l'on obtient par le couplage des données contenues dans les réclamations d'hôpitaux et de celles enregistrées à la suite de visites de médecins ou à la suite d'enquêtes plus coûteuses [30,56]. Les recherches actuelles donnent à penser que la meilleure correction pour la casuistique découle de l'utilisation des réclamations produites avant l'hospitalisation de référence et de celles produites au cours de cette hospitalisation.

Dans l'étude sur la prostatectomie, que nous avons évoquée plus haut, la casuistique a été mesurée à l'aide de toutes les réclamations d'hôpitaux disponibles (c'est-à-dire, aussi bien celles enregistrées dans les six mois ayant précédé l'intervention que celles enregistrées au moment de l'intervention [74]. Les conditions pertinentes pour la période de six mois ayant précédé la prostatectomie étaient les suivantes: hospitalisé avec cancer (sauf cancer de la prostate), hospitalisé avec troubles cardio-vasculaires et pensionnaire d'une maison de soins. Les diagnostics correspondants au moment de l'intervention chirurgicale étaient les suivants: cancer de la prostate (non diagnostiqué auparavant), cancer (sauf cancer de la prostate), maladies cardio-vasculaires et autres diagnostics connexes (plus d'un diagnostic posé). Outre la comparaison entre hôpitaux, cette étude renfermait aussi une comparaison de la prostatectomie sanglante et de la prostatectomie transurétrale en ce qui à trait au taux de mortalité postopératoire. Comme les interventions transurétrales sont généralement recommandées pour les patients trop souffrants pour subir une intervention sanglante, il était évident que l'étape essentielle consistait à définir des covariables appropriées pour la correction entre les deux groupes.

À cette fin, on a groupé des données de réclamations et des données recueillies prospectivement auprès d'un hôpital du Manitoba [8,9]. Dans cet hôpital, une infirmière a interviewé chaque patient à propos de l'utilisation de médicaments avant l'opération (le nombre de médicaments utilisés et le nom de ces médicaments - Digitalis, etc.) et du nombre d'affections préopératoires dont souffrait le patient (obésité, troubles respiratoires, etc.). En outre, les anesthésiologistes ont évalué chaque patient à l'aide de la Classification de la condition physique de l'American Society of Anesthesiologists (ASAPS) [33]. Cette classification repose sur une échelle de 1 (patient en bonne santé) à 5 (patient qui risque de mourir des suites de l'opération).

Deux modèles de régression logistique ont été définis pour déterminer la probabilité de décès dans la période postopératoire. Un des modèles renfermait des variables calculées uniquement à l'aide de données administratives tandis que l'autre comprenait en plus des variables fondées sur des données recueillies prospectivement (ASAPS, etc.). Deux de ces variables (celles fondées sur ASAPS et l'usage de médicaments) étaient affectées de coefficients significatifs dans l'équation, où elles ont remplacé quatre des six variables fondées sur les données de réclamations. Le modèle qui renfermait des données prospectives avait une efficacité prédictive qui n'était que légèrement supérieure à celle de l'autre modèle; par ailleurs, les deux modèles s'ajustaient assez bien aux données. De plus, le coefficient faisant le lien entre le genre d'intervention chirurgicale et le taux de mortalité postopératoire était le même dans les deux modèles, ce qui donnait à penser que les données administratives constituaient des totaux de contrôle très utiles pour la gravité des cas [49]. Ces analyses sont appliquées à d'autres genres d'interventions chirurgicales; s'il s'avère que les observations relatives à la prostatectomie peuvent être généralisées, cela tendra fortement à démontrer que les données de réclamations peuvent produire des mesures de comorbidité presqu'aussi bonnes que celles que produisent les indices fondés sur des données primaires recueillies à grands frais.

Quand les données administratives sont-elles "assez bonnes" pour permettre d'identifier le meilleur de deux traitements, de tester des hypothèses sur la relation entre le volume d'activité chirurgicale et les résultats de traitement et de reconnaître les hôpitaux où les résultats sont particulièrement défavorables (ou particulièrement favorables)? Lorsqu'on étudie le rapport entre le volume d'activité chirurgicale et les résultats de traitements, il suffit de montrer que la casuistique moyenne dans les hôpitaux à fort volume d'activité chirurgicale n'est pas différente de celle observée dans les hôpitaux à faible volume d'activité chirurgicale. Les corrections statistiques sont beaucoup plus nécessaires dans la comparaison de divers traitements ou l'analyse des hôpitaux. Comme on ignore les différences réelles de casuistique et que celle-ci ne peut être corrigée entièrement en fonction des niveaux de connaissance actuels, il est

nécessaire de comparer les résultats de n'importe quelle méthode avec ceux obtenus à l'aide des **meilleures méthodes existantes.**

Comme nous venons de le voir, il est souvent difficile d'améliorer les résultats d'analyses uniquement à l'aide de données administratives. Étant donné les coûts afférents à la collecte prospective de données sur les facteurs de risque, l'expérience semble justifier difficilement une collecte générale de données. Lorsqu'il s'agit de quelques hôpitaux, la collecte prospective peut être utile pour recueillir des données sur les facteurs de risque et les affections postopératoires ou les causes possibles de certains résultats.

Données fondées sur les hôpitaux par comparaison aux données fondées sur la population

Les analyses de résultats sont facilitées par des données démographiques qui portent non seulement sur les événements survenant à l'hôpital où a eu lieu l'opération mais aussi sur l'ensemble des décès et des réadmissions. Par ailleurs, les données de l'hôpital où a eu lieu l'intervention peuvent être plus accessibles et plus faciles à analyser, ce qui accélère la transmission de la rétro-information et facilite le contrôle de la qualité des soins. Les analyses axées sur les hôpitaux (soit parce que les investigateurs sont rattachés à un seul hôpital ou que les personnes étudiées sont identifiées par un numéro propre à un hôpital) peuvent omettre des événements importants dans le suivi de la population de patients.

Ces recherches surviennent à un moment particulièrement opportun parce que l'on craint que le Prospective Payment System de Medicare ne favorise des séjours si brefs que la qualité des soins en souffrirait. Si tel était le cas, les réadmissions et les décès qui auraient pu être évités pourraient être omis ou attribués au mauvais hôpital par un système de contrôle. L'accessibilité à une base de données complète facilite l'amélioration des analyses de résultats. Les analyses transversales, qui sont fondées sur les hôpitaux et ne dépendent pas d'identificateurs individuels ou de fichiers d'inscription, peuvent être comparées aux analyses longitudinales, qui sont fondées sur la population et couvrent l'ensemble des réadmissions et des décès enregistrés dans une période donnée. Ainsi les bases de données plus complètes (que l'on retrouve normalement dans un nombre réduit de territoires administratifs) peuvent nous indiquer le genre d'interventions et la catégorie de patients pour lesquels la mortalité hospitalière constitue ou ne constitue pas un indice de mortalité approprié pour la période postopératoire. Les analyses préliminaires permettent de croire que pour certaines interventions (pontage coronarien, chirurgie de remplacement valvulaire), les décès survenant dans la période d'hospitalisation qui suit l'intervention représentent une très forte proportion des décès enregistrés dans les trois mois suivant l'intervention. Pour d'autres interventions ou affections (prostatectomie, fracture de la hanche), ce n'est pas le cas.

4. CONSIDÉRATIONS PRATIQUES

Les chercheurs qui veulent utiliser des données administratives se posent deux questions fondamentales: comment obtenir de meilleures données (plus complètes, plus détaillées)? et quel genre de logiciel utiliser pour analyser efficacement ces données? Dans cette section, nous voyons comment appliquer des méthodes de couplage d'enregistrements afin d'obtenir de meilleures données et nous analysons les caractéristiques de logiciels destinés à exploiter les bases de données existantes.

Couplage d'enregistrements

Le couplage d'enregistrements a souvent servi dans la recherche en médecine du travail à ajouter des données de mortalité aux données concernant les risques pour la santé au travail [62]. D'autres exemples de couplage avec des données de mortalité

comprennent les cohortes de maladie, les coefficients mode de vie/risque, les essais cliniques et les cohortes de population en général. Lorsqu'on dispose d'un nombre suffisant de variables identificatrices, le couplage d'enregistrements permet d'obtenir des données de la première catégorie pour des études de taux de survie. Cependant, le couplage peut aussi servir de façon plus générale à incorporer des données administratives à d'autres études [66]. Les méthodes statistiques qui sous-tendent le couplage d'enregistrements ont été analysées en profondeur dans d'autres documents. Le présent article montre comment le couplage d'enregistrements en tant qu'approche peut faire entrevoir des possibilités qui, dans d'autres circonstances, seraient insoupçonnées. Si nous avons un identificateur unique (ou une série d'identificateurs de nature différente), il est souvent possible de se servir d'études additionnelles (fondées sur des enquêtes, des analyses de dossiers d'hôpitaux et ainsi de suite) pour enrichir le fichier de chaque personne. Au Manitoba par exemple, l'Étude longitudinale sur le vieillissement a apparié restrospectivement des données d'enquête et des données de réclamations pour mieux faire ressortir la relation entre l'état de santé déclaré par les gens et l'utilisation des services médicaux. Des données du registre des tumeurs et des statistiques de l'état civil ont aussi été combinées à des données du registre de la Commission des services de santé du Manitoba et à des données de réclamations afin d'améliorer la qualité des renseignements contenus dans chaque fichier.

Chercher des possibilités de couplage est un exercice intellectuel utile lorsqu'on exploite des bases de données administratives. Le chercheur devrait peut-être commencer par déterminer les couplages souhaitables et ensuite s'occuper de l'aspect pratique. Cette approche nous a permis de réaliser plusieurs couplages dans notre étude sur des interventions chirurgicales courantes au Manitoba:

1- couplage des réclamations d'hôpitaux et des données primaires sur l'anesthésie et ses résultats afin de produire une série de données très complète;

2- couplage des réclamations d'hôpitaux et des relevés d'honoraires des médecins afin de vérifier l'authenticité des actes chirurgicaux et d'inscrire la date de l'intervention sur les réclamations d'hôpitaux;

3- couplage du fichier d'inscription et des statistiques de l'état civil afin de vérifier les décès et d'en indiquer la cause.

Bien que les méthodes de couplage aient varié d'un cas à l'autre, les fichiers appariés se sont tous avérés utiles. Les couplages permettent l'analyse de cas individuels et ont pour effet d'accroître sensiblement la quantité et la qualité des données.

De telle données peuvent provenir d'ailleurs. Si deux fichiers appariables ne renferment aucun nom ou aucun numéro d'identification, les investigateurs doivent alors trouver quatre ou cinq variables semblables dans les deux bases de données. Ces variables devraient suffire pour apparier les fichiers et appliquer des méthodes éprouvées pour évaluer la qualité des appariements [47,66].

Logiciels

Jusqu'à maintenant, nous avons concentré notre attention sur les aspects théoriques de l'exploitation des données administratives: plans de recherche, qualité et complétude des bases de données, définition des résultats et mesure de la casuistique. Sur le plan pratique, des logiciels appropriés constituent la première étape d'un programme de diffusion des concepts exposés dans ce document. Un système qui facilite l'analyse par région, les analyses longitudinales, le couplage d'enregistrements et ainsi de suite est nécessaire pour exploiter de façon efficace les bases de données existantes. Lorsque des

analystes exigent de l'information structurée de diverses manières, il est primordial d'avoir un logiciel facile à utiliser.

Au Canada, les organismes provinciaux de la santé diffèrent considérablement les uns des autres en ce qui a trait à la quantité de renseigments recueillis et à la structure utilisée pour la gestion des données. Même si les relevés de sortie des hôpitaux semblent contenir les mêmes genres de renseignements dans chaque province, l'analyse se fait systématiquement au niveau provincial plutôt qu'au niveau national. En s'y n efforçant, Statistique Canada pourrait constituer une base de données réduite sur les hôpitaux du pays. De même, la collecte des données de la Partie B du formulaire de Medicare aux États-Unis se fait au niveau des États sinon à un niveau inférieur avec des structures variées. Des groupes comme les collèges provinciaux des médecins et chirurgiens au Canada, les Peer Review Organizations (PRO) aux États-Unis, diverses catégories d'assureurs et les associations d'hôpitaux en Amérique du Nord pourraient être intéressés par un logiciel qui vise à accroître leur potentiel analytique "interne".

Un système comme celui décrit ci-dessus devrait aussi être polyvalent et facile à utiliser et exiger relativement peu de temps machine. Le rythme de développement des systèmes, les possibilités de modification et l'adaptabilité sont aussi des facteurs à considérer. D'après l'expérience des auteurs, le language SAS, qui est très répandu, s'est révélé l'outil idéal pour l'élaboration d'un système d'information de gestion à l'intention des analystes de la santé. Le SAS est caractérisé par sa souplesse; en effet, il réunit plusieurs variables (comme des diagnostics et des interventions) et de bonnes méthodes de formation de sous-groupes de données afin de produire un fichier abrégé contenant des réclamations et des variables particulières. Les caractéristiques du SAS permettent d'exploiter plusieurs bases de données dans la même unité de traitement. Avec l'augmentation de puissance, l'accroissement du temps d'exécution engendré par l'utilisation de languages évolués n'est plus aussi préoccupant [19].

Le macro-processeur du SAS et le macro-language qui l'accompagne ont servi à définir des procédures permettant d'exécuter les fonctions exposées ci-dessus. Ce macro-processeur offre une manière de stocker et d'extraire les travaux de SAS qui doit être adaptée à la situation. Lorsque les circonstances le permettent, les instructions sont les mêmes dans tous les programmes et la syntaxe dans les modules correspond à celle utilisée par le SAS. On se sert des caractéristiques régulières du SAS pour manipuler et analyser les données. On peut ainsi mettre sur pied rapidement un système de logiciels cohérent à partir des languages de quatrième génération qui existent déjà.

5. ANALYSE

Cet article a traité les diverses façons d'envisager le milieu fort renseigné dans lequel évolue les chercheurs, les gestionnaires et les responsables de l'élaboration des politiques. Les études sur la limitation des coûts et la qualité des soins peuvent influer notablement sur la quantité et la répartition des ressources consacrées au domaine de la santé. Notre analyse a fait ressortir les problèmes qui se rattachent à l'identification des établissements dont les résultats requièrent une attention spéciale. La comparaison des résultats de divers traitements peut s'appliquer directement aux deux domaines naissants que sont l'épidémiologie clinique et la prise de décisions médicale.

Étant donné l'accroissement spectaculaire du taux d'utilisation des services chirurgicaux chez les personnes âgées en Amérique du Nord [65] et l'inquiétude qui règne au sujet du coût des soins de santé, les bases de données administratives existantes sont essentielles pour observer les résultats de nombreux actes médicaux. Le fait de définir des variables permettant de prévoir les résultats postopératoires peut éclairer le médecin dans ses décisions. Ainsi, en apprenant que les prostatectomies pratiquées sur des

pensionnaires de maisons de soins se soldaient par un taux de mortalité relativement élevé, les urologues du Maine ont décidé de réduire le nombre de ces interventions [11].

Comme la majorité des essais cliniques ne visent qu'un nombre restreint de personnes, les analyses longitudinales qui utilisent les bases de données sur les réclamations constituent probablement la principale source de données sur les résultats pour les opérations pratiquées sur les membres de certains groupes d'âge. Par exemple, seules les réclamations peuvent fournir une information continuelle sur les cas de décès, de réadmission et de rétablissement sans complications parmi les personnes âgées à la suite d'un pontage coronarien. Cette méthode, dont l'usage s'est répandu rapidement pour les patients de plus de 65 ans, repose sur quelques séries longitudinales mais ne renferme aucun résultat d'essais cliniques randomisés [2,46].

Que faudra-t-il pour améliorer les résultats des opérations chirurgicales? Bien que le Canada et les États-Unis n'aient pas appliqué les mêmes méthodes pour limiter les coûts, les dépenses de santé ont augmenté dans les deux pays depuis le début des années 1970. Les augmentations ont été observées au chapitre de la prestation de services de santé dans des établissements en particulier [58,60] et dans le réseau en général. [15,36]. Le coût des services de santé ne doit pas être envisagé seul mais par rapport rapport à la qualité des soins et aux résultats [58]. Or, il est presque impossible d'obtenir des données illustrant le rapport entre des hausses de coûts particulières et la variation de la qualité des soins. L'augmentation des coûts directs des hôpitaux au cours des dix dernières années aurait dû se traduire par une amélioration de la qualité des soins. L'intensification du suivi des personnes ayant subi une intervention chirurgicale, l'accentuation des soins postopératoires et le perfectionnement du personnel infirmier sont toutes des mesures qui se sont concrétisées en Amérique du Nord. Ces mesures visaient principalement à améliorer la qualité des soins. Au cours des dix dernières années, on a aussi centralisé au Manitoba les opérations à risques élevés et limité la pratique chirurgicale des médecins non spécialistes.

Cependant, des études n'ont révélé aucune amélioration notable en ce qui a trait aux résultats de trois genres d'interventions chirurgicales courantes - hystérectomie, cholecystectomie et prostatectomie - pratiquées au Manitoba pendant dix ans [44]. L'analyse de trois cohortes de chirurgie (toutes les personnes ayant subi une intervention en 1972-1973, 1977-1978 ou 1982-1983) n'a révélé aucune baisse sensible du taux de mortalité postopératoire ou du taux de réadmission dans la période qui suit immédiatement l'intervention ou encore du taux de réadmission pour des complications survenues dans les quinze mois suivant l'intervention. Ces données pourraient nous laisser croire que la science médicale ne progresse plus aussi rapidement [32]. Au début des années 1970, l'organisation et la prestation des soins relatifs à ces trois genres d'interventions étaient d'une qualité telle qu'il y avait peu de possiblités d'améliorer les résultats des opérations chirurgicales malgré la progression soutenue des dépenses dans le secteur hospitalier.

Un suivi attentif des établissements où les résultats sont inférieurs aux prévisions et la transmission de rétro-information à ces établissements auront vraisemblablement beaucoup plus d'effet sur la qualité que la hausse graduelle du budget des dépenses à la grandeur du réseau. Dans un réseau où la qualité générale des soins est passablement élevée, il est difficile de réaliser des progrès, même faibles. En mettant l'accent sur le suivi et la rétro-information, nous pouvons concentrer notre attention sur les hôpitaux où il est plus facile de réaliser des progrès. S'il s'agit de freiner les coûts, les méthodes analysées dans ce document permettent de se concentrer sur certaines régions ou certains hôpitaux plutôt que de s'intéresser à l'ensemble.

La mise à jour, l'analyse et l'amélioration des bases de données existantes représentent un moyen peu coûteux de mieux concevoir l'accessibilité et la qualité des soins. Nous avons la technique et les données à notre portée; il ne reste plus qu'à obtenir les fonds qui

permettront de faire les études nécessaires et à vouloir mettre en application les conclusions de ces études.

6. REMERCIEMENTS

Les auteurs tiennent à exprimer leur reconnaissance à la Commission des services de santé du Manitoba pour son soutien. Cette étude a été subventionnée en vertu du Programme national de recherche et développement en matière de santé (projet no. 6607-1197-44) et a fait l'objet d'une Bourse de chercheur émérite (no. 6607-1314-48). Les interprétations et les opinions contenues dans cet article sont celles des auteurs et ne reflètent pas nécessairement la position de la Commission des services de santé du Manitoba ou de Santé et Bien-Être social Canada. Les auteurs tiennent également à remercier Sandra Sharp et Kerry Meagher pour leur aide dans la préparation du manuscrit.

RÉFERENCES

[1] Abrams, H.D., Detsky, A.S., Roos, L.L., et coll. Is there a role for surgery in the acute management of infective endocarditis? A decision analysis and medical date base approach, Med. Decis. Making, 1988. (à paraître)

[2] Anderson, G.M., et Lomas, J. Monitoring the diffusion of technology: coronary artery bypass graft surgery in Ontario, Am. J. Public Health, 1988. (à paraître)

[3] Barer, M.L., Evans, R.G., Hertzman, C., et coll. Aging and health care utilization: new evidence on old fallacies, Soc. Sci. Med. 24:851, 1987.

[4] Blumberg, M.S. Risk adjusting health care outcomes: a methodologic review, Med. Care 43:351, 1986.

[5] Brook, R.H., et Lohr, K.N. Efficacy, effectiveness, variations, and quality: boundary-crossing research, Med. Care 23:710, 1985.

[6] Caper, P. Variations in medical practice: implications for health policy. Health Affairs 3(2):110, 1984.

[7] Charlson, M.E., Pompei, P., Ales, K.L. et coll. A new method of classifying prognostic comorbidity in longitudianl studies development and validation, J. Chron. Dis. 40:373, 1987.

[8] Cohen, M.M., et Duncan, P.G. Physical status score and trends in anesthetic complications, J. Chron. Dis., 1988. (à paraître)

[9] Cohen, M.M., Duncan, P.G., Pope, W.D.B., et coll. A survey of 112,000 anesthetics at one teaching hospital (1975-1983), Can. Anaesth. Soc. J. 33:22, 1986.

[10] Cretin, S., et Worthman, L.G. Alternative systems for case mis classification in health care financing (Rand R-3457-HCFA), Santa Monica, CA, 1986, Rand Corporation.

[11] Davis, H. "Was surgery needed?" dans The Baltimore Sun, 6 avril 1986.

[12] Demlo, L.K., et Campbell, P.M. Improving hospital discharge data: lessons from the National Hospital Discharge Survey, Med. Care 19:1030, 1981.

[13] Egdahl, R.H. Ways for surgeons to increase the efficiency of their use of hospitals, New Engl. J. Med. 309:1184, 1983.

[14] Eisenberg, J.M. Physician utilization: the state of research about physicians' practice pattern, Med. Care 23:461, 1985.

[15] Evans, R.G. Finding the levers, finding the courage: what have we learned about cost containment in North America? J. Health, Politics, Policy & Law 11:585, 1987.

[16] Feinleib, M. Data bases, data banks and data dredging: the agony and the ecstasy, J. Chron. Dis. 37:783, 1984.

[17] Fetter, R.B., Shin, Y., Freeman, J. L., et coll. Case mis definition by diagnosis-related groupes, Med. Care 18:1 (supp) 1980.

[18] Griffith, J.R., Restuccia, J.D., Tedeschi, P.J. et coll. Measuring community hospital services in Michigan, Health Serv. Res. 16:135, 1981.

[19] Harel, E.C., et McLean, E.R. The effects of using a non-procedural computer language on programmer productivity, MIS Quarterly 9:109, 1985.

[20] Hornbrook, M.C. Techniques for assessing hospital case mix, Annu. Rev. Public Health 6:295, 1985.

[21] Horwitz, R.I. The experimental paradigm and observational studies of cause-effect relationships in clinical medicine, J. Chron. Dis. 40:91, 1987.

[22] Jencks, S.F., et Dobson, S. Refining case-mis adjustment: the research evidence, New Engl. J. Med. 317:679, 1987.

[23] Jencks, S.F., et Dobson, A. Strategies for reforming Medicare's physician payments: physician diagnosis-related groups and othe approaches, New Engl. J. Med. 312:1492, 1985.

[24] Klingman, D., Pine, P., et Simon, J. Outcomes of surgery among Medicaid recipients in Georgia and Michigan: 1981-1982, 1987 (version préliminaire, soumise à Med. Care.)

[25] Luft, H.S. Regionalization in medical care, Am. J. Public Health 75:125, 1985.

[26] Luft, H.S., Garnick, D.W., Mark, D. et coll. Evaluating research on the use of volume of services performed in hospitals as an indicator of quality, Washington, DC, 1987, Office of Technology Assessment, Congress of the United States. (version préliminaire)

[27] Luft, H.S., Bunker, J.P., et Enthoven, A.C. Should operations be regionalized? The empirical relation between surgical volume and mortality, New Engl. J. Med. 301:1364, 1979.

[28] Luft, H.S., et Hunt, S.S. Evaluating individual hospital quality through outcome statistics. J.A.M.A. 255:2780, 1986.

[29] Maerki, S.C., Luft, H.S., et Hunt, S.S. Selecting categories of patients for regionalization: implications of the relationship between volume and outcome, Med. Care 24:148, 1986.

[30] Mossey, J.M., et Roos, L.L. Using insurance claims to measure health status: the illness scale, J. Chron. Dis. 40:41S (suppl), 1987.

[31] Mosteller, F., Gilbert, J.P., et McPeek, B. Reporting standards and research strategies for controlled trials, Con. Clin. Trials 1:37, 1980.

[32] Neuhauser, D. "Cost-effective clinical decision-making implications for the delivery of health services", dans Bunker, J.P., Barnes, B.A., et Mosteller, F. (ed.), Costs, risks, and benefits of surgery, New York, 1977, OXford University Press.

[33] Owens, W.D., Felts, J.A., et Spitznager, E.L., Jr. ASA physical classifications: a study of consistency of ratings, Anesthesiology 49:239, 1978.

[34] Patricelli, R.E. Employers as managers of risk, cost, and quality, Health Affairs 6(3):75, 1987.

[35] Pauker, S.G., et Kassirer, J.P. Decision analysis, New Engl. J. Med. 316:250, 1987.

[36] Reinhardt, U.E. Resource allocation in health care: the allocation of lifestyles to providers, Milbank Mem. Fund Q. 65:153, 1987.

[37] Roos, L.L. Alternative designs to study outcomes: the tonsillectomy case, Med. Care 17:1069, 1979.

[38] Roos, L.L., Cageorge, S.M., Austen, E., et coll. Using computers to identify complications after surgery, Am. j. Public Health 75:1288, 1985.

[39] Roos, L.L., Cageorge, S.M., Roos, N.P., et coll. Centralization, certification, and monitoring: readmissions and complications after surgery, Med. Care 24:1044, 1986.

[40] Roos, L.L., Nicol, J.P., et Cageorge, S.M. Using administrative data for longitudinal research: comparisons with primary data collection, J. Chron. Dis. 40:41, 1987.

[41] Roos, L.L., Nicol, J.P., Johnson, C., et coll. Using administrative data banks for research and evaluation: a case study, Eval. Q. 3L236, 1979.

[42] Roos, L.L., Roos, N.P., Cageorge, S.M., et coll. How good are the date? Relaibility of one health care data bank, Med. Care 20:266, 1982.

[43] Roos, L.L., Roos, N.P., et Henteleff, P.D. Assessing the impact of tonsillectomies, Med. Care 16:502, 1978.

[44] Roos, L.L., Roos, N.P., et Sharp, S.M. Monitoring adverse outcomes of surgery using administrative date, Hearth Care Fin. Rev. 7:5 (suppl), 1987.

[45] Roos, L.L., et Sharp. S.M. Becoming more efficient at outcomes research, Intl. J. Tech. Asses. Health Care, 1988. (à paraître)

[46] Roos. L.L., et Sharp, S.M. Innovation, centratization, and growth: coronary artery bypass graft surgery in Manitoba, 1987. (soumis à des fins de publication.)

[47] Roos, L.L., Sharp, S.M., et Wajda, A. Assessing data quality: a computerized approach, Soc. Sci. Med., 1988. (à paraître)

[48] Roos, L.L., Wajda, A., et Nicol, J.P. The art and science of record linkage: methods that work with few indentifiers, Comput. Bio. Med. 16:45, 1986.

[49] Roos, N.P. Differential use of outpatient surgery by hospital physicians: what are the potential savings? Journal de l'Ass. médical canadienne, 1988. (à paraître)

[50] Roos, N.P. Hysterectomies in one Canadian province. A new look at risks and benefits, Am. J. Public Health 74:39, 1984.

[51] Roos, N.P., et Danzinger, R.G. Assessing surgical risks in a population: patient histories before and after cholecystectomy, Soc. Sci. Med. 22:571, 1986.

[52] Roos, N.P., Flowerdew, G., Wajda, A., et coll. Variations in physicians' hospitalization practices: a population-based study in Manitoba, Canada, Am. J. Public Health 76:45, 1986.

[53] Roos, N.P., et Lyttle, D. Hip arthroplasty surgery in Manitoba: 1973-1978, Clin. Orthop. (199):248, 1985.

[54] Roos, N.P., et Ramsey, E. A population-based study of prostatectomy: long term outcomes associated with differing surgical approaches, J. Urol. 137:1184, 1987.

[55] Roos, N.P., et Roos, L.L. Hign and low surgical rates: risk factors for area residents, Am. J. Public Health 71:591, 1981.

[56] Roos, N.P., Roos, L.L., Mossey, J.M., et coll. Using administrative data to predict important health outcomes: entry to hospital, nursing home and death, Med. Care 1988. (à paraître)

[57] Schwartz, W.B. The inevitable failure of current costcontainment strategies: why they can provide only temporary relief, J.A.M.A 257:220, 1987.

[58] Scitovsky, A.A. Changes in the costs of treatment of selected illnesses, 1971-1981, Med. Care 23:1345, 1985.

[59] Showstack, J.A., Rosenfeld, K.E., Garnick, D.W., et coll. Association of volume with outcome of coronary artery bypass graft surgery: scheduled vs. nonscheduled operations, J.A.M.A. 257:785, 1987.

[60] Showstack, J.A., Stone, M.H., et Schroeder, S.S. The role of changing clinical practices in the rising costs of hospital care, New Engl. J. Med. 313:1201, 1985.

[61] Sloan, F.A., Perrin, J.M., et Valvona, J. In-hospital mortality of surgical patients: Is there an empiric basic for standard setting? Surgery 99:446, 1986.

[62] Smith, M.E. "Record linkage: organizing the facts together" dans Bennett, B. M. et Trute, B. (ed. Mental health information systems: problems and prospects, New York, 1984, Edwin Mellen Press.

[63] Stern, R.S., et Epstein, S.M. Institutional responses to prospective payment bases on diagnosis-related groups: implications for cost, quality, and access, New Engl. J. Med. 312:621, 1985.

[64] Thomas, J.W., Ashcraft, M.L.S., et Zimmerman, J. An evaluation of alternative severity of illness measures for use by university hospitals, Ann Arbor, Mich, 1986, Department of Health Services Management and Policy, School of Public Health, University of Michigan.

[65] Valvona, J., et Sloan, F. Rising rates of surgery among the elderly. Health Affairs 4(3):108, 1985.

[66] Wajda, A., et Roos, L.L. Simplifying record linkage: software and strategy, Comput. Biol. Med. 17:239, 1987.

[67] Wennberg, J.E. Commentary: on patient need, equity, supplier-induced demand and the need to assess the outcome of common medical practices, Med. Care 23:512, 1985.

[68] Wennberg, J. E. Which rate is right? New Engl. J. Med. 310:310, 1986.

[69] Wennberg, J. E., Blowers, L., Parker, R., et coll. Changes in tonsillectomy rates associated with feedback and review, Pediatrics 59:821, 1977.

[70] Wennberg, J. E., et Fowler, F. J. A test of consumer contribution to small area variations in health care delivery, J. Maine Med. Assoc. 68:275, 1977.

[71] Wennberg, J. E., Freeman, J., et Culp, W. J. Are hospital services rationed in New Haven or over-utilised in Boston? Lancet, May 23:1185, 1987.

[72] Wennberg, J. E., McPherson, K., et Caper, P. Will payment based on diagnosis-related groups control hospital costs? New Engl. J. Med. 311:295, 1984.

[73] Wennberg, J. E., Roos, N. P., Sola, L., et coll. Use of claims data systems to evaluate health care outcomes: mortality and reoperation following prostatectomy, J.A.M.A. 257:933, 1987.

[74] West, R., Sherman, G.J., et Downey, W. A record linkage study of valproate and malformations in Saskatchewan, Revue canadienne de santé publique, 76:226, 1985.

[75] Wennberg, J. E. Commentary: using claims to measure health status, J. Xhron. Dis. (suppl) 40:51S, 1987.

[76] Wennberg, J. E. Which rate is right? New Engl. J. Med. 310:310, 1986.

[77] Wennberg, J. E., Blowers, L., Parker, R. et coll. Changes in tonsillectomy rates associated with feedback and review, Pediatrics 59:821, 1977.

[78] Wennberg, J. E. et Fowler, F. J. A test of consumer contribution to small area variations in health care delivery, J. Maine Med. Assoc. 68:275, 1977.

[79] Wennberg, J. E. Freeman, J. et Culp, W. J. Are hospital services rationed in New Haven or over-utilised in Boston? Lancet, May 23:1185, 1987.

[80] Wennberg, J. E., McPherson, K. et Caper, P. Will payment based on diagnosis-related groups control hospital costs? New Engl. J. Med. 311:295, 1984.

[81] Wennberg, J. E., Roos, N. P., Sola, L. et coll. Use of claims data systems to evaluate health care outcomes: mortality and reoperation following prostatectomy, J.A.M.A. 257:933, 1987.

[82] West, R., Sherman, G.J. et Downey, W. A record linkage study of valproate and malformations in Saskatchewan, Revue canadienne de santé publique, 76:226, 1985.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

IDENTIFICATEURS MANQUANTS ET JUSTESSE DE L'OBSERVATION SUIVIE

MARTHA E. FAIR et PIERRE LALONDE[1]

RÉSUMÉ

Chaque fois qu'il y a des formulaires à remplir, les personnes qui cherchent à réduire au minimum le travail à effectuer et à s'assurer que les gens à qui on demande des renseignements ne réagissent pas de façon négative se demandent s'il est vraiment nécessaire de recueillir certains identificateurs personnels. Dans cette étude, nous allons tenter de déterminer dans quelle mesure la présence ou l'absence de divers identificateurs a une influence sur le taux d'erreur de couplage. Le test consiste à supprimer temporairement certains des identificateurs les plus importants (p. ex. les prénoms au complet, la date de naissance et le nom de jeune fille de la mère) parmi ceux que l'enregistrement de départ contient, individuellement ou en groupe. Les recherches sont répétées et, selon le cas, certains éléments d'information sont utilisés ou non. Les données ayant servi à l'essai sont tirées d'une étude de la mortalité chez les mineurs ontariens. À cet égard, la date de naissance complète revêt une importance particulière, suivie des prénoms au complet. Ces données ont une incidence considérable sur l'élaboration des questionnaires d'enquête ainsi que sur l'étude de la faisabilité de projets de couplage et du taux de réussite prévu de ces projets. Une méthode permettant de compléter les identificateurs auxquels on a facilement accès lorsque leur nombre est limité est aussi examinée dans cette étude. Celle-ci consiste à utiliser un fichier supplémentaire (p. ex. le fichier des numéros d'assurance sociale) pour faciliter le couplage et en améliorer la qualité.

1. INTRODUCTION

Cette communication a pour objet de décrire les résultats de tests que nous avons effectués dans le but de produire des données quantitatives mesurant l'incidence que la présence ou l'absence de divers identificateurs personnels ou groupes d'identificateurs a sur le degré d'exactitude d'un couplage probabiliste d'enregistrements provenant de deux fichiers de données administratives, par exemple un fichier sur les mineurs ontariens et la base de données sur la mortalité au Canada (BDMC). Cette étude se situe dans la lignée de travaux effectués à Statistique Canada avec l'appui de la Commission de contrôle de l'énergie atomique, du ministère ontarien du Travail et de la Commission des accidents

[1] Martha E. Fair et Pierre Lalonde, Statistique Canada, Immeuble R.H. Coats, 18e étage, Parc Tunney Ottawa (Ontario) K1A 0T6.

du travail (CAT), lesquels visent à mettre au point des méthodes de couplage et de collecte de données qui permettent de réaliser une observation suivie complète de cohortes de travailleurs, par exemple les personnes exposées aux rayonnements ionisants (Fair et coll., 1988a et 1988b). Nous avons également eu la collaboration d'Emploi et Immigration Canada.

Le contenu de la communication est divisé en six parties:

1. Description du contexte et présentation de l'étude;
2. Examen des principaux résultats et conclusions;
3. Description des fichiers utilisés;
4. Description de la méthode appliquée;
5. Exposé des résultats détaillés;
6. Discussion de l'incidence que cette étude aura pour ce qui est de:
   - l'estimation des taux d'erreur prévus lorsqu'il y a couplage des données contenues dans deux fichiers;
   - l'utilisation d'un fichier intermédiaire pour faciliter le couplage;
   - l'élaboration d'une méthode de collecte de données recommandée pour les études.

Voici à quel genre de questions nous essayons de répondre:

1. De combien les chances de trouver le bon enregistrement diminuent-elles lorsqu'un des fichiers faisant l'objet du couplage contient seulement l'année de naissance, plutôt que la date complète (année, mois, jour)?
2. De combien les chances de trouver le bon enregistrement augmentent-elles si l'on ajoute des éléments d'information comme le lieu de naissance et le nom de jeune fille de la mère au moment de la collecte des données?
3. Quelle est l'incidence sur les chances de trouver le bon enregistrement si l'on dispose seulement des initiales et de l'année de naissance plutôt que des prénoms et de la date de naissance au complet?
4. Dans quelle mesure est-il important de demander, au moment de la collecte des données d'enquête, les prénoms et la date de naissance au complet plutôt que simplement les initiales et l'âge si l'on prévoit de procéder plus tard au couplage de ces données et de données provenant d'autres sources?
5. Si l'on prévoit que les chances de trouver le bon enregistrement seront faibles, peut-on augmenter ces chances en passant en plus par un fichier "intermédiaire" comme le fichier index des numéros d'assurance sociale (NAS)?
6. L'utilisation des identificateurs que contient le fichier principal des numéros d'assurance sociale augmente-t-elle de façon marquée les chances de trouver l'enregistrement indiquant qu'une personne est décédée? Si oui, comment peut-on tenir compte de cette information dans la recherche de l'enregistrement indiquant qu'une personne est décédée?

2. DESCRIPTION DU CONTEXTE ET PRÉSENTATION DE L'ÉTUDE

La société et les législateurs sont souvent confrontés à des questions pressantes concernant les effets à long terme sur la santé d'agents nocifs présents sur les lieux de travail ou ailleurs. Récemment, le désastre de Tchernobyl a amené les médias et le public dans le monde entier à s'interroger sur les effets éventuels des rayonnements ionisants sur la santé des personnes qui y sont exposées et sur celle de leurs enfants. On s'est

également interrogé un peu partout sur les risques éventuels pour la santé des personnes qui, au moment de l'accident, étaient loin de l'endroit où il s'est produit.

De fortes pressions se font sentir dans la société aujourd'hui pour que soient déterminés et révélés les risques pour la santé auxquels le public est exposé, particulièrement lorsque l'effet néfaste est cumulatif ou ne se manifeste pas avant longtemps, comme dans le cas du cancer. Ces pressions sont exercées par les médias, les organismes investis d'un pouvoir de réglementation, les syndicats, les commissions spéciales et les chercheurs. On juge que les travailleurs et le grand public ont le droit d'être protégés contre les situations qui présentent un risque connu ou soupçonné pour la santé.

Le genre de statistiques dont il est question ici sont différentes de celles qu'on établit à partir des données de recensement ou d'enquête, lesquelles font essentiellement le portrait d'une situation à un moment donné. Lorsqu'on fait une observation médicale suivie de longue durée, on choisit une population et on suit son évolution pendant peut-être 20 à 30 ans pour se renseigner sur la santé de ses membres et savoir lesquels sont encore en vie et lesquels sont décédés. Pour avoir les moyens statistiques de déceler certains effets éventuels sur la santé, il faut parfois suivre des populations à effectif élevé, comparer les résultats de plusieurs études afin de déterminer s'il est possible d'obtenir des résultats analogues et éventuellement regrouper des données internationales.

Les épidémiologistes et les statisticiens qui font des études de mortalité portant sur des cohortes de travailleurs doivent pratiquement jouer au détective pour obtenir les données dont ils ont besoin. Les systèmes d'information sur la santé et les décès qui ont été mis au point à des fins administratives et les bases de données qu'ils ont permis de constituer sont une mine de renseignements utiles sur l'état de santé de la population qui nous intéresse (Last 1986).

Au Canada, on a de plus en plus recours depuis quelques années au couplage informatisé des enregistrements lorsqu'on réalise des observations médicales suivies de longue durée. Dans plus de 50 cas, on a utilisé la base de données sur la mortalité au Canada (Smith et Newcombe, 1980). Il s'agit d'un fichier longitudinal contenant plus de 6 millions d'enregistrements faisant état de décès survenus au Canada depuis 1950 et indiquant par un code la cause du décès. La base de données sur la mortalité au Canada, le Système itératif général de chaînage d'articles et les techniques de couplage probabiliste utilisées ont déjà été décrits en détail ailleurs (Smith 1981, 1986; Hill et Pring-Mill 1985; Howe et Lindsay 1981).

La plupart des études de mortalité par cohorte visent essentiellement à déterminer si la population choisie est exposée à un risque trop élevé de décès pendant la période d'observation. Pour ce faire, il faut connaître à la fois l'effectif de la cohorte, la période d'observation et le nombre prévu de décès associés à l'ensemble des causes ou à certaines causes en particulier. La cohorte de travailleurs est suivie, et il est déterminé à la fin de la période d'observation lesquels de ses membres sont vivants et lesquels sont décédés (figure 1) (Simon et Toulbee, pages 25 à 90; Groupe de travail sur la santé et la sécurité des travailleurs de laboratoire; Redmond et coll., 1969; Monson 1980). La proportion de certificats de décès obtenus pour les cas de décès connus, la proportion de travailleurs pour lesquels la recherche a été infructueuse (c.-à-d. ceux pour lesquels il a été impossible de savoir à la fin de l'étude s'ils étaient vivants ou décédés) et la proportion de la cohorte qu'on a observée sont autant de données qui devraient être présentées en même temps que les résultats de l'étude. En effet, comme elles sont révélatrices de la quantité de données manquantes, elles ont une incidence directe sur la validité des conclusions. La proportion de la cohorte observée pour laquelle la recherche a été infructueuse constitue la limite supérieure du nombre de personnes dont le décès n'a pas été déterminé.

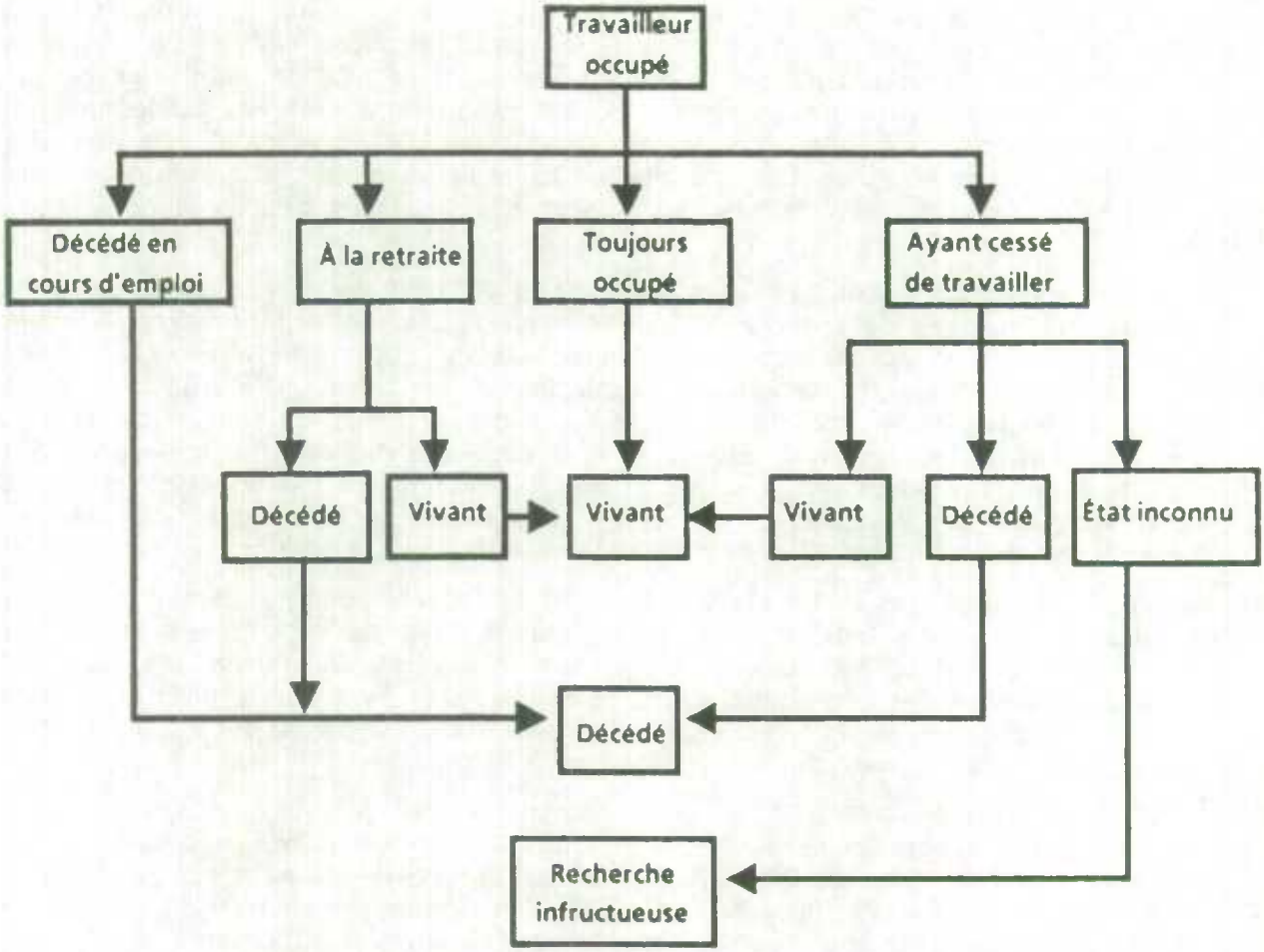


Figure 1. Observation suivie d'une cohorte de travailleurs visant à déterminer s'ils sont décédés

Les chercheurs qui effectuent ce genre d'étude ont des exigences très variées en ce qui a trait à la qualité des résultats qu'ils visent. Pour certaines statistiques, on se contente d'une estimation grossière de la proportion de décès. Dans d'autres situations, il est souvent important de régler les cas sur lesquels plane le doute et de déterminer dans la mesure du possible s'il s'est passé quelque chose qui fait qu'on ne pourra pas trouver d'enregistrement indiquant qu'une personne est décédée, par exemple si celle-ci a émigré. Les épidémiologistes qui font des analyses par cohorte exigent souvent une observation suivie très poussée de chaque membre de la cohorte en vue de savoir s'il est vivant ou non, particulièrement s'il est question d'établir des normes réglementaires ou d'élaborer des critères d'indemnisation. Il est alors souhaitable de retracer l'histoire de la plus forte proportion possible de membres de la cohorte, car le fait de ne pas pouvoir établir de façon certaine si une partie appréciable de la cohorte est vivante ou non peut donner lieu à des conclusions erronées ou trompeuses. Si moins de 95% des recherches aboutissent, on peut se demander si les résultats ne sont pas biaisés.

Pour le test, nous avons utilisé les résultats d'une étude de mortalité portant sur une cohorte de mineurs ontariens et dont la période d'observation s'étendait de 1950 à 1977

(Muller et coll., 1983). Les résultats que nous avons obtenus ont influé sur la conception et la mise au point d'une méthode pouvant convenir à plusieurs autres études, particulièrement celles de cohortes de travailleurs dont on connaît le numéro d'assurance sociale, et sur l'élaboration de la procédure de collecte de données que nous recommandons et qui a été conçue à l'intention des chercheurs qui voudront mener des études de mortalité, des études sur l'incidence du cancer ou des études génétiques.

Il convient de souligner qu'aucun indentificateur individuel numérique ne figurait dans les deux premiers fichiers où se trouvaient les enregistrements que nous cherchions à rapprocher, de sorte qu'il a fallu avoir recours aux noms, aux dates de naissance et autres données de ce genre. Les registres canadiens de l'état civil contiennent rarement le numéro d'assurance sociale sous une forme assimilable par machine et, dans certaines provinces, on n'a pas prévu d'endroit où inscrire ce numéro sur le document de base original. Il nous semblait que le fichier principal des numéros d'assurance sociale contenait de meilleurs identificateurs personnels. Nous avons donc pensé que, lorsque le NAS se trouve dans le fichier se rapportant à la cohorte de mineurs, la recherche des enregistrements indiquant lesquels de ces derniers sont décédés serait probablement facilitée par l'utilisation de fichiers index comme fichiers "intermédiaires". Le but de notre étude était donc de déterminer si les identificateurs contenus dans le fichier principal des NAS pouvaient être utilisés comme nous le proposions et de mesurer l'effet que cette utilisation aurait sur les chances de trouver les enregistrements indiquant quels membres de la cohorte étaient décédés. Pour le couplage des données contenues dans le fichier sur les mineurs ontariens et dans la base de données sur la mortalité au Canada, nous nous sommes servis du Système itératif général de chaînage d'articles et nous avons appliqué des techniques de couplage probabiliste.

Nos tests visaient particulièrement à évaluer la qualité de notre produit statistique selon le nombre d'identificateurs personnels dont nous disposions dans le fichier pour effectuer le couplage (tableau 1). Par le passé, nous avions recommandé que les données recueillies en vue d'un rapprochement avec la base de données sur la mortalité au Canada contiennent le nom de famille à la naissance et celui utilisé par la suite, les prénoms ainsi que la date et le lieu de naissance au complet. Le nom de jeune fille de la mère et les variables se rapportant aux parents pouvaient aussi être très utiles. Nous souhaitions plus précisément connaître l'ampleur des conséquences négatives si certains identificateurs étaient absents. Il est connu que la date de naissance et les prénoms sont des éléments d'information très importants, mais nous n'avions pas beaucoup de données quantitatives sur les erreurs causées par l'absence de divers identificateurs dans les fichiers. Nous voulions également étudier la possibilité d'utiliser le fichier principal des NAS pour faciliter la recherche des enregistrements indiquant lesquels des mineurs ontariens étaient décédés, car si ces variables étaient plus utiles que les autres, nous pourrions à la fois améliorer la qualité du produit et réduire la quantité de travail manuel à faire dans cette étude.

Tableau 1

Identificateurs figurant dans les fichiers de la CAT, des NAS et de la BDM

Identificateur	Fichier de la CAT	Fichier des NAS	Fichier de la BDM
Numéro d'assurance sociale	+	+	nfp
Noms*	+	+	+
Date de naissance	+	+	+*
Lieu de naissance	+**	nfp	+
Sexe	+	+	+
Province de résidence	+	nfp	+
Nom de jeune fille de la mère	nfp	+	+*
Vivant jusqu'à l'année X au moins	+	nfp	+

\+ Figure dans le fichier
* Pas toujours complet
** Le même code est utilisé pour le Canada et les États-Unis
nfp = ne figure pas dans le fichier

3. PRINCIPAUX RÉSULTATS ET CONCLUSIONS

On estime que sur les 30,000 enregistrements concernant des mineurs qui avaient déjà fait l'objet d'un rapprochement avec les enregistrements de la base de données sur la mortalité au Canada, 2,243 avaient été correctement appariés (le décès avait été établi de façon certaine). Là-dessus se trouvaient 705 paires dont chacun des enregistrements contenait tous les identificateurs, en l'occurrence la date de naissance au complet, deux prénoms au complet et le nom de jeune fille de la mère. Pour déterminer dans quelle mesure les taux d'erreur se ressentent de la présence ou de l'absence de divers identificateurs dans les fichiers, nous avons retiré certains des identificateurs les plus importants que les fichiers contiennent individuellement ou en groupe. Nous avons répété la recherche d'enregistrements indiquant lesquels des mineurs étaient décédés, parfois en utilisant tous les identificateurs, parfois en n'en utilisant pas certains.

La méthode utilisée pour procéder au couplage probabiliste des données en vue d'établir le décès aboutit à l'attribution d'un poids total. Celui-ci est la somme des poids attribués selon qu'il y a concordance, concordance partielle ou non-concordance des éléments d'information comparés. Un seuil est établi au-delà duquel le couplage est accepté. Pour les besoins de notre étude, nous avons défini le poids constituant le seuil optimum comme étant le niveau auquel les faux cas positifs et les faux cas négatifs se rapprochent le plus en nombre. Avec ce seuil optimum, l'erreur totale (c.-à-d. la somme des faux cas positifs et des faux cas négatifs) tend à être réduite, autant que possible, au minimum.

Les résultats de la recherche des enregistrements établissant le décès sont présentés au tableau 2. Il ressort clairement que la date de naissance au complet est un élément d'information particulièrement important, suivie des prénoms au complet. En fait, même un identificateur d'une importance apparemment moindre comme le deuxième nom ou l'initiale, qu'on ne trouve d'ailleurs pas souvent dans les documents d'enquête, a une influence marquée sur le résultat de la recherche. Un identificateur aussi peu courant que le nom de jeune fille de la mère joue un rôle considérable si les identificateurs plus courants sont absents (comparer les lignes 5 à 8 aux lignes 1 à 4 du tableau 2). Le tableau

3 montre ce qui arrive aux taux d'erreur lorsque des identificateurs sont retirés de l'enregistrement de départ et lorsque des identificateurs sont indifféremment présents ou absents.

Pourquoi n'obtient-on pas plus facilement de meilleurs résultats lorsqu'on fait une recherche en vue d'établir si des personnes sont décédées? La réponse est que tout dépend de la qualité et de la disponibilité des variables d'identification présentes dans les documents originaux. Certains renseignements permettant d'identifier les personnes ne sont pas consignés correctement. Soulignons aussi que les données présentées ici concernent uniquement des mineurs de sexe masculin.

Tableau 2

Incidence sur les taux d'erreur du retrait de certains identificateurs des enregistrements "hybrides"[1]

No du passage	Date de Naissance			Prénoms						Seuil optimal	Nombre de cas			Nombre total de cas erronés par rapport à 705 paires bien assorties (%)
				I	R	I	R				Faux	Faux	Total	
	A	M	J	1	1	2	2	NJFM	LN		Positif	Négatif	erronés	
	Retrait des données sur la date de naissance													
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
2	+	+	-	+	+	+	+	+	.	46 +	12	12	24	3.4
3	+	-	-	+	+	+	+	+	.	50 +	23	23	46	6.5
4	-	-	-	+	+	+	+	+	.	37 +	45	45 + 3*	93	13.2
5	+	+	+	+	+	+	+	-	.	37 +	14	14	28	4.0
6	+	+	-	+	+	+	+	-	.	25 +	18	19	37	5.2
7	+	-	-	+	+	+	+	-	.	24 +	33	34 + 1	68	9.6
8	-	-	-	+	+	+	+	-	.	4 +	76	78 + 9*	163	23.1
	Retrait des données sur le nom													
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
9	+	+	+	+	+	-	-	+	.	45 +	15	14	29	4.1
10	+	+	+	+	-	-	-	+	.	45 +	20	20	40	5.7
5	+	+	+	+	+	+	+	-	.	37 +	14	14	28	4.0
11	+	+	+	+	+	-	-	-	.	19 +	20	21 + 1*	42	6.0
12	+	+	+	+	-	-	-	-	.	7 +	25	25 + 1*	51	7.2
	Retrait du nom de jeune fille de la mère													
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
5	+	+	+	+	+	+	+	-	.	37 +	14	14	28	4.0
	Retrait du lieu de naissance													
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
13	+	+	+	+	+	+	+	+	-	49 +	8	8	16	2.3

Tableau 2

Incidence sur les taux d'erreur du retrait de certains identificateurs des enregistrements "hybrides"[1] (suite)

No du passage	Date de Naissance A	M	J	Prénoms I 1	R 1	I 2	R 2	NJFM	LN	Seuil optimal	Nombre de cas Faux Positif	Faux Négatif	Total erronés	Nombre total de cas erronés par rapport à 705 paires bien assorties (%)
	Retrait de plusieurs identificateurs													
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
3	+	-	-	+	+	+	+	+	.	50 +	23	23	46	6.5
15	+	-	-	+	+	-	-	+	.	43 +	41	41 + 1*	83	11.8
16	+	-	-	+	+	-	-	+	-	41 +	40	41 + 2*	83	11.8
17	+	-	-	+	+	-	-	-	.	1 +	70	66 + 5*	141	20.0
20	+	-	-	+	-	-	-	-	.	- 14 +	98	96 + 8*	202	28.6
18	-	-	-	+	+	-	-	+	.	21 +	87	89 + 10*	186	26.4
19	-	-	-	+	+	-	-	-	.	- 19 +	144	143 + 32*	319	45.2
14	+	+	+	+	+	+	+	-	-	36 +	13	13 + 1*	27	3.8

* Voir les notes concernant les tableaux 2 et 3

[1] Uniquement en fonction des paires bien assorties, les deux enregistrements de la paire ayant la série complète d'identificateurs (date de naissance au complet, deux prénoms au complet et nom de jeune fille de la mère). Le lieu de naissance était présent ou absent dans les deux enregistrements. Les pourcentages sont calculés par rapport à 705 paires bien assorties.

Explication des titres de colonne:

I	=	Initiales du premier (I1) ou du second (I2) prénom
R	=	Reste du premier (R1) ou du second (R2) prénom
NJFM	=	Nom de jeune fille de la mère
LN	=	Lieu de naissance
AMJ	=	Année, mois, jour de naissance

Explication des signes:

Identificateur:	Retiré	(-)
	Présent	(+)
	Présent ou absent	(.)

Tableau 3

Incidence sur les taux d'erreur du retrait de certains identificateurs des enregistrements "hybrides"[1]

No du passage	Date de Naissance A	M	J	Prénoms I 1	R 1	I 2	R 2	NJFM	LN	Seuil optimal	Nombre de cas Faux Positif	Faux Négatif	Total erronés	Nombre total de cas erronés par rapport à 705 paires bien assorties (%)
	Retrait des données sur la date de naissance													
1	.	.	.	.	.	.	.	.	.	50 +	54	55	109	4.9
2	.	.	-	.	.	.	.	.	.	40 +	82	81 + 1*	164	7.3
3	.	-	-	.	.	.	.	.	.	38 +	125	123 + 10*	258	11.5
4	-	-	-	.	.	.	.	.	.	16 +	229	233 + 41*	503	22.4
	Retrait des données sur le nom													
1	.	.	.	.	.	.	.	.	.	50 +	54	55	109	4.9
5	.	.	.	.	.	-	-	.	.	42 +	61	63 + 1*	125	5.6
6	.	.	.	.	-	-	-	.	.	35 +	79	81 + 1*	161	7.2
	Retrait d'autres identificateurs													
7	.	.	.	.	.	.	.	.	-	40 +	45	46	91	4.1
8	.	.	.	.	.	.	.	-	.	35 +	52	51	103	4.6

* Voir les notes concernant les tableaux 2 et 3

[1] En fonction de l'utilisation de tous les 30,000 enregistrements de départ susceptibles d'être appariés. Les pourcentages sont calculés par rapport à 2,243 paires bien assorties.

Explication des titres de colonne:

I	=	Initiales du premier	(I1)	ou du second	(I2) prénom
R	=	Reste du premier	(R1)	ou du second	(R2) prénom
NJFM	=	Nom de jeune fille de la mère			
LN	=	Lieu de naissance			
AMJ	=	Année, Mois, Jour de naissance			

Explication des signes:

Identificateur: Retiré (-); Présent (+);Présent ou absent (.)

Notes concernant les tableaux 2 et 3

1.* Dans les tableaux 2 et 3, les "faux négatifs" accompagnés d'un astérisque sont le résultat du couplage d'enregistrements de départ susceptibles d'être appariés et de mauvais enregistrements. Les poids attribués à ces paires d'enregistrements se situent normalement en-deçà du seuil optimum; il ne s'agit donc pas de paires bien appariées dont le poids serait en-deçà du seuil optimal et qui se trouveraient ainsi classées comme des "faux négatifs".

2. Le seuil "optimal" est défini comme étant celui où les nombres de cas "faux positifs" et de cas "faux négatifs" sont presque égaux en valeur absolue.

3. On estime que l'opération de couplage a été réussie pour 2,243 des 30,000 enregistrements se rapportant aux mineurs, grâce à l'utilisation des enregistrements "hybrides". Sur les 2,243 paires d'enregistrements obtenues, 705 contenaient de part et

d'autre la série complète d'identificateurs. Les résultats présentés dans le tableau 2 sont basés sur les 705 paries d'enregistrements contenant la série complète d'identificateurs et les résultats présentés dans le tableau 3, sur les 2,243 paries bien assorties, dont l'enregistrement de départ était un des 30,000 enregistrements susceptibles d'être appariés.

4. FICHIERS UTILISÉS

Dans le fichier de la Commission des accidents du travail, il y a environ 50,279 enregistrements se rapportant aux mineurs ontariens de la cohorte à l'étude. Exactement 30,000 de ces enregistrements contenaient un NAS valide permettant le couplage avec le fichier principal des NAS. Les enregistrements provenant de ces deux sources et les enregistrements de la base de données sur la mortalité au Canada qui contiennent le nom et des renseignements de ce genre peuvent éventuellement faire l'objet d'un couplage.

L'état civil des 30,000 mineurs dont le NAS était connu (en l'occurrence, le fait de savoir lesquels étaient vivants et lesquels étaient décédés) avait été établi auparavant par un processus minutieux en trois étapes: a) recherche en vue d'établir quelles personnes étaient décédées, b) recherche portant sur les dossiers fiscaux de 1977-1978 en vue d'établir quelles personnes étaient vivantes et c) vérification manuelle des contradictions apparentes afin d'en découvrir la cause et de connaître l'état réel des personnes en question.

D'après les résultats de la recherche visant à établir quels membres de la cohorte étaient décédés entre 1950 et 1977, entreprise au cours de l'étape intermédiaire de comparaison de l'opération de couplage à l'aide du Système itératif général de chaînage d'articles, un fichier relativement restreint d'enregistrements faisant état des décès avait été sauvegardé; on avait conservé tous les enregistrements (voir le fichier "Prod 7-DATB" du tableau 4) le moindrement susceptibles d'être rapprochés des enregistrements se rapportant aux mineurs, quels que soient les résultats éventuels de ce rapprochement. Comme le NAS a été introduit en 1964, on s'est servi du fichier pour les décès survenus à partir de cette année. Les enregistrements devaient se conformer au système de codes phonétiques connu sous le sigle NYSIIS (New York State Identification and Intelligence System phonetic code), lequel pour être inclus dans ce fichier réduit.

Tableau 4

Nombre d'enregistrements de la CAT concernant des mineurs et nombre d'enregistrements faisant état des décès ayant servi à l'évaluation de l'utilité du NAS

Enregistrements	Total
Enregistrements de la CAT concernant des mineurs (hommes)	
Nombre total dans le fichier	50,279
Nombre total contenant un NAS valide	30,000
État civil des 30,000 mineurs dont l'enregistrement contenait le NAS:	
Vivant - confirmé	26,736
Décès établi par couplage (1964-1977) confirmé	2,254
Recherche infructueuse	1,010
Enregistrements parcourus pour établir le décès (hommes)	
Nombre approximatif total dans la BDM (1964-1977)	1,300,000
Nombre total dans le fichier "prod7-DATB" (1950-1977)	46,679
Nombre total dans le fichier "prod7-DATB" (1964-1977)	35,251

5. MÉTHODE

Pour réaliser notre étude, nous avons eu recours à quatre fichiers, un fichier faisant état des décès et trois ensembles d'enregistrements se rapportant aux mineurs ontariens. C'est à partir de ces derniers qu'on a entrepris la recherche visant à déterminer quelles personnes étaient décédées. Voici plus précisément en quoi consistaient ces fichiers:

1. Enregistrements tirés de la BDMC faisant état des décès pour la période 1964-1977 et ayant été retenus après la phase de comparaison de la recherche antérieure;
2. Enregistrements de la CAT contenant les NAS;
3. Enregistrements provenant du fichier principal des NAS dont les NAS correspondaient à ceux des enregistrements de la CAT;
4. Enregistrements hybrides composés d'éléments provenant du fichier de la CAT et de celui des NAS.

Sur les 30,000 enregistrements de la CAT contenant les NAS, seulement 2,254 concernaient des mineurs dont le décès avait été établi. Cependant, les 30,000 ont servi à l'essai, car il était possible que certains fassent l'objet d'un couplage erroné. Pour chaque enregistrement de la CAT dont on s'est servi, on est allé chercher l'enregistrement correspondant dans le fichier des NAS. Ces enregistrements ont été utilisés soit séparément, soit ensemble.

Le plan était de répéter la recherche visant à établir quelles personnes étaient décédées en se servant:

1. seulement des identificateurs de la CAT;
2. seulement des identificateurs du fichier des NAS;
3. des identificateurs contenus dans les enregistrements hybrides.

Il avait été convenu que les enregistrements hybrides contiendraient le nom de jeune fille de la mère, tiré des enregistrements des NAS car ceux de la CAT ne l'avaient pas, et le code correspondant au lieu de naissance, tiré des enregistrements de la CAT car il n'était pas dans ceux du NAS. Lorsque les renseignements permettant d'identifier une personne (nom de famille, année, mois et jour de naissance ainsi que deux prénoms) n'étaient pas inscrits de la même façon dans les enregistrements provenant des deux sources, deux enregistrements distincts ont été créés. Lorsque les six éléments d'information étaient identiques, un seul enregistrement était conservé. Les renseignements étaient différents dans près de 60% des cas.

6. RÉSULTATS

Le but premier de l'étude était d'apporter une réponse à une question simple, en l'occurrence: "L'utilisation d'identificateurs personnels complets comme ceux qui figurent dans le fichier principal des NAS augmente-t-elle de façon marquée les chances de trouver les enregistrements indiquant quelles personnes sont décédées?" La réponse simple à cette question est que l'utilisation des identificateurs du fichier des NAS augmente de beaucoup les chances de trouver les bons enregistrements dont on connait d'avance l'existence (voir les tableaux 5 et 6). En se servant uniquement des identificateurs de la CAT, on a complètement raté 56 des 2,254 bons enregistrements dont l'existence était connue. En utilisant en plus des identificateurs de la CAT ceux des NAS, on n'en a ratés seulement 7 (c.-à-d. qu'on est passé d'un pourcentage d'échec de 2.5 à un pourcentage de 0.3).

Il faut choisir des poids qui constituent un seuil au-delà duquel le couplage peut se faire. Comme nous l'avons dit plus haut, nous avons choisi un seuil "optimal", c'est-à-dire un niveau auquel les faux positifs et les faux négatifs se rapprochent le plus en nombre. Pour trouver ce seuil, nous avons classé les paires d'enregistrements formées (c.-à-d. les paires constituées de l'enregistrement se rapportant à un mineur et de l'enregistrement qui avait le plus de chances d'être le bon) par ordre décroissant de poids total. Nous les avons regroupées par tranches de dix unités de poids et par tranches plus petites pour les poids se rapprochant de zéro. À partir de cette répartition, il est possible de choisir un seuil optimum à utiliser pour chacune des trois façons d'effectuer la recherche (tableau 7). Les trois seuils varient selon la quantité de renseignements dont on dispose pour identifier les personnes. Ils ont également tendance à se situer un peu au-dessus du niveau où, théoriquement, le nombre de faux positifs est égal au nombre de faux négatifs (probabilité 50-50 ou poids total = zéro).

On peut déterminer à l'aide de ce poids optimal le nombre de paires d'enregistrements mal assorties qui sont acceptées parce que leur poids total se situe au-delà du seuil et le nombre de paires d'enregistrements susceptibles d'être bien assorties qui sont rejetées parce que leur poids total se situe en-deçà du seuil.

Les tests visant à déterminer dans quelle mesure les chances de trouver le bon enregistrement varient selon qu'on n'utilise pas divers identificateurs personnels tendent à démontrer l'utilité d'avoir recours à une source d'enregistrements où les identificateurs sont complets et de bonne qualité.

Tableau 5

Nombre de "meilleurs" paires selon le genre d'enregistrements de départ — décès seulement

(Les recherches se rapportant toutes aux 2,254 mineurs dont on sait qu'ils sont morts pendant la période 1964-1977.)

Genre de paire obtenue	Nombre d'enregistrements de départ		
	Fichier de la CAT	Fichier des NAS	Fichier hybrides
Chiffres			
C'était la bonne paire et la "meilleure"	2,193	2,238	2,243
C'était la bonne paire, mais pas la "meilleure"	5	3	4
Aucune paire n'a été constituée	56	13	7
Nombre total d'enregistrements de départ	2,254	2,254	2,254
Pourcentages			
C'était la bonne paire et la "meilleure"	97.29	99.29	99.51
C'était la bonne paire, mais pas la "meilleure"	.22	.13	.18
Aucune paire n'a été constituée	2.48	.58	.31
Nombre total d'enregistrements de départ	100.00	100.00	100.00

NOTA: La "meilleure" paire est, par définition, celle qui a la plus forte probabilité calculée (c.-à-d qui a le poids le plus élevé) d'être une bonne paire.

Tableau 6

Nombre de bonnes paires obtenues selon le genre d'enregistrements de départ (d'après les mêmes données que le tableau 5.)

Bonnes paires obtenues à partir d'un enregistrement: Du fichier de la CAT	Du fichier des NAS	Du fichier Hybride	Nombre de bonnes paires*	Pourcentage de bonnes paires*
Oui	Oui	Oui	2,192	97.25
Non	Oui	Oui	49	2.17
Oui	Non	Oui	6	.27
Non	Non	Non	7**	.31
Total			2,254	100.00
Combiné				
Oui	--	--	2,198	97.52
--	Oui	--	2,241	99.42
--	--	Oui	2,247	99.69

* Toutes les bonnes paires sont comptées, même celles, peu nombreuses, dont le poids était inférieur à celui d'une autre paire possible qui n'était pas la bonne.

** Les 7 paires qui auraient dû être formées ne l'ont pas été quel que soit l'enregistrement de départ utilisé parce que:
Dans 4 cas, le code NYSIIS correspondant au nom de famille n'était pas le même
Dans 2 cas, le décès a été enregistré trop tard
Dans 1 cas, le décès est survenu à l'extérieur du pays (en Allemagne).

Tableau 7

Paires faussement positives et paires faussement négatives, selon un seuil optimal

Nombre d'enregistrements de départ (Susceptibles d'être appariés)		Seuil optimal	Faux positif		Faux négatif		Total	
			No.	(%)	No.	(%)	No.	(%)
En fonction des enregistrements de départ ayant donné lieu à des paires bien assorties (Sans tenir compte des enregistrements pouvant être appariés qui ne l'ont pas été au cours de cette opération.)								
Fichier de la CAT	(2198)	+15	80	(3.6)	83	(3.8)	163	(7.4)
Fichier des NAS	(2241)	+16	55	(2.5)	54	(2.4)	109	(4.9)
Fichier hybride	(2247)	+50	55	(2.4)	54	(2.4)	109	(4.9)
En fonction de l'ensemble des enregistrements de départ (En tenant compte des enregistrements pouvant être appariés qui ne l'ont pas été au cours de cette opération.)								
Fichier de la CAT	(2254)	+15	80	(3.5)	139	(6.2)	219	(9.7)
Fichier des NAS	(2254)	+16	55	(2.4)	67	(3.0)	122	(5.4)
Fichier hybride	(2254)	+50	55	(2.4)	61	(2.7)	116	(5.1)

7. UTILITÉ POUR LES RECHERCHES À VENIR

Les résultats de l'étude sont applicables aux études de mortalité par cohortes professionnelles qui pourraient être menées à l'avenir, notamment celles qui s'intéressent aux effets à long terme des rayonnements ionisants. Le couplage des données sera nettement meilleur si les identificateurs recueillis sont complets.

L'utilisation de fichiers "intermédiaires" pour faciliter le couplage s'est avérée particulièrement utile. Nous avons trouvé qu'il était pratique pour nos travaux de créer à partir de ces fichiers intermédiaires des enregistrement hybrides constitués de zones provenant de chaque fichier. Parmi les fichiers qui pourraient servir de fichier intermédiaire, mentionnons le registre des mariages, qui permettrait de retrouver les données se rapportant aux personnes dont le nom aurait changé au cours de la période d'observation.

Les personnes suivantes ont souvent besoin de ce genre de renseignements:

1. Celles qui travaillent à la révision de formulaires (p. ex. les travaux en cours concernant les actes d'état civil);
2. Celles qui font des recommandations relatives à la collecte de données ou effectuent cette collecte (p. ex. dans le cadre de recherches sur l'hygiène du travail et du milieu);
3. Celles qui conçoivent des questionnaires (p. ex. pour des enquêtes ou pour des études sur l'incidence du cancer);
4. Celles qui évaluent les biais pouvant se glisser dans l'analyse des résultats d'études (p. ex. le traitement des cas pour lesquels la recherche a été infructueuse dans les études de mortalité);
5. Celles qui évaluent la faisabilité et les taux d'erreur éventuels d'une étude selon la présence de certains identificateurs;
6. Celles qui s'interrogent sur l'incidence que l'utilisation d'un fichier intermédiaire pour faciliter le couplage des enregistrements peut avoir sur les chances de trouver les enregistrements qui les intéressent.

REMERCIEMENTS

Les auteurs tiennent à remercier M. Newcombe pour les avoir aidés à préparer cette communication.

BIBLIOGRAPHIE

Fair, M.E., Newcombe, H.B., Lalonde, P., et Poliquin, C. (1988). "Alive" searches as complementing death searches in the epidemiological follow-up of Ontario miners. Rapport de recherche rédigé à Statistique Canada en vertu d'un contrat conclu avec la Commission de contrôle de l'énergie atomique. (En voie de publication, Commission de contrôle de l'énergie atomique, 1988a).

Fair, M.E., Newcombe, H.B., et Lalonde, P. (1988). "Improved mortality searches for Ontario miners using Social Insurance Index identifiers". Rapport de recherche rédigé à Statistique Canada en vertu d'un contrat conclu avec la Commission de contrôle de l'énergie atomique. (En voie de publication, Commission de contrôle de l'énergie, 1988b).

Hill, T., et Pring-Mill, F. (1985). "Generalized iterative record linkage system". Dans *Record Linkage Techniques - 1985. Proceedings of the Workshop on Exact Matching Methodologies*, (B. Kills et W. Alvey, éditeurs), Dept. of the Treasury, Internal Revenue Service, 327-333.

Howe, G.R., et Lindsay, J. (1981). "A generalized iterative record linkage computer system for use in medical follow-up studies". Dans *Computers and Biomedical Research*, 14, 327-340.

Last, J.M. (1986). "Individual privacy and health information: an ethical dilemma?" Dans *Revue canadienne de santé publique*, 77, 168-169.

Monson, R.A. (1980). *Occupational Epidemiology*. Boca Raton, Florida: CRC Press, Inc..

Muller, J., Wheeler, W.C., Gentleman, J.F., Suranyi, G., et Kusiak, R.A. (1983). "Study of Mortality of Ontario Miners, 1955-1977 Part I". Ministère du Travail de l'Ontario, Direction des études spéciales et des services de santé, 400, avenue University, Toronto (Ontario) M7A 1T7.

Redmond, C.K., Smith, E.M., Lloyd, J.W., et Rush, H.W. (1969). "Long-term mortality study of steelworkers. III Follow-up". Dans *Journal of Occupational Medicine*, 11: 513-521.

Smith, M.E., et Newcombe, H.B. (1980). "Automated follow-up facilities in Canada for monitoring delayed health effects". Dans *American Journal of Public Health*, 70, 1261-1268.

Smith, M.E., et Silins, J. (1981). "Generalized iterative record linkage system". Dans *Social Statistics Section, Proceedings of the American Statistical Association*, 128-137.

Smith, M.E. (1981). "Long-term medical follow-up in Canada. Dans *Quantification of Occupational Cancer*. Banbury Report N°. 9. Cold Spring Harbor, New York, Cold Spring Harbor Laboratory, 675-688.

Smith, M.E. (1986). "Future needs and directions for computerized record linkage in health research in Canada. a) Future study plans and studies using the Canadian mortality data base". Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (G.R. Howe et R.A. Spasoff, éditeurs) University of Toronto Press, 211-230 et 249-257. Adresser les demandes au: Département d'épidémiologie et de médecine sociale, Université d'Ottawa, Faculté des sciences de la santé, Ottawa (Ontario), K1H 8M5.

Symons, M.J., et Taulbee, M.J. (1984). "Statistical evaluation of the risk of cancer mortality among industrial populations". Dans *Statistical Methods for Cancer Studies*, 270 Madison Avenue, New York, Marcel Dekker Inc., 25-90.

Working Group on Health and Safety of Laboratory Workers (1986). "Health and safety of lab workers in Canada - A code of Conduct". *Occupational Health in Ontario* 7,2, 92-108.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

MÉTHODES DE CALCUL UTILISÉES POUR L'APPLICATION DU MODÈLE D'APPARIEMENT DES ENREGISTREMENTS DE FELLEGI-SUNTER AUX LISTES D'ENTREPRISES

WILLIAM E. WINKLER[1]

RÉSUMÉ

Soit **AXB** l'espace produit de deux ensembles **A** et **B**, qui est formé de **concordances** (paires dont les éléments représentent la même entité) et de **non-concordances** (paires dont les éléments représentent des entités différentes). Les règles d'appariement divisent **AXB** en **liens** (concordances désignées), en **cas indéterminés** (paires pour lesquelles nous reportons une décision) et en **non-liens** (non-concordances désignées). Suivant un intervalle fixe pour les taux d'erreur, Fellegi et Sunter (1969) ont défini une règle d'appariement optimale, c'est-à-dire une règle qui réduit au minimum l'ensemble des cas indéterminés. L'optimalité dépend de la connaissance de certaines probabilités de sélection composées utilisées dans un rapport de vraisemblance déterminant. En appliquant le modèle d'appariement des enregistrements, on pose souvent des hypothèses qui permettent d'estimer les probabilités de sélection composées. Si les hypothèses ne sont pas satisfaites, il se peut qu'une méthode d'appariement qui utilise des estimations calculées suivant ces hypothèses ne soit pas optimale. Dans cet article, nous analysons des méthodes qui permettent de modifier les règles d'appariement lorsque les hypothèses ne sont pas valides. À cette fin, nous faisons une analyse empirique de listes d'entreprises pour lesquelles l'authenticité des concordances a été vérifiée. Le nombre de cas indéterminés que produisent les méthodes de calcul habituelles et les méthodes révisées peut varier selon les échantillons. Cette relation est analysée au moyen de méthodes "bootstrap" (Efron, 1987).

1. INTRODUCTION

Cet article contient une analyse des méthodes de calcul utilisées pour l'application du modèle d'appariement des enregistrements de Fellegi-Sunter aux listes d'entreprises.

Étant donné deux listes, nous voulons utiliser des identificateurs qui nous permettront de distinguer les paires d'enregistrements dont les éléments se rattachent à la même entité (**concordances**) et celles dont les éléments se rattachent à des entités différentes (**non-concordances**). Nous chercherons donc à définir une règle d'appariement qui nous permettra de diviser l'espace produit de paires en trois groupes: **liens** (concordances désignées), **cas indéterminés** (paires pour lesquelles une décision est reportée) et **non-liens** (non-concordances désignées).

[1] William E. Winkler, Statistical Research Division, U.S. Bureau of the Census.

Selon un intervalle fixe pour le nombre de concordances et de non-concordances erronées, Fellegi et Sunter (1969, théorème) élaborent une méthode qui, en théorie, réduit au minimum le nombre de cas indéterminés. L'optimalité dépend de la connaissance de certaines probabilités de sélection composées utilisées dans un rapport de vraisemblance fondamental.

L'objet fondamental de l'application du modèle est d'identifier automatiquement les liens et les non-liens à l'aide de l'ordinateur. On procède ensuite à une analyse manuelle afin de recueillir (parfois) plus de renseignements sur les cas indéterminés et de les identifier ensuite comme liens ou non-liens.

Le modèle de Fellegi-Sunter est particulièrement utile en ce qu'il permet d'apporter progressivement des corrections au logiciel pour ce qui a trait à la description et à la comparaison des renseignements qui doivent entrer dans la procédure d'appariement des enregistrements. Par exemple, supposons que la paire ci-dessous soit définie comme un

NOM	RUE	VILLE	ÉTAT	CODE POSTAL
Zabrinsky, Robert A	16 Sycamore St	Dayton	Ohio	53342
Zabrinsky, R	167 W Sycamore	Dayton	Ohio	53342

cas indéterminé. Cette paire a été obtenue à l'aide d'une méthode qui utilise une abréviation du nom de famille. Avec les logiciels existants, les autres éléments de l'adresse ne peuvent être comparés qu'à raison d'un caractère à la fois.

S'il existait un logiciel qui permettrait de comparer "SYCAMORE" avec "SYCAMORE", la méthode informatique qui intègre les renseignements supplémentaires permettrait peut-être d'identifier la paire comme un lien. Essentiellement, nous cherchons à développer un logiciel qui imitera le processus décisionnel que suivent les humains.

Newcombe et coll. (1983) montrent jusqu'à quel point les méthodes informatiques peuvent surpasser les méthodes manuelles. Ils prennent l'exemple de fichiers de population renfermant de grandes quantités de renseignements qui peuvent être efficacement comparés.

De nombreuses listes contiennent toutefois des renseignements qui, par leur forme, se prêtent mal à une comparaison efficace. Nous voulons donc explorer des méthodes générales qui nous permettent de prendre des décisions toujours meilleures. Étant donné un intervalle fixe pour les taux d'erreur, de **meilleures** règles d'appariement auront pour effet de réduire l'ensemble des cas indéterminés. Nous pouvons obtenir de telles règles lorsqu'il est possible, grâce à un logiciel supplémentaire, d'utiliser une plus forte proportion des renseignements disponibles ou de modifier les méthodes de calcul si les hypothèses ne sont pas parfaitement valides.

Dans les sections qui suivent, nous allons présenter des méthodes qui permettent d'appliquer le modèle de Fellegi-Sunter aux listes d'entreprises. Plus précisément, nous allons faire une analyse de la taille des régions de cas indéterminés compte tenu d'un intervalle fixe pour les taux d'erreur. Au point de vue de l'application, nous allons utiliser des paires de listes pour lesquelles l'authenticité ou la fausseté des appariements a été vérifiée.

Bien que les méthodes de calcul que nous allons utiliser se rattachent proprement aux genres de listes que l'on trouve à l'Energy Information Administration (EIA), nous allons souligner les ressemblances générales qui existent entre ces méthodes et celles utilisées dans d'autres cas d'appariement d'enregistrements.

La deuxième section de cet article est divisée en quatre sous-sections. La première contient une description de la base de données et des sous-zones particulières qui sont comparées. La seconde renferme une description sommaire du modèle de Fellegi-Sunter et expose les aspects qui rendent ce modèle facile à modifier pour des applications précises. La troisième met en lumière les hypothèses générales et les méthodes de calcul utilisées dans les projets d'appariement d'enregistrements au Canada et aux États-Unis. Elle expose aussi en détail les méthodes de calcul qui se rattachent spécifiquement à l'objet de cet article.

La quatrième sous-section décrit les méthodes d'évaluation. La méthode d'évaluation fondamentale consiste à suivre l'évolution de la taille de la région des cas indéterminés lorsque divers genres de règles d'appariement sont appliquées suivant un intervalle fixe pour les taux d'erreur. La taille des régions de cas indéterminés est une statistique qui peut dépendre des échantillons ayant servi à uniformiser les règles d'appariement. La distribution de cette statistique est évaluée au moyen de la méthode bootstrap d'Efron (1987, 1982, 1979).

La troisième section de cet article renferme les résultats de l'analyse. Dans la quatrième section, nous analysons d'autres genres de comparaisons et discutons de l'application de critères de groupage additionnels. Enfin, nous concluons l'article par un résumé.

2. BASE DE DONNÉES, MODÈLE D'APPARIEMENT, MÉTHODES DE CALCUL ET D'ÉVALUATION

Cette section décrit la base de données, le modèle d'appariement des enregistrements de Fellegi-Sunter de même que les méthodes de calcul et d'évaluation.

2.1 Base de données

La description de la base de données se fait en deux étapes. Nous allons d'abord décrire les caractéristiques générales de la base, puis énumérer les sous-zones particulières qui font l'objet d'une comparaison.

2.1.1 Description générale

La base de données renferme 54,850 enregistrements, qui représentent chacun une entreprise, et 3,050 doubles, pour un total de 57,900 enregistrements. Une paire d'enregistrements formée d'une entreprise et du double correspondant est désignée comme une concordance; toutes les autres paires sont des non-concordances.

La base de données a été construite à l'aide de 11 listes de l'EIA et de 47 listes fournies par les États et l'industrie; ces listes renfermaient 176,000 enregistrements. Les doubles ont été identifiés par des techniques élémentaires de rappel (le numéro de téléphone est parfois indiqué) et de sondage.

Cet article examine comment on peut modifier le modèle de Fellegi-Sunter de manière à obtenir une description précise des doubles qui sont difficiles à identifier. Nous ne considérerons pas dans cet article les doubles faciles à identifier (c'est-à-dire ceux pour lesquels il n'y a généralement pas de problème de concordance des caractères).

Voici un exemple d'un double difficile à identifier:

NOM	RUE	VILLE	ÉTAT	CODE POSTAL
Zabrinsky Fuel	16 W Sycamore St	Dayton	OH	53315
Zabrinky Cmpny	167 Sycamere St	Springfield	OH	53315

En examinant le deuxième enregistrement, nous constatons que "Zabrinsky" et "Sycamore" sont mal écrits, que le terme "Cmpny" est une abréviation inusitée et que le code postal pour Springfield (Ohio), une banlieue de Dayton, est 53315.

2.1.2 Sous-zones particulières comparées

Quatre ensembles de sous-zones particulières font l'objet d'une comparaison dans chaque paire d'enregistrements. Le premier ensemble est celui que l'on peut obtenir par des comparaisons simples de sous-séquences. On forme la zone LM-NOM, qui figure dans l'énumération ci-dessous, en classant les mots de la zone NOM par ordre décroissant de longueur et en scindant les liaisons par un tri alphabétique.

ZONE	COLONNES DE SOUS-ZONES COMPARÉES
NOM	1-4, 5-10, 11-20, 21-30
RUE	1-6, 7-15, 16-30
CODE POSTAL	1-3, 4-5
VILLE	1-5, 6-10, 11-15
ÉTAT	1-2
TÉLÉPHONE	1-3, 4-6, 7-10
LM-NOM	1-4, 5-10, 11-20, 21-30

Le deuxième ensemble est le résultat des quatre comparaisons des deux plus longs mots de la zone NOM. Là encore, les liaisons sont scindées par un tri alphabétique.

Les deux derniers ensembles sont formés de sous-ensembles des zones RUE et NOM qui sont définis par des logiciels très perfectionnés. Par exemple, ZIPSTAN, qui est un logiciel du Census Bureau (Département du commerce des États-Unis, 1978b), sert à définir des sous-zones correspondantes de la zone RUE. Ces sous-zones sont le numéro civique, les préfixes 1 et 2, le nom de la rue, les suffixes 1 et 2 et l'unité. Les préfixes sont des directions comme "East" ou "North". Les suffixes sont des génériques comme "Street" ou "Road". L'unité représente des identificateurs comme le numéro d'appartement ou le numéro de pièce.

Le module NSKGEN5 du logiciel utilisé dans le Registre des entreprises du Canada (Statistique Canada, 1984, 1982) sert à définir des sous-zones correspondantes de la zone NOM. NSKGEN5 crée trois groupes de mots. Le premier groupe comprend trois abréviations, dont la première correspond au NOM DE FAMILLE si celui-ci est indiqué. Le deuxième groupe est formé de deux mots, dont le premier correspond au prénom. Le troisième groupe est constitué en fait d'un seul mot qui est obtenu par l'enchaînement et l'abréviation de mots contenus dans la zone NOM. On trouvera plus de détails à ce sujet dans Winkler (1987a) ou dans Statistique Canada (1984, 1982).

Intuitivement, une comparaison donnée de sous-zones correspondantes de la zone NOM peut dépendre de la comparaison de différentes sous-zones correspondantes de la zone NOM.

2.2 Modèle de Fellegi-Sunter

Le modèle de Fellegi-Sunter utilise une approche théorique décisionnelle qui confirme la validité des principes mis en application par Newcombe (Newcombe et coll., 1959). Pour donner un aperçu du modèle, nous le décrivons sous forme de paires ordonnées dans un espace produit. Notre description suit de près celle de Fellegi et Sunter (1969, p. 1184-1187).

Soient deux populations **A** et **B** dont les éléments seront désignés respectivement par **a** et **b**. Nous supposons que certains éléments sont communs aux deux populations.

Par conséquent, l'ensemble de paires ordonnées

$$AXB = \{(a,b): a\varepsilon A,\ b\varepsilon B\}$$

est l'union de deux ensembles disjoints, soit l'ensemble des **correspondances**

$$M = \{(a,b): a=b,\ a\varepsilon A,\ b\varepsilon B\}.$$

et l'ensemble des **non-correspondances**

$$U = \{(a,b): a\neq b,\ a\varepsilon A,\ b\varepsilon B\}.$$

Les enregistrements relatifs à **A** et à **B** sont désignés respectivement par $\alpha(a)$ et $\beta(b)$,. Le **vecteur de comparaison** τ associé aux enregistrements est défini par:

$$\tau[(\alpha(a),\beta(b))] \equiv \{\tau^1[(\alpha(a),\beta(b))],\tau^2[(\alpha(a),\beta(b))],\ldots,\tau^K[(\alpha(a),\beta(b))]\}.$$

Pour éviter toute confusion, nous désignerons la fonction τ sur **AXB** par $\tau(\alpha,\beta)$, $\tau(a,b)$, ou τ. L'ensemble des réalisations possibles de τ est désigné par Γ.

La probabilité conditionnelle de $\tau(a,b)$ si $(a,b)\varepsilon M$ est donnée par

$$m(\tau) \equiv P\{\tau[(\alpha(a),\beta(b))]\,|\,(a,b)\varepsilon M\}$$

$$\sum_{(a,b)\varepsilon M} P\{\tau[(\alpha(a),\beta(b))]\}.P[(a,b)\,|\,M].$$

De même, nous désignerons la probabilité conditionnelle de τ si $(a,b)\varepsilon U$ par $u(\tau)$.

Nous observons un vecteur d'information $\tau(a,b)$ associé à la paire (a,b) et voulons identifier une paire comme lien (ensemble A_1), cas indéterminé (ensemble A_2) ou non-lien (ensemble A_3). Posons L comme une règle d'appariement qui divise **AXB** en trois groupes: A_1, A_2, et A_3. Nous disons qu'une **erreur de type I** est commise si la règle L place $m\varepsilon M$ dans A_3 et qu'une **erreur type de II** est commise si L place $u\varepsilon U$ dans A_1. Fellegi et Sunter (1969) définissent une règle d'appariement L_0 optimale avec les ensembles correspondants A_1, A_2, et A_3; l'optimalité de cette règle se lit de la façon suivante:

THÉORÈME (Fellegi-Sunter, 1969). Soit L' une règle d'appariement avec les ensembles correspondants A_1', A_2', et A_3' de sorte que $P(A_3'\,|\,M) = P(A_3\,|\,M)$ et $P(A_1'\,|\,U) = P(A_1\,|\,U)$. Alors, $P(A_2\,|\,U) \leq P(A_2'\,|\,U)$ et $P(A_2\,|\,M) \leq P(A_2'\,|\,M)$.

Autrement dit, si L' et L_0 sont deux règles concurrentes qui ont les mêmes taux d'erreur de type I et de type II (lesquels sont des probabilités conditionnelles), alors la probabilité conditionnelle (pour l'ensemble U ou l'ensemble M) que l'on ne prenne pas de décision selon la règle L' est toujours supérieure à la probabilité conditionnelle que l'on ne prenne pas de décision selon la règle L_0.

De fait, la règle d'appariement de Fellegi-Sunter est optimale par rapport à n'importe quel ensemble Q de paires ordonnées dans **AXB** si nous définissons des probabilités d'erreur P_Q et une règle d'appariement L_Q qui dépendent de Q. Ainsi, nous pourrions peut-être définir des sous-ensembles de **AXB** auxquels nous appliquerions divers genres de renseignements en quantités variables.

Par exemple, si nous avons un ensemble de paires pour lesquelles le numéro de téléphone est indiqué, nous pourrions utiliser le numéro de téléphone et quelques caractères du nom pour définir les liens. Pour d'autres paires, nous pourrions devoir utiliser aussi les renseignements contenus dans les zones RUE et VILLE.

Des **critères de groupage** sont souvent à l'origine des ensembles Q de paires ordonnées auxquels est appliquée la règle d'appariement de Fellegi-Sunter. Les critères de groupage sont des indicatifs de tri qui servent à réduire le nombre de paires considérées. Au lieu de considérer toutes les paires contenues dans **AXB**, nous pourrions considérer seulement les paires dont les éléments ont en commun les trois premiers chiffres du code postal ou l'abréviation du nom de famille (pour autant que cette abréviation soit acceptable).

2.3 Méthodes de calcul

L'objectif premier de l'appariement d'enregistrements est de trouver des méthodes de calcul applicables qui utilisent efficacement les données contenues dans les fichiers, qui favorisent une mise à jour relativement simple de ces fichiers par des méthodes améliorées et qui peuvent être vérifiées.

Cette sous-section se divise en six parties. La première contient une description de la règle d'appariement générale du modèle de Fellegi-Sunter. La seconde contient une description de la version simplifiée des méthodes de calcul lorsqu'une hypothèse d'indépendance conditionnelle est posée.

Les méthodes de calcul décrites dans les deux premières parties ont été appliquées (à quelques différences près) aux fichiers de population du U.S.Bureau of Census (Kelley 1985a, 1985b et 1986; Département du commerce des États-Unis, 1978a), au U.S. National Death Index (Rogot et coll., 1983) et au California Automated Mortality Linkage System (Arellano 1985) et ont servi au Département de l'agriculture des États-Unis (Coulter 1977; Département de l'agriculture des États-Unis, 1979), à la U.S. Energy Information Administration (Winkler 1985b, 1987a) et à Statistique Canada (Smith et Silins 1981; Smith, Newcombe et Dewar 1983).

Dans la troisième partie, nous discutons de la validité de l'hypothèse d'indépendance conditionnelle. La quatrième partie renferme une observation sur des règles d'appariement équivalentes. Dans la cinquième partie, nous décrivons deux méthodes générales permettant de modifier les méthodes de calcul. Enfin, la sixième partie contient une description des méthodes de calcul utilisées précisément pour cet article.

2.3.1 Définition générale de la règle d'appariement

Pour comprendre pourquoi on utilise des méthodes de calcul particulières, considérons le rapport de vraisemblance suivant

$$R \equiv R[\tau(a,b)] = m(\tau)/u(\tau). \qquad (2.1)$$

Si, dans l'équation ci-dessus, le numérateur est positif et le dénominateur nul, nous attribuons au rapport une valeur arbitraire très élevée. La règle d'appariement de Fellegi-Sunter s'énonce alors comme suit:

Si $R > \text{UPPER}$, (a,b) est définie comme un lien.

Si $\text{LOWER} \leq R \leq \text{UPPER}$, (a,b) est définie comme un lien possible. (2.2)

Si $R < \text{LOWER}$, (a,b) est définie comme un non-lien.

Les bornes LOWER et UPPER sont déterminées par l'intervalle de taux d'erreur voulu.

2.3.2 Simplification suivant l'hypothèse de l'indépendance conditionnelle

En pratique, on simplifie le calcul de deux façons. La première consiste à poser l'hypothèse de l'indépendance conditionnelle de Fellegi et Sunter (1969):

$$m(\tau) = m_1(\tau^1).m_2(\tau^2)\ldots m_k(\tau^K) \text{ et}$$

$$u(\tau) = u_1(\tau^1).u_2(\tau^2)\ldots u_k(\tau^K)$$

où, pour $i = 1, 2, \ldots, K$

$$m_i(\tau^i) = P(\tau^i | (a,b)\varepsilon M) \text{ et}$$

$$u_i(\tau^i) = P(\tau^i | (a,b)\varepsilon U)$$

La deuxième façon de simplifier le calcul est d'utiliser une fonction du rapport défini en (2.1), qui se prête bien à des calculs. Nous avons choisi la fonction Log_2. Nous avons donc

$$W \equiv W(\tau) = \text{Log}_2[m(\tau)/u(\tau)] \qquad (2.3)$$

$$= W^1 + W^2 + \ldots + W^K,$$

où $W^1 \equiv \text{Log}_2 [m_i(\tau^i)/u_i(\tau^i)]$ pour $i = 1, 2, \ldots, K$. Nous appelons W le **poids de comparaison total** lié à une paire et W_i, $i = 1, 2, \ldots, K$, les **poids de comparaison individuels.**

Pour le reste de cet article, nous supposerons que chaque composante τ^i, $i = 1, 2, \ldots K$, de τ est une variable binaire (par exemple concordance/non-concordance) et nous définirons les événements de comparaison marginale par

$$B^i \equiv \{(a,b) | \tau^i(a,b) = \tau^i 0\}$$

pour une valeur donnée de $\tau^i 0$. Suivant l'hypothèse de l'indépendance conditionnelle, nous devons estimer 2K probabilités du genre

$$P(\tau\varepsilon B_i | M) \text{ and } P(\tau\varepsilon B_i | U), \; i = 1, 2, \ldots, K. \qquad (2.4)$$

Si nous avons un ensemble de paires pour lesquelles l'authenticité ou la fausseté des correspondances est vérifiée, nous devons diviser cet ensemble, pour chaque caractéristique de concordance B_i, $i = 1, 2, \ldots, K$, de manière à obtenir les quatre sous-ensembles définis par (2.4) avant de procéder à l'estimation.

En l'absence d'hypothèse, nous devons estimer 2.2^K-1 probabilités de sélection combinées (le numérateur et le dénominateur de la formule (2.1)) et diviser l'ensemble de paires pour lesquelles l'authenticité ou la fausseté des correspondances est vérifiée en 2.2^K-1 sous-ensembles. Même avec un petit nombre de comparaisons (disons six au maximum), nous pourrions ne pas être en mesure d'obtenir des échantillons suffisamment grands pour estimer avec précision les probabilités de sélection combinées.

Dans des applications antérieures, Newcombe et Kennedy (1962) ont fait plus de 60 comparaisons, Rogot et coll. (1983) en ont fait 11 et Winkler (1985b, 1987a) en a fait plus de 30.

Fellegi et Sunter (1969) ont proposé deux méthodes pour calculer des probabilités comme celles définies en (2.4). La première méthode renferme plusieurs hypothèses qui reposent sur une connaissance préalable des caractéristiques des fichiers. La seconde méthode permet de calculer directement les probabilités à partir des caractéristiques des fichiers et n'exige pas que l'on ait vérifié l'authenticité des correspondances tirées des échantillons.

2.3.3 Validité de l'hypothèse de l'indépendance conditionnelle

Fellegi et Sunter précisent que si l'hypothèse de l'indépendance conditionnelle n'est pas valide, on ne pourra plus interpréter d'une manière strictement probabiliste les estimations des poids calculés selon la formule (2.3). Autrement dit, la règle d'appariement du théorème de Fellegi-Sunter pourrait ne pas réduire au minimum le nombre de cas indéterminés. Néanmoins, Fellegi et Sunter croient à la solidité de leur modèle dans les cas où l'hypothèse de l'indépendance ne tiendrait plus.

Winkler (1985b) a montré que l'hypothèse de l'indépendance conditionnelle n'est pas valide pour des comparaisons simples de portions des zones NOM et RUE pour des listes d'entreprises. En utilisant les mêmes portions des zones NOM et RUE, Kelley (1986) a montré que l'hypothèse de l'indépendance conditionnelle n'était pas valide pour des fichiers de personnes. De plus, Kelley et Winkler ont montré, chacun de leur côté, que l'efficacité d'appariement dépendait beaucoup de l'ensemble de paires pour lequel des probabilités comme celles définies en (2.4) étaient calculées.

Suivant l'hypothèse de l'indépendance, la probabilité de sélection combinée équivaut au produit de probabilités comme celles définies en (2.4). Si nous avons un ensemble de paires pour lesquelles l'authenticité ou la fausseté des correspondances est vérifiée, il est alors possible de corriger les probabilités de sélection combinées pour les cas où l'hypothèse de l'indépendance ne tiendrait plus. Si les corrections sont appréciables, cela pourrait mettre en doute la solidité du modèle de Fellegi-Sunter.

2.3.4 Règles équivalentes

Nous constatons que pour chaque ensemble de comparaisons $B_1, B_2, \ldots, B_K$ (et l'espace de comparaisons correspondant $\Gamma = \{\tau(a,b): (a,b) \in AXB\}$), l'ensemble des poids calculés selon (2.3) engendre un ordre linéaire dans Γ, les liaisons étant ordonnées de façon arbitraire. Tout autre ensemble de nombres réels qui est appliqué à Γ et engendre le même ordre linéaire produira des règles d'appariement optimales comme celle définie en (2.2) **pour chaque intervalle de taux d'erreur** auquel s'applique la règle d'appariement originale de Fellegi-Sunter.

Si l'intervalle est fixe, nous n'avons qu'à appliquer les nombres réels R', les bornes LOWER' et UPPER' et les règles d'appariement du genre de celle définie en (2.2) de manière que les ensembles de liens et de non-liens concordent avec ceux qui sont dérivés de la règle optimale de Fellegi-Sunter. Nous avons donc beaucoup de latitude en ce qui concerne la manière d'appliquer les règles d'appariement.

2.3.5 Méthodes générales de correction

Deux méthodes générales de correction se rattachent aux méthodes de calcul des poids de comparaison individuels. La première consiste à subdiviser en plusieurs parties le sous-ensemble de paires de AXB pour lequel des poids de comparaison individuels sont calculés. On détermine la règle d'appariement en faisant en sorte que la règle fondamentale de Fellegi-Sunter corresponde uniquement aux divers sous-ensembles pour lesquels des poids sont calculés. Les poids de comparaison individuels peuvent varier largement selon les sous-ensembles.

La seconde méthode consiste à modifier les poids de comparaison individuels. Si, dans l'hypothèse de l'indépendance, nous considérons l'équation:

$$W \equiv \text{Log}_2(P(\tau\varepsilon B_i \cap B_2 \cap \ldots \cap B_K | M)/P(\tau\varepsilon B_1 \cap B_2 \cap \ldots \cap B_k | U))$$

$$= W^1 + W^2 + \ldots + W^K,$$

où $W^i \equiv \text{Log}_2\ (P(\tau\varepsilon B_j | M)/P(\tau\varepsilon B_j | U))$ pour i = 1, 2, et K, nous voulons trouver des méthodes flexibles qui permettent de corriger les W^i, i= 1, 2, ..., K, de telle sorte que leur somme donne de meilleures règles d'appariement.

S'il existe un échantillon pour lequel l'authenticité ou la fausseté des correspondances a été vérifiée, nous pouvons alors estimer les poids de comparaison individuels (Tepping, 1968) et les corrections.

La méthode de correction la plus simple est la méthode de la plus grande ascension (voir, par exemple, Cochran et Cox 1957). Nous commençons par utiliser les renseignements relatifs à l'authenticité ou à la fausseté des correspondances dans un échantillon afin d'estimer des probabilités comme celles définies en (2.4). Ces probabilités servent ensuite à calculer des poids de comparaison individuels, que l'on additionne afin d'obtenir une estimation du poids total (2.3). On peut déterminer les bornes UPPER et LOWER de l'expression (2.2) pour chaque intervalle de taux d'erreur de type I ou de type II qui est fixe. On en déduit aussitôt le nombre de liens possibles pour des règles comme celle définie en (2.2).

Dans un deuxième temps, nous choisissons un poids de comparaison individuel, que nous modifions d'une valeur fixe (par exemple ± 1), nous recalculons le poids total de (2.2) au moyen du nouveau poids individuel, puis nous déterminons les nouvelles bornes UPPER et LOWER ainsi qu'une nouvelle région de liens possibles.

Si, avec un intervalle fixe, la taille de la région des liens possibles diminue, nous corrigeons (à la hausse ou à la baisse) le poids de comparaison individuel jusqu'à ce que la taille cesse de décroître. Nous poursuivons en modifiant les autres poids individuels de la même façon.

Si la taille de la région des cas indéterminés décroît de façon appréciable, alors nous savons que l'hypothèse de l'indépendance conditionnelle n'est pas valide pour l'ensemble des comparaisons.

Une règle d'appariement qui repose sur des poids de comparaison individuels corrigés dépend de l'échantillon utilisé pour la méthode de la plus grande ascension.

2.3.6 Méthodes particulières

Nous avons besoin de renseignements additionnels pour décrire les méthodes particulières par lesquelles nous calculons les poids et définissons les règles d'appariement correspondantes.

Les seuls ensembles de paires considérés sont ceux qui répondent aux critères de groupage suivants.

CRITÈRES DE GROUPAGE
1. CODE POSTAL — 3 chiffres, NOM — 4 caractères
2. CODE POSTAL — 5 chiffres, RUE — 6 caractères
3. TÉLÉPHONE — 10 chiffres
4. Zone LM-NOM; ensuite, appliquer le critère 1.*

* Ce critère renferme aussi un mécanisme de suppression qui empêche l'appariement avec des mots qui reviennent souvent comme OIL, FUEL, CORP. et DISTRIBUTOR.

Nous divisons en cinq classes l'ensemble de paires formées à l'aide des quatre séries de critères de groupage;

- Classe 1 (1,021 paires) : conformes au critère n° 1 et à aucun autre ou conformes aux critères 1 et 4 et à aucun autre.
- Classe 2 (624 paires) : conformes au critère n° 2 et à aucun autre ou conformes aux critères 2 et 3 et à aucun autre.
- Classe 3 (256 paires) : conformes au critère n° 3 seulement.
- Classe 4 (244 paires) : conformes au critère n° 4 seulement.
- Classe 5 (2,240 paires) : conformes à au moins un critère mais n'appartenant à aucune des classes précédentes.

La classe 5 renferme des paires qui répondent habituellement à au moins deux critères de groupage. Les cinq classes englobent 2,991 concordances et 1,494 non-concordances et auraient dû comprendre 59 autres concordances reconnues comme telles. Winkler (1985a, 1987a) examine en détail la formation des classes et la définition des séries de critères de groupage.

Les règles d'appariement sont classées selon les méthodes de calcul des poids de comparaison individuels et la manière dont les nouvelles règles d'appariement sont définies:

Le premier type de calcul de poids (AA) est un agrégat qui porte sur toutes les paires. Le second type de calcul (A) est un agrégat qui porte sur les 4 premières classes. Le troisième (U) produit des valeurs distinctes à l'intérieur des 4 premières classes. Le quatrième (C) renferme une condition qui modifie le calcul de type U.

À mesure que sont définies successivement les règles d'appariement, le calcul des poids devient de plus en plus complexe. Les concordances qui ne sont pas incluses dans l'une ou l'autre des 5 classes ne sont pas considérées dans la section des résultats parce que leur nombre est le même pour chacune des quatre règles d'appariement.

TYPE	CALCUL DE POIDS INDIVIDUELS	RÈGLE D'APPARIEMENT
AA	Appliqué uniformément à toutes les paires contenues dans les 5 classes	S'applique à toutes les paires
A	Appliqué uniformément à toutes les paires contenues dans les classes 1 à 4	Identifie les paires de la classe 5 comme des liens; applique la méthode Fellegi-Sunter aux paires des 4 autres classes
U	Appliqué uniformément dans chacune des 4 premières classes	Identifie les paires de la classe 5 comme des liens; applique la méthode Fellegi-Sunter à chacune des paires des 4 premières classes
C	Appliqué uniformément dans chacune des 4 premières classes	Même que U, mais modifie les poids pour tenir compte de la non-indépendance.

2.4 Méthodes d'évaluation

La méthode d'évaluation fondamentale consiste à suivre l'évolution de la taille de la région des cas indéterminés lorsqu'on applique les divers genres de règles d'appariement suivant un intervalle de taux d'erreur fixe.

Comme il n'est pas possible, avec des règles aussi complexes, de construire des modèles pour des paramètres statistiques tels que le nombre de cas indéterminés, nous nous servons de la méthode "bootstrap" (Efron 1987, 1982, 1979) pour évaluer la distribution de ces paramètres.

S'il y a des ensembles de paires pour lesquelles l'authenticité ou la fausseté des concordances est vérifiée, nous pouvons utiliser la méthode "bootstrap" d'Efron pour estimer la variation des paramètres selon les règles suivantes:

1. Tirer (avec remise) des échantillons de référence de même taille.
2. Estimer les poids de comparaison individuels définis en (2.4) en se servant des renseignements relatifs à l'authenticité ou à la fausseté des concordances dans l'échantillon et utiliser ces poids pour estimer le poids total par la formule (2.3).
3. Calculer les bornes LOWER et UPPER à l'aide de chaque échantillon (en l'occurrence, nous limitons la proportion de liens classés comme non-concordances à 2% et la proportion de non-liens classés comme concordances à 3%).
4. À l'aide des poids de comparaison individuels estimés à l'étape 2, calculer un poids de comparaison total pour chaque paire contenue dans l'ensemble prélevé. Utiliser les bornes calculées à l'étape 3 pour identifier les paires comme liens, cas indéterminés ou non-liens.
5. À l'aide des estimations d'échantillons, déterminer la moyenne et la variance des poids limites, des taux d'erreur de classification et du nombre de cas indéterminés.

Les bornes (2 et 3%, étape 3) visent à faire en sorte que les taux d'erreur de classification pour toute la base de données soient inférieurs à 5%.

Les calculs et les corrections doivent être effectués uniformément pour tous les échantillons de référence. Les poids individuels corrigés, les poids totaux et les bornes doivent tous être obtenus à l'aide des mêmes méthodes de correction. Si un poids individuel est corrigé à la hausse (étape 2) d'une quantité x ou d'un pourcentage y pour un échantillon, la même correction doit s'appliquer pour les autres échantillons.

Comme les distributions sous-jacentes peuvent n'être pas normales ou être biaisées et asymétriques, nous pouvons utiliser de nouvelles méthodes proposées par Efron (1982, 1987) pour déterminer les intervalles de confiance.

3. RÉSULTATS

Cette section se divise en trois parties. La première est une comparaison générale des quatre méthodes de pondération décrites dans la sous-section 2.3.6. La seconde approfondit les deux meilleures méthodes parmi ces quatre. Enfin, la troisième contient les résultats de l'évaluation par la méthode "bootstrap".

3.1 Comparaison générale

Nous fixons une borne supérieure de 5% pour la proportion de concordances classées par erreur dans les non-concordances et une borne supérieure de 2% pour la proportion de non-concordances classées par erreur dans les concordances. Comme nous utilisons des

données discrètes, les taux d'erreur réels n'égaleront généralement pas les bornes supérieures (tableau 1, colonnes 2 et 3).

Tableau 1

Taux d'erreur et nombre de cas indéterminés pour diverses méthodes de pondération

Type de poids	Proportion des concordances classées par		Total des paires classés		Cas indéterminés
	non-concordances	concordances	non-concordances	concordances	
AA	.047	.020	964	2009	1512
A	.041	.015	952	2481	1052
U	.050	.020	1083	2707	695
C	.033	.019	1441	2947	97

Nous constatons qu'à mesure que s'accroît la complexité de la méthode de pondération, le nombre de cas indéterminés (taille de la région soumise à une révision manuelle) diminue de façon appréciable, passant de 1,512 à 97. Cela indique que la complexité accrue des méthodes de calcul des poids engendre de meilleures règles de décision.

Nous constatons aussi que les deux dernières méthodes, qui consistent à calculer séparément des poids de comparaison individuels dans les quatre premières classes, produisent les plus petits ensembles de cas indéterminés (695 et 97 respectivement).

3.2 Les meilleures méthodes

Considérons plus en détail les deux meilleures méthodes, soit les règles d'appariement fondées sur les poids de type U et de type C. Les résultats obtenus par l'application des poids de type U et de type C sont présentés dans les tableaux 2 et 3 respectivement. En déterminant les poids limites pour chaque classe, nous fixons une limite supérieure approximative de 5% pour la proportion de non-concordances classées par erreur dans les concordances et une limite supérieure approximative de 2% pour la proportion de concordances classées par erreur dans les non-concordances. La borne supérieure générale est conservée.

Tableau 2

Résultats de l'application d'une règle d'appariement fondée sur des poids de type U pour définir les concordances et les non-concordances (Taux d'erreur de classification global de 5%)

Classe	Poids limites		Concordances classées par		Total des paires classées		Total des paires non-classées	Total des enregistrements
	Lower	Upper	non-concordances	concordances	non-concordances	concordances		
1	0.5	6.5	39	14	674	264	83	1021
2	-4.5	3.5	2	4	100	115	409	624
3	-4.5	6.5	2	1	55	42	159	256
4	2.5	11.5	11	2	254	46	44	344
Totaux			54	21	1083	467	695	2245

Tableau 3

Résultats de l'application d'une règle d'appariement fondée sur des poids de type C pour définir les concordances et les non-concordances (Taux d'erreur de classification global de 3%)

Classe	Poids limites		Concordances classées par		Total des paires classées		Total des paires non-classées	Total des enregistrements
	Lower	Upper	non-concordances	concordances	non-concordances	concordances		
1	4.5	7.5	28	8	692	274	55	1021
2	2.5	2.5	5	3	379	245	0	624
3	-0.5	4.5	5	6	104	110	42	256
4	8.5	8.5	9	4	266	78	0	344
Totaux			47	21	1441	707	97	2245

Si nous comparons les colonnes 4 et 5 des tableaux 2 et 3, nous constatons que les totaux correspondants sont comparables. Ce résultat est conforme à la méthode de délimitation. Dans chaque classe, la règle d'appariement fondée sur des poids de type C produit un moins grand nombre de cas indéterminés que la règle d'appariement fondée sur des poids de type U.

De fait, le nombre d'enregistrements considérés comme cas indéterminés est moindre pour les classes 1 et 4 (83 contre 55 et 44 contre 0 respectivement) et considérablement moindre pour les classes 2 et 3 (409 contre 0 et 159 contre 42 respectivement).

La méthode fondée sur des poids de type C permet de classer la totalité des paires contenues dans les classes 2 et 4.

La règle d'appariement fondée sur des poids de type C se distingue de celle fondée sur des poids de type U à deux points de vue. D'une part, nous modifions les poids de concordance qui se rattachent aux quatre sous-zones du NOM après avoir classé les mots par ordre décroissant de longueur (tableau 4). Les seules variations appréciables (supérieures à 2.5 sur l'échelle $\log_2$) sont observées dans la classe 2.

Tableau 4

Correction des poids de concordance pour les sous-zones tirées de la zone LM/NOM[1]

	SOUS-ZONES			
CLASSE	1	2	3	4
1	.	.	-	+
2	++	++	+	+
3	+	+	-	++
4	.	+	-	+

[1] "." signifie un écart inférieur à 1.0,
"+" et "-" signifient un écart supérieur à 1.0 et inférieur à 2.5
"++" signifie un écart supérieur à 2.5.

D'autre part, on n'utilise le poids de concordance que si quatre sous-zones correspondantes, soit les trois sous-zones de VILLE et la sous-zone ÉTAT, concordent. De fait, la variation de poids accroît généralement le pouvoir de différenciation relatif des concordances/non-concordances dans toutes les sous-zones sauf celles de la zone VILLE.

La plus forte réduction du nombre de cas indéterminés (de 409 à 0) est observée dans la classe 2. Les non-liens qui ont une zone VILLE conforme sont proportionnellement légèrement plus nombreux que les liens (.95=359/379 contre .95=223/245).

Nous donnons ci-dessous l'exemple d'une concordance qui n'est pas reconnue comme un lien selon la règle fondée sur des poids de type U mais qui l'est selon la règle fondée sur des poids de type C.

NOM	RUE	VILLE	ÉTAT	CODE POSTAL
Roberts Heat Oils	167 Sycamore St.	Dayton	OH	53315
Maxwell S Robert Heat Oil	167 Sycamore St.	Dayton	OH	53315

Les six premiers chiffres du numéro de téléphone concordaient également.

Voici un exemple d'une fausse concordance déduite à l'aide de poids de type C.

NOM	RUE	VILLE	ÉTAT	CODE POSTAL
Molar Petro	167 Sycamore St.	Dayton	OH	53315
Petrochem	167 Sycamore St.	Dayton	OH	53315

Ces deux entreprises sont situées à la même adresse et ont le même numéro de téléphone.

Voici un exemple d'une fausse non-concordance déduite à l'aide de poids de type C.

NOM	RUE	VILLE	ÉTAT	CODE POSTAL
Johns Geo M	167 Sycamore St.	Springfield	OH	53315
Geo M Johns Jobber	167 Sycamore	Spring Field	OH	53315

À cause de l'insertion ou de la suppression de blancs dans des zones correspondantes, l'ordinateur a tendance à définir les paires d'enregistrements en question comme des non-concordances.

3.3 Application de la méthode "bootstrap"

Les résultats présentés dans cette sous-section découlent de l'application de méthodes de calcul d'intervalles de confiance "boostrap" de complexité croissante (tableau 5). Pour chaque classe, on utilise 500 échantillons répétés pour calculer des intervalles de confiance de 90% pour les estimations du nombre d'enregistrements définis comme des cas indéterminés. Le taux d'erreur de classification maximum est fixé à 5%.

Tableau 5

Intervalles de confiance bootstrap de 90% pour le nombre de cas indéterminés 500 échantillons répétés

Type de Poids	Classe	Intervalle ordinaire	Intervalle BC	Intervalle BC_a
C	1	(42,117)	(37,108)	(37,108)
C	2	(0, 0)	(7, 7)	(7, 7)
C	3	(31,154)	(34,156)	(34,156)
C	4	(0, 36)	(0, 39)	(0, 39)
U	1	(122,192)	(128,296)	(128,296)
U	2	(383,501)	(383,501)	(383,501)
U	3	(149,201)	(142,197)	(142,197)
U	4	(35, 82)	(33, 81)	(33, 81)

Le premier genre d'intervalle est l'intervalle "bootstrap" ordinaire qui repose en partie sur la théorie courante (Efron, 1979). Le second genre d'intervalle, désigné par BC, est un intervalle qui comporte une correction pour le biais (Efron 1979, 1982). Le troisième genre d'intervalle, désigné par BC_a, est déterminé au moyen de corrections pour le biais et l'asymétrie (Efron 1987).

En examinant le tableau 5, nous constatons que les trois intervalles sont à peu près de la même longueur pour une classe donnée. Si la méthode de correction utilisée pour obtenir des poids de type C dépendait largement des échantillons de référence, nous nous attendrions que les intervalles de confiance rattachés aux poids de type C soient plus grands que ceux rattachés aux poids de type U.

Le fait que l'on observe de grands intervalles d'un côté comme de l'autre indique que les résultats dépendent largement des échantillons de référence. En outre, le fait que les intervalles de confiance ordinaires soient comparables aux intervalles BC et BC_a correspondants indique que les distributions respectives ne sont ni biaisées ni asymétriques.

Le nombre de cas indéterminés contenus dans les intervalles fondés sur les poids de type C est presque toujours inférieur au nombre de cas indéterminés contenus dans les intervalles fondés sur les poids de type U. Seuls les intervalles calculés pour les classes 3 et 4 montrent un léger chevauchement. Il est donc raisonnable d'accepter l'hypothèse selon laquelle la règle d'appariement fondée sur des poids de type C surpasse constamment la règle fondée sur des poids de type U.

4. ANALYSE

Cette section se divise en trois parties. Dans la première, nous considérons l'utilité de faire des comparaisons qui dépendent en partie d'autres comparaisons. Ensuite, nous proposons une définition élargie simple des variables binaires que nous avons utilisées plus tôt dans cet article. Enfin, dans la troisième partie, nous décrivons des méthodes qui permettent de définir de nouvelles séries de critères de groupage.

4.1 Utilité des comparaisons qui dépendent d'autres comparaisons

Intuitivement, il est justifié de faire un certain nombre de comparaisons, dont certaines peuvent dépendre en partie d'autres comparaisons, parce qu'elles sont susceptibles de créer un pouvoir de différenciation additionnel lorsqu'elles servent à des règles établies convenablement. Newcombe et Kennedy (1962, voir aussi Newcommbe et coll. 1983) donnent des exemples de comparaison de portions de zones réservées au nom, où, intuitivement, ces comparaisons peuvent dépendre d'autres comparaisons. Néanmoins, les comparaisons additionnelles peuvent produire des règles d'appariement meilleures que les règles qui n'utilisent pas ces mêmes comparaisons.

Ce qui compte surtout lorsqu'on utilise des comparaisons additionnelles, c'est de savoir exploiter convenablement le pouvoir de différenciation additionnel qui en résulte. La série de comparaisons que nous avons faites dans cet article -- notamment la comparaison des sous-zones de la zone réservée au nom -- n'est pas indépendante au sens de l'équation (2.2). Le but premier de cette série de comparaisons est d'illustrer des méthodes qui permettent d'obtenir systématiquement de meilleures règles d'appariement lorsque l'hypothèse de l'indépendance conditionnelle n'est pas valide.

4.2 États de comparaison: nombre et catégories

Pour mieux exposer les principes fondamentaux des règles d'appariement, nous nous en sommes tenus aux variables binaires (par exemple concordance/non-concordance). Fellegi et Sunter (1969, p. 1194-1195) proposent une méthode qui élargit la notion de variable binaire en subdivisant l'état "concordance" en états "concordance/valeur de concordance".

Dans les exemples typiques (Newcombe et Kennedy 1962, Newcombe et coll. 1983), on attribue des poids de concordance/valeur de concordance qui correspondent à l'inverse de la fréquence relative d'apparition d'un enchaînement de caractères donné, souvent le nom de famille.

La méthode de Fellegi-Sunter évoquée ci-dessus s'applique aussi aux règles d'appariement définies dans cet article. En pratique, les comparaisons que l'on juge les plus appropriées comme poids de concordance/valeur de concordance peuvent être définies comme la série principale. Ces poids sont estimés surtout à l'aide de la méthode fondamentale de Fellegi-Sunter. Les autres poids sont estimés par les méthodes de correction décrites dans cet article.

De plus, il est possible d'utiliser une formule de comparaison d'enchaînements (Winkler, 1985a) pour attribuer une cote aux comparaisons qui renferment de légères variations d'orthographe, celles-ci étant causées par l'insertion, la suppression ou la transposition de caractères. La paire (Zabrinsky, Zabrinky) en est un exemple type. Si le terme Zabrinsky était accompagné d'un poids fondé sur la fréquence d'apparition, nous pourrions rajuster ce poids à la baisse pour tenir compte de la faute d'orthographe.

4.3 Autres critères de groupage

Lorsqu'on utilise une série de critères de groupage pour réduire le nombre de paires de **AXB** qui doivent faire l'objet d'un traitement plus poussé, on vise deux objectifs contradictoires. D'une part, on cherche à réduire fortement le nombre de paires qui doivent être traitées et à obtenir un ensemble (de paires) où les règles d'appariement peuvent départager clairement les concordances et les non-concordances. D'autre part, on cherche à obtenir un ensemble qui renferme autant d'éléments de M (concordances) que possible.

Pour savoir s'il est souhaitable de définir des critères de groupage additionnels, il faut d'abord obtenir des estimations du nombre de concordances qui échappent à une série

donnée de critères de groupage. Si les estimations sont raisonnablement faibles, il ne sera pas nécessaire de définir de nouveaux critères.

Pour estimer le nombre de concordances qui échappent à une série donnée de critères, Scheuren (1983) propose d'utiliser les méthodes régulières de saisie-resaisie telles qu'elles sont décrites dans Bishop, Fienberg et Holland (1975, chapitre 6).

Winkler (1987a) a appliqué ces méthodes aux données empiriques et aux quatre séries de critères de groupage présentées dans cet article. À l'aide du modèle loglinéaire du meilleur ajustement pour les comptages d'enregistrements saisis et d'enregistrements non-saisis selon les critères de groupage, il a calculé un intervalle de confiance de 95% (27,160).

Pour n'importe quel ensemble de paires donné (particulièrement celles obtenues par groupage), nous pouvons utiliser la méthode II de Fellegi et Sunter pour estimer le nombre de concordances dans cet ensemble ainsi que les proportions de concordance $P(\tau \varepsilon B_j | M)$ et $P(\tau \varepsilon B_j | U)$, $i = 1, 2,$ et 3, où B_j, $i = 1, 2,$ et 3 désigne un ensemble de trois événements de concordance quelconques. La validité du calcul dépend de l'hypothèse de l'indépendance conditionnelle, en supposant que le nombre de concordances dans l'ensemble de paires soit supérieure à 0 et en supposant que $P(\tau \varepsilon B_j | M) \neq P(\tau \varepsilon B_j | U)$, $i = 1, 2,$ and 3.

Même si l'hypothèse de l'indépendance conditionnelle n'était pas valide, il serait possible d'obtenir une estimation grossière de la proportion de concordances dans le nouvel ensemble de paires. Winkler (1978b) utilise à cette fin l'algorithme EMM (espérance mathématique-maximisation) avec une hypothèse moins poussée.

Si la proportion de concordances est trop faible ou si les règles d'appariement fondées sur l'information contenue dans le nouvel ensemble de paires ne permettent de définir avec suffisamment de précision les concordances, on pourrait devoir définir des critères de groupage additionnels ou accepter qu'un nombre déterminé de concordances échappent à la série de critères existants.

5. RÉSUMÉ

Les résultats exposés dans cet article indiquent que si nous voulons réduire la taille de la région des cas indéterminés lorsque l'hypothèse de l'indépendance conditionnelle n'est pas valide, nous devons corriger les estimations des poids de comparaison individuels de manière que leur somme produise de meilleures règles d'appariement.

BIBLIOGRAPHIE

Arrelano M. (1985). "An Implementation of the Two-Population Fellegi-Sunter Probability Linkage Model, "dans *Record Linkage Techniques 1985*, colligé par W. Alvey et B. Kilss, U.S. Internal Revenue Service, publication 1299 (2-86), p. 255-258.

Bishop, Y. M. M., Fienbert, S. E., et Holland, P. W. (1975), *Discrete Multivariate Analysis*, MIT Press, Cambridge, MA.

Cochran, W.G. et Cox, G.M. (1957), *Experimental Designs*, J. Wiley and Sons, New York.

Coulter, R.W. (1977), "An Application of a Theory for Record Linkage," Rapport technique produit par le Statistical Reporting Service du Département de l'agriculture des É.-U.

Coulter, R.W. et Mergerson, J.W. (1977), "An Application of a Record Linkage Theory in Constructing a List Sampling Frame", Rapport technique produit par le Statistical Reporting Service du Département de l'agriculture des É.-U.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Ann. Stat.*, 7, p. 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Methods*, SIAM, Philadelphie, PA.

Efron, B. (1987), "Better Bootstrap Confidence Intervals (with discussion)", *JASA, 82*, p. 171-185.

Fellegi, I. P., et Sunter, A. B. (1969), "A Theory for Record Linkage", *JASA, 40*, p. 1183-1210.

Kelley, R. P. (1985a), "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," dans *Record Linkage Techniques 1985*, colligé par W. Alvey et B. Kilss, U.S. Internal Revenue Service, publication 1299 (2-86), p. 199-203.

Kelley, R. P. (1985b), "Bayesian Adjustment of the Matching Discriminant Function", communication présentée au congrès annuel de l'ASA à Las Vegas, NV.

Kelley, R. P. (1986), "Robustness of the Census Bureau's Record Linkage System," ASA 1986 Proceedings of the Section on Survey Research Methods, p. 620-624.

Newcombe, H.B., Kennedy, J.M. Axford, S.J., et James, A. P. (1959), "Automatic Linkage of Vital Records", *Science, 130*, p. 954-959.

Newcombe, H.B. et Kennedy, J.M. (1962), "Record Linkage", *Communications of the ACM, 5*, p. 563-566.

Newcombe, H.B., Smith, ME.E, Howe, G.R., Mingay, J., Strugnell, A., et Abbatt, J.D. (1983)., "Reliability of Computerized Versus Manual Searches in a Study of the Health of Eldorado Uranium Workers", *Comptu. Biol. Med., 13*, p. 157-169.

Rogot, E., Schwartz, S., O'Connor, K., et Olsen, C. (1983), "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," *ASA 1983 Proceedings of the Section on Survey Research Methods*, p. 319-324.

Scheuren, F. (1983), "Design and Estimation for Large Federal Surveys using Administrative Records," *ASA 1983 Proceedings of the Section on Survey Research Methods*, p. 377-381.

Scheuren, F. (1985), "Methodological Issues in Linkage of Multiple Date Bases", dans *Record Linkage Techniques 1985*, colligé par W. Alvey et B. Kilss, U.S. Internal Revenue Service, publication 1299 (2-86), p. 155-167.

Smith, M. et Silins, J. (1981), "Generalized Iterative Record Linkage System," *ASA 1981 Proceedings of the Social Statistics Section*, p. 128-137.

Smith, M., Newcombe, H.B., et Dewar, R. (1983) "Automated Nationwide DeathClearance of Provincial Cancer Registry Files - The Alberta Cancer Registry Study," ASA 1983 Proceedings of the Section on Survey Research Methods, p. 300-305.

Statistique Canada, Division des systèmes informatiques (1982), "Record Linkage Software".

Statistique Canada, Division de la planification et du soutien en informatique (1984), "Record Linkage Software".

Tepping, B. J. (1968), "A Model for Optimum Linkage of Records," *JASA 63*, p. 1321-1332.

Département de l'agriculture des É.-U./Statistical Reporting Service (1979), "List Frame Development: Procedures and Software."

Département du commerce des É.-U., Bureau of the Census/Survey Research Division (1978a), "UNIMATCH: A Record Linkage System."

Département du commerces des É.-U., Bureau of the Census/Survey Research Division (1978b), "ZIPSTAN: Generalized Address Standardizer."

Winkler, W. E. (1985a), "Preprocessing of Lists and String Comparison," dans *Record Linkage Techniques 1985*, colligé par W. Alvey et B. Kilss, U.S. Internal Revenue Service, publication 1299 (2-86), p. 181-187.

Winkler, W.E. (1985b), "Exact Matching Lists of Businesses: Blocking, Subfield Identification, and Information Theory," dans *Record Linkage Techniques 1985*, colligé par W. Alvey et B. Kilss, U.S. Internal Revenue Service, publication 1299 (2-86), p. 227-241.

Winkler, W. E. (1985c), "Exact Matching Lists of Businesses," *ASA 1985 Proceedings of the Section on Survey Research Methods*, p. 438-443.

Winkler, W. E. (1987a), "An Application of the Fellegi-Sunter Model of Record Linkage to Lists of Businesses". Energy Information Administration, rapport technique.

Winkler, W. E. (1987b), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage", rapport technique.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

CONCEPTS ET TECHNIQUES POUR L'AMÉLIORATION DE L'APPARIEMENT PROBABILISTE

HOWARD B. NEWCOMBE, MARTHA E. FAIR et PIERRE LALONDE[1]

RÉSUMÉ

On peut améliorer les méthodes automatisées d'appariement probabiliste des enregistrements concernant des personnes ou d'autres entités en simulant la souplesse et les procédés d'un cerveau humain perspicace accomplissant la même tâche. On peut comparer des identificateurs qui ne sont pas de la même sorte; il peut arriver qu'une correspondance ne favorise pas l'appariement; et il faut parfois analyser les enregistrements à plusieurs niveaux pour exploiter pleinement leur pouvoir de différenciation. On peut créer de grands fichiers de paires d'enregistrements **non appariables** formées au hasard pour les comparer directement avec des fichiers de paires d'enregistrements **appariés,** ces derniers fichiers étant analogues aux impressions qui subsistent dans la mémoire d'une personne après qu'elle a effectué les appariements manuellement. Les deux fichiers permettent de calculer directement des probabilités pour toutes les comparaisons auxquelles on pourrait penser sous forme de rapport de fréquence de résultat entre des paires d'enregistrements **appariés** et des paires d'enregistrements **non appariables.** L'empirisme dispense de faire certains des calculs théoriques, qui deviennent laborieux et peuvent facilement engendrer des erreurs quand la définition des résultats est complexe. La simplicité favorise davantage l'intuition que le recours au raisonnement pour faire des comparaisons efficaces et déceler les problèmes que posent les méthodes actuellement utilisées. Nous donnons des exemples.

1. INTRODUCTION

Il y a deux façons de considérer l'appariement d'enregistrements par ordinateur:

a) soit comme une tentative pour imiter **intuitivement** le fonctionnement du cerveau humain quand il accomplit la même tâche,

b) soit comme une activité fondée sur un modèle **mathématique** avec des règles et des choix définis de façon précise.

---

[1] Howard B. Newcombe est un consultant; il a déjà été chef de la Population Research Branch d'Énergie atomique du Canada Ltée à Chalk River. Son adresse actuelle est C.P. 135, Deep River (Ontario), K0J 1P0. Martha E. Fair est chef de la Sous-section de l'hygiène du travail et du milieu à Statistique Canada et Pierre Lalonde, gestionnaire de projet dans cette même sous-section (Section de la statistique de l'état civil et de la santé, Division de la santé, Immeuble R.H. Coats, 18e étage, Parc Tunney, Ottawa (Ontario), K1A 0T6).

Ce sont là deux points de vue complémentaires de la même chose, de sorte que l'arithmétique peut être la même. Par ailleurs, c'est souvent l'aspect sur lequel on veut insister qui déterminera l'option qu'on choisira.

Si l'on veut imiter le cerveau humain, il faut consacrer à l'observation de son fonctionnement un temps qui peut paraître très long. Les appariements sont d'abord faits manuellement pour donner un aperçu. Les opérations automatisées sont conçues pour tenir compte de cette expérience. Elles doivent être testées, mais pas seulement pour détecter les erreurs; il est tout aussi important de voir quels indices ont échappé à la machine. Si un humain peut les découvrir, il arrivera souvent qu'on pourra donner d'autres instructions à la machine pour qu'elle en fasse autant. Les conséquences pratiques de cette préoccupation - et c'est en effet une préoccupation - sont importantes.

Nous allons considérer trois aspects importants de toute opération d'appariement complexe et praticable. Ces aspects sont les suivants:

1. L'utilisation de **comparaisons à plusieurs niveaux** et de **comparaisons entre des identificateurs qui ne sont pas de la même sorte** pour mieux exploiter le pouvoir de différenciation offert. Ces deux techniques sont utilisées couramment depuis longtemps à Statistique Canada.
2. L'utilisation de **fichiers réels de paires d'enregistrements NON APPARIABLES formées au hasard,** qui par ailleurs seraient comparables à des paires d'enregistrements APPARIÉS, pour calculer directement des **probabilités** (c'est-à-dire des fréquences de résultat dans des paires d'enregistrements APPARIÉS par opposition à des paires d'enregistrements NON APPARIABLES, donc **des rapports de fréquence,** dont le logarithme est appelé **poids).** Cette méthode est relativement nouvelle à Statistique Canada.
3. Une procédure appropriée pour **ajuster** les probabilités de façon qu'elles reflètent le pouvoir de différenciation plus ou moins élevé de certains caractéristiques, par exemple des noms rares par rapport aux noms plus répandus, quand elles **concordent entièrement ou partiellement.** C'est une nouvelle technique, qui n'est pas encore utilisée, mais qui est nécessaire du fait que les méthodes actuelles introduisent de grandes erreurs lorsqu'on les applique à des **concordances partielles.**

Ces développements sont le résultat d'un effort soutenu pour imiter le cerveau humain. La logique qu'ils emploient devrait donc être intuitivement évidente, l'arithmétique devrait être simple et la précision résultante devrait refléter celle d'un chercheur humain perspicace.

2. LES CONCEPTS

Tous les appariements, qu'ils soient faits par des humains ou par des machines, consistent à comparer des identificateurs (noms, dates de naissance, lieux de naissance, etc.) pour voir si les résultats (concordances, ressemblances, dissemblances, etc.) sont plus typiques de paires d'enregistrements **APPARIÉS** ou **NON APPARIABLES** formées au hasard (voir tableau 1). Les **rapports de fréquence** résultants sont plutôt comme des **cotes** dans une course de chevaux. Ou, pour être plus précis, ce sont les facteurs par lesquels chaque élément de renseignement modifie les **probabilités** globales.

C'est la logique par laquelle ce concept intuitivement évident est utilisé qui nous intéresse ici. Il y a d'excessives et trop fréquentes simplifications qui masquent la souplesse et la force réelles de ce concept et qu'il faudrait corriger. Par exemple:

- même des identificateurs **de différentes sortes** peuvent être comparés;

- souvent des résultats simples comme la **concordance** ou la **non-concordance** n'exploitent pas très bien le pouvoir de différenciation;
- la **concordance** ne favorise pas toujours l'appariement tandis que la **non-concordance** ne l'empêche pas toujours.
- l'arithmétique n'a pas besoin d'être compliquée pour assurer la précision des probabilités calculées.

Tableau 1

Concepts d'appariement: Probabilité = Rapport de fréquence

Pour tout résultat donné, défini de quelque façon que ce soit

$$\textbf{Probabilité} = \frac{\text{Fréquence dans des paires d'enregistrements } \textbf{appariés}}{\text{Fréquence dans des paires d'enregistrements } \textbf{non appariables}}$$

Application: Utilisé par des humains ou par des machines
Dénominateur tire de la théorie ou observé

En fait, la précision dépend d'un type de raisonnement étonnamment souple, comme on peut le voir dans les exemples suivants:

2.1 Comparaison d'identificateurs différents (tableau 2)

En appariant les extraits de naissance et les extraits d'acte de mariage des parents, on peut comparer le RANG DE NAISSANCE d'un enfant avec la DURÉE DU MARIAGE au moment de la naissance. Les petits numéros de rangs de naissance sont fréquents dans les premières années du mariage, et la naissance d'un premier enfant tard dans le mariage est un événement assez rare pour être considéré comme notable. Inversement, la naissance d'un cinquième enfant pendant la première année du mariage serait certainement inhabituel, à moins bien entendu qu'il ne s'agisse d'un second mariage. Cette "connexité" entre des identificateurs de différentes sortes est une source de pouvoir de différenciation et devrait être utilisée.

Tableau 2

Concepts d'appariement: Comparaison d'identificateurs **de différentes sortes**
Extrait de naissance **par rapport** à extrait d'acte de mariage des parents

Rang de naissance	Durée du mariage (années)					
	1	2	3	4	5	6+
1	+	—	—	—	—	X
2	—	+	—	—	—	—
3	—	—	+	—	—	—
4	—	—	—	—	—	—
5	—	—	—	—	—	—
6 et plus	X	—	—	—	—	—

\+ = Courant
X = Moins courant
Concordance / non-concordance a peu de signification

2.2 La matrice des "niveaux" et des "valeurs" de résultat

Dans l'exemple qui précède, on peut construire une matrice de combinaisons possibles de RANGS DE NAISSANCE et de DURÉE DU MARIAGE. Le cerveau humain ne mémorise pas toute la matrice, mais il voit clairement quelles combinaisons sont courantes et normales et quelles combinaisons sont peu probables ou même extraordinaires. L'idée de **concordance** ou de **non-concordance** ne présente pas d'intérêt particulier dans ce cas-ci; il faut plutôt mesurer le degré de **connexité** ou de **non-connexité** empirique, et c'est ce à quoi sert le **rapport de fréquence** (ou rapport entre la fréquence d'un résultat donné dans des paires d'enregistrements APPARIÉS et la fréquence dans des paires d'enregistrements NON APPARIABLES).

2.3 Interprétation des résultats des comparaisons (tableau 3)

Il est préférable de **ne pas** supposer à l'avance quels résultats **favoriseront** l'appariement et quels résultats **ne le favoriseront pas.** Par exemple, quand on apparie des extraits de naissance et des enregistrements correspondant aux enfants d'une même fratrie, il est utile de comparer les rangs de naissance indiqués dans les deux enregistrements. Dans ce cas, les concordances **ne favorisent pas** l'appariement, parce qu'il ne devrait pas y avoir deux premiers-nés dans la même fratrie, ni deux deuxièmes enfants, et ainsi de suite. En fait, il est inutile de porter des jugements anticipés parce que l'interprétation correcte devient évidente dès qu'on a assez de données pour calculer un rapport de fréquence. Dans un certain sens, la machine trouve par elle-même à mesure que son "expérience" s'accroît.

Tableau 3

Concepts l'appariement:il arrive que la concordance ne **justifie pas l'appariement**
Extrait de naissance **par rapport** à extrait de naissance d'un enfant de mêmes parents

Rang de Order	Rang de naissance d'un enfant de mêmes parents					
	1	2	3	4	5	6+
1	X	—	—	—	—	—
2	—	X	—	—	—	—
3	—	—	X	—	—	—
4	—	—	—	X	—	—
5	—	—	—	—	X	—
6+	—	—	—	—	—	—

X = peu probable (La concordance **ne favorise pas** l'appariement)

2.4 Les rapports de fréquence peuvent être obtenus empiriquement

Les rapports de fréquence ou probabilités, pour tout résultat donné, peuvent être calculés directement à partir de vrais fichiers de paires d'enregistrements APPARIÉS et de paires d'enregistrements NON APPARIABLES. Le calcul demeure simple quelle que soit la complexité de la comparaison, pourvu qu'on puisse obtenir ces deux fichiers. De toute évidence, le cerveau humain fait quelque chose de très similaire quand il garde en mémoire une image de ce à quoi ressemblaient les paires d'enregistrements qu'il **a appariés** et de la façon dont elles différaient des paires d'enregistrements qu'il a rejetés

comme NON APPARIABLES. Ces deux sortes de résidus mnémoniques forment sans doute la base empirique sur laquelle reposent la plupart du temps les jugements quantitatifs humains. La même démarche peut être utilisée par les ordinateurs; ces derniers ont l'avantage d'avoir une meilleure mémoire que les humains, tout en pouvant être plus systématiques. Comme cet empirisme simplifie la logique et les calculs, il est maintenant utilisé de façon courante à Statistique Canada.

2.5 Combien de niveaux de résultat? (tableau 4)

Pour la plupart des comparaisons, il y a plusieurs niveaux de résultat possibles, dont certains doivent être regroupés pour des raisons de simplicité. Toutefois, il faut éviter de combiner des niveaux où les **rapports de fréquence** (ou **probabilités**) diffèrent beaucoup parce que cela fait perdre une partie du pouvoir de différenciation. Par exemple, la comparaison du MOIS DE NAISSANCE produit 144 combinaisons possibles ou "valeurs" de résultat, qui peuvent être regroupées en un spectre de 12 "niveaux" (c'est-à-dire avec des écarts de 0, 1, 2, 3, ..., 11 mois). Habituellement, les rapports de fréquence pour des écarts de 1, 2, 3 et même 4 mois formeront une série graduée intermédiaire entre la **concordance totale** à une extrémité du spectre et l'**extrême non-concordance** à l'autre extrémité. En règle générale, on ne doit pas combiner de niveaux où les rapports de fréquence diffèrent de 200 pour cent ou plus. Toutefois, pour des écarts aussi importants que 5 mois ou plus, les rapports de fréquence (ou probabilités) seront habituellement si semblables que toute perte de pouvoir de différenciation attribuable au regroupement sera négligeable. Le principe s'applique à tout spectre de niveaux de résultat obtenus à partir de la comparaison de n'importe quel identifiant. Seuls les tests empiriques montreront quels niveaux on pourra combiner sans danger et quels niveaux il vaut mieux garder séparés.

Il n'est pas nécessaire de démontrer que le cerveau humain fonctionne comme on pense qu'il fonctionne. Ce qui importe, c'est de montrer que ce que nous avons appris en faisant les appariements nous-mêmes augmente effectivement la précision des probabilités calculées. On peut habituellement mesurer l'amélioration en comparant les rapports de fréquence obtenus à partir d'une méthode plus grossière. L'interprétation devrait être évidente.

Tableau 4

Concepts d'appariements: Quelques résultats peuvent être **regroupés**
Mois de naissance: Nombre de mois d'écart

		Mois figurant sur l'enregistrement B											
		J	F	M	A	M	J	J	A	S	O	N	D
Mois figurant sur l'enregistrement A	J	0	1	2	3	4	5	6	7	8	9	10	11
	F	1	0	1	2	3	4	5	6	7	8	9	10
	M	2	1	0	1	2	3	4	5	6	7	8	9
	A	3	2	1	0	1	2	3	4	5	6	7	8
	M	4	3	2	1	0	1	2	3	4	5	6	7
	J	5	4	3	2	1	0	1	2	3	4	5	6
	J	6	5	4	3	2	1	0	1	2	3	4	5
	A	7	6	5	4	3	2	1	0	1	2	3	4
	S	8	7	6	5	4	3	2	1	0	1	2	3
	O	9	8	7	6	5	4	3	2	1	0	1	2
	N	10	9	8	7	6	5	4	3	2	1	0	1
	D	11	10	9	8	7	6	5	4	3	2	1	0

Conclusion: ne pas regrouper les niveaux à moins que les rapports de fréquence ne soient similaires.

3. LA PERTINENCE DES RÉSULTATS MULTIPLES

Pour presque toutes les comparaisons d'identificateurs, la reconnaissance de la multiplicité des résultats est la seule façon d'utiliser tout le pouvoir de différenciation offert. Le cas des identificateurs figurant sur un avis de décès servira à illustrer ce point (voir tableau 5.).

Différents degrés de **ressemblance** ou de **connexité** sont observés dans le cas des NOMS DE FAMILLE, des PRÉNOMS, des DATES DE NAISSANCE et des INITIALES du fait que ces renseignements sont parfois inversés. Le LIEU DE RÉSIDENCE, le LIEU DE TRAVAIL et l'ÉTAT MATRIMONIAL sont susceptibles de changer de façon non aléatoire au cours de la vie d'une personne, de sorte que l'intervalle de temps entre les deux enregistrements peut jouer un rôle important dans la répartition. Dans le cas du LIEU DE NAISSANCE, qui lui ne change pas, l'endroit déclaré n'est pas toujours très précis et certains écarts observés sont en réalité des concordances partielles, comme dans le cas où des provinces ou des pays voisins sont pris l'un pour l'autre. Même le SEXE, qu'on pourrait considérer comme la seule exception, peut produire quatre combinaisons de valeurs (MM, FF, MF, FM).

Tableau 5

Pertinence des résultats multiples: **Identificateurs sur les actes de décès**

Noms:	Personnes décédée; Père, Mère; Conjoint(e); Répondant
Lieux:	Naissance de la personne décédée, du père, de la mère; Résidence; Décès
Dates:	Naissance; Décès
Renseignements Personnels:	Sexe; État matrimonial; Origine ethnique; Profession; Cause du décès

Conclusion: Toutes les comparaisons peuvent avoir des résultats multiples

Bref, des résultats multiples font ressortir tout le pouvoir de différenciation qui pourra être utilisé.

4. EXEMPLES DE RÉSULTATS MULTIPLES

Il n'y a pas vraiment pas de limite au nombre de résultats possibles qu'on peut vouloir observer dans une comparaison donnée. Toutefois, dans le passé, une limite pratique s'imposait en raison de la somme de travail que représentait la prévision, à partir de la théorie, de ce que seraient les fréquences aléatoires dans un fichier hypothétique de paires d'enregistrements NON APPARIABLES. Cette contrainte ne s'applique désormais plus à Statistique Canada depuis que des méthodes commodes ont été mises au point par un des nôtres (P.L.) qui permettent de créer et d'utiliser facilement des fichiers réels de paires d'enregistrements NON APPARIABLES formées au hasard. C'est ainsi qu'on peut maintenant observer directement les dénominateurs des rapports de fréquence.

Les raisons pour lesquelles on distingue différents niveaux de résultat sont habituellement évidentes. Dans le cas du MOIS DE NAISSANCE et du JOUR DE NAISSANCE, la répartition se fait selon l'importance des écarts (par exemple, 0, 1, 2, 3, 4 et 5 mois ou jours ou plus). Dans le cas de l'ANNÉE DE NAISSANCE, les groupes d'âge peuvent aussi être importants. En ce qui concerne l'ÉTAT MATRIMONIAL, il devrait y avoir au moins neuf niveaux représentant toutes les combinaisons possibles de **jamais marié, marié** et **déjà marié,** à cause du pouvoir de différenciation négatif très élevé des résultats peu probables (par exemple, **marié** suivi de **jamais marié**). Chaque fois qu'on ne

sait pas très bien quelle est la juste mesure entre le raffinement utile et la complexité non souhaitée, les tests empiriques seront d'un grand secours.

Les NOMS et les IDENTIFICATEURS GÉOGRAPHIQUES méritent une attention spéciale parce qu'ils contiennent une bonne partie du pouvoir de différenciation total et que la **ressemblance** et la **connexité** peuvent prendre beaucoup de formes différentes. Il y a beaucoup de plans de comparaison possibles, et les deux exemples que nous donnons ici ne sont pas nécessairement les meilleurs (voir les tableaux 6 et 7).

Tableau 6

Exemple d'une comparaison à plusieurs niveaux:
Prénoms (comparés seulement si les INITIALES concordent)

Concordances Totales	(SAMUEL - SAMUEL)
2-7 caractères, concordance totale	
Troncations	(SAMUEL - SAM)
2, 3, 4-6 concordent + troncation	
Concorde + ne concorde pas	(SAMUEL - SAMPSON)
1, 2, 3, 4-6 concordent + non-concordance	

Tableau 7

Exemple d'une comparaison à plusieurs niveaux:
Lieu de travail par rapport au lieu de décès

Décédé:	Même ville ou lieu
	Autre lieu mais dans la même industrie même province
	Autre province - ouest du Canada
	Autre province - est du Canada
	Autre pays

(une répartition plus fine encore est possible)

Si l'on veut mettre au point de meilleurs plans, il faut observer quels indices sont jugés utiles quand des appariements difficiles sont effectués manuellement. Bien entendu, toute nouvelle procédure doit être testée et, pour passer le test avec succès, elle doit permettre de distinguer les degrés de **connexité** que n'aurait pas fait ressortir l'approche plus grossière qu'elle doit remplacer.

Cet empirisme fait que la logique de la machine reste simple et facile à comprendre. Ici, il y a un parallèle évident à faire avec le pragmatisme du cerveau humain quand il utilise ce dont il se souvient des paires d'enregistrements qu'il a jugés APPARIABLES ou pas.

5. UN EXEMPLE DE RAPPORTS DE FRÉQUENCE OBTENUS EMPIRIQUEMENT

Un seul exemple nous servira à illustrer l'utilisation d'un fichier réel de paires d'enregistrements NON APPARIABLES. Il s'agit des PREMIÈRES et DEUXIÈMES INITIALES, dont la comparaison peut être plus complexe et plus importante qu'on pourrait penser.

1. Le fait que la **DEUXIÈME INITIALE MANQUE** dans les deux enregistrements d'une paire d'enregistrements comporte en soi un pouvoir de différenciation positif qu'il faut utiliser.

2. **Les non-concordances des deux premières et des deux deuxièmes initiales** tendent à être corrélées, et le biais qui en résulte doit être éliminé.
3. **L'inversion des deux INITIALES** est fréquente et oblige à faire des comparaisons avec ordre inversé. Il en est de même des **changements de position,** qui se produisent quand le premier des trois prénoms est utilisé irrégulièrement.
4. **Des PREMIÈRES et des DEUXIÈMES INITIALES qui concordent ou qui concordent avec ordre inversé pourraient être utilisées comme indices** pour indiquer si les autres prénoms doivent être comparés directement ou avec ordre inversé.

Les méthodes antérieures permettaient de faire des comparaisons avec ordre inversé, mais il n'existait pas de façon correcte de calculer les rapports de fréquence dans les cas où des **non-concordances directes** étaient suivies de **concordances avec ordre inversé.** (Un "coup réussi" précédé d'un "coup manqué" n'est pas la même chose qu'un "coup réussi" au premier essai et ne devrait pas être considéré comme tel.)

La solution élégante est de comparer en même temps directement et avec ordre inversé. Cela produit dans l'exemple qui nous intéresse 22 résultats possibles (si l'on ne tient pas compte des enregistrements où il manque les deux INITIALES et si l'on place en première position la DEUXIÈME INITIALE dans les cas où c'est la seule qui paraît), dont bon nombre peuvent être regroupés. On peut calculer leurs fréquences aléatoires respectives de façon théorique, mais il est plus facile de les obtenir directement à partir d'un vrai fichier de paires d'enregistrements NON APPARIABLES.

Des exemples de rapports de fréquence obtenus à partir d'une méthode utilisant simultanément des comparaisons directes et des comparaisons avec ordre inversé sont présentés dans le tableau 8 et les erreurs ainsi évitées sont présentées dans le tableau 9.

Tableau 8

Comparaisons à plusieurs niveaux: 1e et 2e initiales

Résultats	Rapport de fréquence Liens / Pas de lien	=	Probabilité
	%		
Les deux 2e initiales figurent			
2 Concordances	34.60 / .13	=	266 / 1
2 Concordances avec ordre inversé	2.06 / .10	=	21 / 1
1 Concordance, 1 non-concordance	1.51 / 3.48	=	1 / 2
1 Concordance avec ordre inversé	.39 / 3.00	=	1 / 8
Aucune concordance	.07 / 19.10	=	1 / 273
Une 2e initiale figure sur un enregistrement mais pas sur l'autre			
1 Concordance	21.30 / 3.51	=	6 / 1
1 Concordance avec ordre inversé	2.01 / 3.17	=	1 / 2
Aucune concordance	.30 / 43.66	=	1 / 146
Les deux 2e initiales se figurent pas			
1 Concordance	37.15 / 1.69	=	22 / 1
Aucune concordance	.61 / 22.16	=	1 / 36
Fréquences combinées	100.00 100.00		

La méthode permet un meilleur usage du pouvoir de différenciation si elle utilisée de la façon suivante:

a) Une vraie multiplication ou division **par 3** des probabilités, due à **la présence ou à l'absence corrélée** de la DEUXIÈME INITIALE, peut être utilisée avec la méthode de comparaison à plusieurs niveaux.

b) Le biais dû à la fausse multiplication **par 36** des probabilités calculées et attribuable à la **non-concordance corrélée** des INITIALES est évité si l'on enchaîne les initiales.

c) La fausse multiplication **par 12** des probabilités calculées dans le cas de la **concordance avec ordre inversé** des INITIALES quand elles sont comparées directement et comparées avec ordre inversé l'une après l'autre dans une série d'étapes est également évitée si l'on enchaîne les initiales.

Ce sont les vrais fichiers de paires d'enregistrements NON APPARIABLES qui rendent cette sorte de perfectionnement simple et commode à utiliser en pratique.

Tableau 9

Comparaisons à plusieurs niveaux: 1e et 2e initiales

Probabilités vraies (initiales comparées ensemble)
par rapport à
Probabilités fausses (initiales comparées séparément)

	Facteur d'erreur (probabilités vraies / fausses)
Effet de vraie corrélation exploité	
2e initiale figurant sur les deux enregistrements	1.5 / 1
2e initiale ne figurant sur aucun des deux enregistrements	1.6 / 1
2e initiale figurant sur un enregistrement mais pas sur l'autre	1 / 2.1
Effet de fausse corrélation évité	
2 premières et 2 deuxièmes initiales concordent	1 / 1.1
2 premières et 2 deuxièmes initiales ne concordent pas	9.1 / 1
2 initiales concordent, les 2 autres initiales ne concordent pas	1 / 4.0
Fausses probabilités combinées évitées	
concordance avec ordre inversé (par exemple JW-WJ)	1 / 12.5

6. FAIRE QUE LES PROBABILITÉS SOIENT PROPRES À DES VALEURS DONNÉES (MÉTHODE CORRECTE)

Les **probabilités** ou **rapports de fréquence** tels que définis jusqu'ici ne tiennent pas compte du fait que le pouvoir de différenciation ne sera pas le même selon que les noms, les initiales, etc., sont communs ou rares. Il faut donc les convertir pour les rendre propres à des valeurs données (ou spécifiques). Malheureusement, on sait maintenant que la méthode généralement utilisée fausse beaucoup les probabilités quand on l'applique aux **concordances partielles.** Une nouvelle technique a donc été mise au point qui convient

aussi bien aux **concordances totales** qu'aux **concordances partielles.** La meilleure façon de la décrire, c'est de se reporter à la formule du RAPPORT DE FRÉQUENCE vue précédemment (voir tableau 1).

Quand la définition des résultats utilisée dans la formule ne précise pas de nom particulier (ou une partie de nom en particulier), les probabilités seront associées à un nom qui a une fréquence "moyenne", c'est-à-dire ni particulièrement rare, ni particulièrement commun. Il faut donc procéder à un **ajustement, à la hausse** quand on veut que les probabilités soient associées à un nom rare, **à la baisse** quand on veut qu'elles soient associées à un nom commun. En fait, la probabilité devrait refléter la rareté ou la fréquence relative d'un nom particulier (ou de sa partie qui concorde) par rapport à un nom "moyen".

Aucune partie du rapport de fréquence "global" n'est perdue parce que l'information contenue dans le rapport est irremplaçable. (Ces rapports de fréquence non propres à des valeurs données sont appelés "globaux" parce qu'ils englobent toutes les définitions possibles de résultat, y compris les composantes **concordance** et **non-concordance.**) Au lieu de laisser tomber quelque information que ce soit, nous multiplions tout simplement les rapports par le **facteur d'ajustement** qui convient.

Les détails sont simples. La fréquence "moyenne" ou **fréquence générale,** par exemple pour les prénoms masculins, est une moyenne pondérée des fréquences de toutes les valeurs. Elle est habituellement fondée sur le fichier qu'on explore et correspond à la somme des carrés de chacune des fréquences. Les **fréquences générales** sont obtenues séparément pour 1, 2, 3, 4 et 7 caractères. Le **FACTEUR D'AJUSTEMENT** pour la partie concordante d'un nom particulier sera égal à la **fréquence générale** divisée par la **fréquence spécifique.** Le facteur d'ajustement sera plus grand qu'un dans le cas des noms, initiales, etc., rares et plus petit qu'un dans le cas des valeurs communes, comme on pouvait s'y attendre (voir tableau 10).

Tableau 10

Exemples de facteurs d'ajustement spécifiques

Valeur	Frequence générale	Frequence specifique	Facteur d'ajustement 6 en/specif
A (initiale)	1 / 10	1 / 9.2	1 / 1.1
Amos (1)	1 / 11.5	1 / 249.0	21.7 / 1
Amos - Combiné (2)	1 / 115	1 / 2280.8	19.8 / 1
J (initiale)	1 / 10	1 / 6.4	1 / 1.6
John (1)	1 / 11.5	1 / 2.9	1 / 4.0
John - Combiné (2)	1 / 115	1 / 18.6	1 / 6.2

(1) Comparativement à tous les prénoms commençants par cette initiale
(2) Comparativement à tous les prénoms

Sur le plan conceptuel, on peut considérer qu'un rapport de fréquence propre à une valeur donnée repose uniquement sur des paires d'enregistrements dans lesquelles l'enregistrement du fichier parcouru comporte en soi la valeur spécifiée pour la partie de l'identificateur qui concorde. En principe, les dénominateurs pourraient ensuite être obtenus à partir de très gros fichiers de paires d'enregistrements NON APPARIABLES.

Toutefois, dans le cas où ni la définition du résultat (par exemple, = "concordance de 4 caractères suivie d'une troncation relative", comme dans ALEX - ALEXANDER) ni la valeur particulière (par exemple, partie du nom qui concorde = ALEX) ne sont communes, il faudra un fichier de paires d'enregistrements NON APPARIABLES excessivement gros. Il ne fait aucun doute que des fichiers de 50,000 paires comme ceux qu'utilisent Statistique Canada sont beaucoup trop petits.

Par conséquent, il faut une méthode correcte permettant de convertir des **rapports de fréquence globaux** (et de leurs logarithmes, appelés **poids globaux**) en leurs équivalents propres à des valeurs données pour pouvoir effectuer toute opération d'appariement par ordinateur qui tienne compte des niveaux de résultat multiples.

Même s'ils sont corrects en général, les FACTEURS D'AJUSTEMENT peuvent être raffinés si on les applique à des cas bien précis. La méthode proposée, et qu'on décrit ici, utilise exactement les mêmes fréquences générales et les mêmes fréquences spécifiques (et leurs poids) que celles qu'on calcule selon les méthodes actuelles, c'est-à-dire que pour des raisons de simplicité on ne fait pas de différence entre les noms d'une longueur donnée et les noms de même longueur une fois tronqués. Il est donc possible de raffiner les FACTEURS D'AJUSTEMENT, mais il faut pour cela d'autres tableaux de référence.

7. LA MÉTHODE DE CONVERSION INCORRECTE

La méthode de conversion incorrecte des probabilités globales en leurs équivalents propres à des valeurs données ne présente pas seulement un intérêt local. Dans quelque pays que ce soit, les personnes qui font de l'appariement d'enregistrements sont susceptibles de se trouver devant le même problème logique. On a toujours considéré, à tort, qu'un raccourci acceptable pour les **concordances totales** pouvait aussi convenir pour des **concordances partielles.**

Le raccourci dont il s'agit ici consiste à remplacer **le dénominateur d'un rapport de fréquence global** (c'est-à-dire la fréquence d'un résultat donné dans des paires d'enregistrements NON APPARIABLES) par une **fréquence propre à une valeur donnée** obtenue à partir du fichier exploré (et dont le logarithme est appelé **poids de fréquence**). Mais la fréquence abandonnée est associée à des **enregistrements appariés,** tandis que la fréquence qui la remplace est associée à des **enregistrements non appariés,** de sorte que ces fréquences sont loin d'être équivalentes même quand elles sont numériquement égales comme dans les **concordances totales.**

Si le raccourci ne fonctionne pas avec les **concordances partielles,** c'est parce que la définition des résultats inclut à la fois une composante de **concordance** et une composante de **non-concordance.** Jamais une fréquence obtenue à partir d'**enregistrements non appariés** ne peut tenir compte de la partie **dissemblance** de la définition d'un résultat ayant trait à des **enregistrements appariés.**

Pour voir en quoi les deux méthodes diffèrent, appliquons chacune d'elles à des **concordances partielles** des prénoms JOHN et JOHNNY et AMOS et AMOSE. Dans ces deux cas, la définition globale est "une concordance des quatre premiers caractères suivie d'une **troncation** relative" (voir tableau 11). Les données nécessaires, calculées comme on le fait actuellement, sont:

Rapport de fréquence global (4 premièrs conc.)= 1.4% / 0.024% = 58.3 / 1.

Fréquence générale (4 premiers caractères) = 0.87%

Fréquences spécifiques: AMOS = 0.037%; JOHN = 5.304%.

Tableau 11

Comparaison de la méthode de conversion correcte et de la méthode de conversion correcte

Partie concordante (plus partie non concordante)	Rapport de fréquence Liens / pas de lien	=	Probabilité
	%		
Méthode incorrecte (dénominateur a été changé pour fréquence spécifique)			
Globale 4-conc (+ Troncation)	1.4 / 0.024	=	58.3 / 1
Amos (espace vide - E)	1.4 / 0.037	=	37.6 / 1
John (espace vide - NY)	1.4 / 5.304	=	1 / 3.8
Méthode correcte (multiplication par le facteur d'ajustement spécifique)			
Globale 4-conc (+ troncation)	1.4 / 0.024	=	58.3 / 1
Amos (espace vide - E)	(fois 0.87 / 0.037)	=	1370.8 / 1
John (espace vide NY)	(fois 0.87 / 5.304)	=	9.6 / 1

Ici, la méthode incorrecte produit des probabilités **36 fois** trop faibles. (Par exemple, une valeur rare comme AMOS devrait faire augmenter les probabilités, non les faire baisser). Cette erreur tend à être plus grande avec la multiplicité des niveaux de résultat. Plus précisément, cette erreur se produit chaque fois que la **fréquence générale** (0.87% dans ce cas-ci) diffère de la **fréquence globale** dans les paires d'enregistrements NON APPARIABLES (0.024%) et est égale au ratio des deux fréquences (0.87% / 0.024% = **36.25 fois).**

Ce cas n'est **pas** un exemple extrême. Quand on prend divers niveaux de **concordance partielle,** on peut voir que la méthode de conversion incorrecte introduit un biais dans les probabilités qui peut aller de 2 à 2000 fois tableau 12). Environ 17 % des comparaisons de PRÉNOMS sont touchées dans les paires d'enregistrements appariés. (Si on prend les NOMS DE FAMILLE, la proportion augmente à 20 pour cent.)

Tableau 12

Grandeurs des erreurs dues à des conversions incorrectes (En comparant des prénoms de 2 à 7 caractères; les initiales concordent)

niveau de concordance	fréquence générale (%)	fréquence globale (pas de lien) (%)	pourcentage d'erreur affectées (gén./glob.)	paires appariées (%)
concordance totale	.98 est.	.982	1.0-fois	83.1
4-6 conc + tronc	.87	.024	36.3-fois	1.6
3 conc + tronc	1.96	.014	140. -fois	.3
2 conc + tronc	3.69	.0016	2306.3-fois	.1
4-6 conc + non-conc	.87	.085	10.2-fois	4.2
3 conc + non-conc	1.96	.142	13.8-fois	3.3
2 conc + non-conc	3.69	1.383	2.7-fois	2.2
1 conc + non-conc	9.85	4.395	2.2-fois	5.2

La mécanique n'est pas tellement différente d'une méthode à l'autre. La méthode incorrecte utilise des tableaux de référence de **fréquences spécifiques** (dont les logarithmes sont appelés **poids de fréquence)** pour les diverses valeurs des noms, initiales, etc., et pour leurs versions abrégées. Dans la nouvelle méthode, qui elle est correcte, il est seulement nécessaire que ces tableaux de référence contiennent à la place des fréquences spécifiques les **facteurs d'ajustement spécifiques** (ou leurs logarithmes, appelés **poids d'ajustement spécifiques)** et qu'on garde les rapports de fréquence globaux intacts et ajustés par les facteurs appropriés.

Bref, en convertissant correctement les **rapports de fréquence globaux** (et **les poids globaux)** en leurs équivalents propres à des valeurs données, on évite, sans coût ni inconvénient additionnels, de répéter les grosses erreurs qu'on faisait habituellement en utilisant l'ancienne méthode qui était incorrecte. Les poids eux-mêmes et les facteurs d'ajustement peuvent encore être raffinés, mais il s'agit d'une autre question qu'on ne traite pas ici.

Ce qui rend ce test simple et concluant, c'est l'utilisation de fichiers réels de paires d'enregistrements NON APPARIABLES formées au hasard pour observer directement la fréquence selon laquelle les divers résultats de **concordance partielle** se produiront par pur hasard. En réalité, la nature et la grandeur de l'erreur n'ont été découvertes que parce que l'utilisation de fichiers réels de paires d'enregistrements NON APPARIABLES permet d'éviter de recourir à des expressions mathématiques compliquées et ainsi de se représenter plus facilement ce qui ne fonctionnait pas bien.

8. LE BUT DU RAFFINEMENT

Le but visé par presque tous les raffinements des méthodes d'appariement est nécessairement modeste, parce que la majorité des enregistrements pouvant être appariés seront correctement appariés quelque grossière que soit l'approche utilisée et parce qu'il y aura toujours des enregistrements qu'on ne pourra pas apparier, si élégantes que soient les méthodes utilisées. Par conséquent, les raffinements n'influeront sur le résultat final que si le pouvoir de différenciation nécessaire est présent dans les deux enregistrements qu'on veut apparier mais difficile à faire ressortir. Il s'ensuit que les mesures de la précision globale de n'importe quelle méthode d'appariement ont tendance à en dire plus sur l'état des fichiers que sur l'efficacité des méthodes.

Malgré cela, on a encore de bonnes raisons pour améliorer l'efficacité des méthodes d'appariement chaque fois que cela est possible sans que les coûts soient excessifs. Les deux principales sources d'inefficacité sont les **erreurs logiques** et l'**incapacité d'exploiter le pouvoir de différenciation.** Dans le passé, ces deux sources d'inefficacité étaient associées à la trop grande importance qu'on donnait aux résultats simples de **concordance et de non-concordance.**

Les erreurs logiques et les possibilités inexploitées sont habituellement découvertes à l'examen visuel des liens plus difficiles. Ces examens permettent d'observer les étapes logiques qui s'enchaînent dans notre cerveau. Ce type de raisonnement intuitif et partiellement inconscient est moins influencé par les contraintes artificielles que ne l'est un raisonnement formel. Mais ce qui est plus important, c'est que chaque fois que le cerveau fait manifestement un meilleur travail que l'ordinateur, cela vaut la peine de chercher à savoir pourquoi. Le secret réside habituellement dans une certaine forme de procédé mental inconscient au départ. Une fois le secret révélé, on peut la plupart du temps donner des instructions à l'ordinateur pour qu'il l'utilise.

Un des buts réalistes qu'on pourrait se donner serait donc de raffiner les méthodes d'appariement automatisé, à mesure que l'usage des anciennes méthodes devient courant et que des progrès sont réalisés, jusqu'à ce qu'elles égalent ou surpassent les capacités d'un esprit perspicace.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

LE COUPLAGE DES ENREGISTREMENTS ET SES APPLICATIONS

MIKE EAGEN et TED HILL[1]

RÉSUMÉ

Le couplage des enregistrements est une discipline bien établie. Tout au long des années 70 et 80, de nombreux perfectionnements ont été apportés: des règles de couplage améliorées ont permis de mieux utiliser l'information disponible; la comparaison croisée des zones s'est répandue; on a élaboré des procédures pour traiter des zones enfreignant les suppositions relatives à l'indépendance des données; on a élaboré d'autres modes de pondération; et, au Canada, on a élaboré un système informatique général permettant de diminuer les dépenses à cet égard.

Tous ces progrès ont amené un niveau de perfectionnement proche de la limite possible, compte tenu des données disponibles. Il reste quand même un écart entre la méthodologie disponible et son utilisation.

Les défis encore à relever par les experts en couplage sont:

i) l'élaboration d'approches méthodologiques intégrées; et

ii) l'amélioration des outils informatiques de manière à permettre la mise en oeuvre sans complication des approches intégrées.

Nous proposons une stratégie pour relever ces défis.

1. INTRODUCTION

Le couplage des enregistrements consiste habituellement à comparer les données de deux fichiers d'entrée pour voir s'ils représentent la même personne ou la même entité. Le présent article ne touche que ce genre d'application; en fait, nous nous sommes limités aux situations où chaque enregistrement est couplé à un seul ou à aucun enregistrement de l'autre fichier. Cette comparaison nécessite l'application d'un certain nombre de règles de couplage. Une règle de couplage peut être très simple (p. ex., comparer les noms de famille) ou plus complexe (comparer les adresses, chacune occupant un certain nombre de zones dans le fichier). L'application de toute règle de couplage produit un résultat. La comparaison des noms de famille, par exemple, peut donner les résultats suivants:

i) le nom de famille est le même;

ii) les quatre premiers caractères et le nom codé phonétiquement (p. ex., NYSIIS, SOUNDEX, **Name Search Key**) sont les mêmes;

iii) le nom codé est le même;

[1] Mike Eagen, Goss, Gilroy & Associates Ltd., 400-222 rue Queen, Ottawa, Canada K1P 5V9. Ted Hill, Statistique Canada, 2405 Edifice Principal, Ottawa, Canada K1A 0T6.

iv) le premier caractère est le même;

v) le nom n'est pas le même;

vi) le nom de famille manque dans l'un ou l'autre fichier.

Les deuxième, troisième et quatrième résultats ci-dessus représentent ce qu'on appelle généralement une concordance partielle. Les règles de couplage peuvent aussi donner une valeur à certains résultats. Si l'on compare deux fichiers où le nom de famille est SMITH, le résultat est la CONCORDANCE et la valeur est SMITH.

Dans la pratique, il n'est ni réaliste ni nécessaire d'effectuer toutes les comparaisons possibles. Pour réduire les frais, on n'applique pas les règles de couplage à toutes les paires d'enregistrements. Certaines paires d'enregistrements sont rejetées après application de seulement quelques règles de couplage.

Aux paires d'enregistrements qui restent, on attribue des poids ou coefficients de pondération fondés sur les résultats de chaque règle de couplage et des valeurs qui en découlent. La pondération totale est calculée pour la paire d'enregistrement et la décision de coupler est fondée sur ce calcul. (Les paires d'enregistrements dont la pondération est élevée sont considérées comme étant manifestement des couplages; celles dont la pondération est peu élevée sont considérées comme n'étant manifestement pas des couplages. Il y a ordinairement un groupe, au milieu, de couplages possibles.)

L'objectif du praticien est de minimiser le nombre de couplages possibles sans accroître le risque de décisions erronnées.

2. HISTORIQUE DU COUPLAGE DES ENREGISTREMENTS

Le couplage des enregistrements a une longue et glorieuse histoire. Les premiers travaux ont été effectués dans les années 50 (Newcombe, H.B., Kennedy, J.M., Axford, S.L. et James, A.P., 1959). Durant les années 60, l'ordinateur est devenu l'outil fondamental de couplage des enregistrements. Durant la même décennie s'est développée une théorie du couplage des enregistrements, qui a atteint son apogée en 1969 dans le document classique d'I.P. Fellegi et A.B. Sunter, qui appuyait les praticiens dans leur travail et présentait les techniques à un public beaucoup plus vaste.

Le couplage des enregistrements a pris passablement d'ampleur au cours des années 70. Les applications sont devenues plus nombreuses, surtout dans le domaine de l'épidémiologie, où le couplage des enregistrements est devenu une technique fondamentale de détermination du risque environnemental. Cette technique a rendu possible des recherches qui autrement n'auraient pas pu être effectuées, ce qui a contribué à épargner des vies et à alléger les soucis des travailleurs et de leurs familles. L'épidémiologie reste d'ailleurs une très importante application du couplage des enregistrements. Toujours au cours des années 70, les agences statistiques centrales telles que Statistique Canada ont commencé à se servir de cette technique pour constituer et dédoubler des plans d'enquête et des registres centraux des entreprises et des exploitations agricoles. La croissance des applications a nécessité la mise au point de systèmes informatiques puissants et souples tels que le Système itératif généralise de couplage d'enregistrements (Hill, T. et Pringmill, F., 1985). La méthodologie a été améliorée à plusieurs reprises au cours des années 70. Plus précisément, on a élaboré des règles de comparaison qui donnaient des résultats multiples, permettant ainsi de mieux utiliser la puissance discriminante des données (Newcombe, Fair, Lalonde, 1987). Le couplage des enregistrements est donc apparu comme une véritable discipline en ce sens qu'il est devenu un domaine d'exercice dominé par des spécialistes qui, forts de leur expérience et de leurs connaissances, pouvaient obtenir des résultats sensiblement supérieurs à ceux de novices.

L'amélioration de la méthodologie s'est poursuivie dans les années 80 et, à notre avis, elle a pratiquement atteint son apogée. En outre, des groupes d'experts très qualifiés dans l'application de cette méthodologie raffinée se sont constitués. Au Canada, par exemple, la Division de la santé de Statistique Canada et la Fondation ontarienne du cancer ont mis sur pied des unités chargées principalement d'appliquer le couplage des enregistrements aux problèmes épidémiologiques. Grâce à leur vaste expérience et à leurs moyens raffinés, ces organismes peuvent présenter des résultats rigoureusement scientifiques qui sont totalement acceptés dans le milieu médical. Néanmoins, le couplage des enregistrements reste pour la plupart des praticiens une technique très difficile, où le risque d'erreur est élevé. La difficulté découle de la pondération, surtout lorsque sont appliquées des règles complexes de couplage. Plus précisément, nous estimons que les stratégies et les moyens de pondération n'ont pas du tout avancé au même rythme que les autres domaines, par exemple en ce qui a trait à une plus grande discrimination dans les règles. Malheureusement, ce retard signifie qu'il est très difficile, sauf pour les praticiens les plus avancés, de concrétiser les améliorations possibles que promettent les progrès de la méthodologie.

Dans le présent document, nous décrivons et nous évaluons les diverses approches à la pondération dans quelques situations ordinaires; nous expliquons la nécessité d'une approche coésive plutôt que fragmentée à la pondération; nous décrivons celle que nous préconisons (ainsi que la nécessité de l'appliquer de façon cohérente); nous proposons des orientations pour la recherche à venir, et enfin, nous présentons nos conclusions.

3. LA THÉORIE DE LA PONDÉRATION

La théorie fondamentale de la pondération est simple. Il faut en premier lieu que les règles du couplage soient définies de manière à être statistiquement indépendantes, c'est-à-dire de sorte que les probabilités de divers résultats d'une règle quelconque ne dépendent pas des résultats des autres règles. Ainsi, on peut obtenir des pondérations pour chaque règle, indépendamment des autres. Pour la règle i et le résultat j, la pondération est donnée par le rapport de probabilité suivant:

$$W_{ij} = \frac{\text{Prob(j/les enregistrements sont couplés)}}{\text{Prob(j/les enregistrements ne sont pas couplés)}} \qquad (1)$$

Il est clair que les résultats qui sont plus probables entre couplages auront une pondération supérieure à un. Le résultat plus susceptible de résulter de l'absence de couplage, au contraire, aura une pondération entre 0 et 1. Si l'on prend les logarithmes, ce qui est l'usage normal, ces pondérations seront positives et négatives respectivement.

La pondération totale de deux enregistrements quelconques mis en comparaison sera simplement le produit des pondérations (1) (ou la somme des logarithmes) pour toutes les règles de couplage.

Le problème, bien sûr, consiste à estimer les probabilités à utiliser pour calculer les pondérations. On a eu recours à deux approches fondamentales. La première, suggérée par Fellegi et Sunter (1969), consiste à utiliser les formules théoriques. La deuxième, suggérée par de nombreux auteurs dont Newcombe, Fair et Lalonde (1987), comporte l'estimation des probabilités directement à partir de fichiers-échantillons de couplages et de non-couplages, créés à cette fin. Cette méthode est celle que nous appelons la méthode directe. Les avantages et les inconvénients relatifs de ces deux approches sont illustrées le mieux si l'on examine divers genres de règles de couplage.

3.1 Concordance exacte

On voit aisément que

Prob (concordance exacte sur la valeur k | articles couplés)

= Prob (concordance exacte | articles couplés) × Prob (valeur k | concordance exacte | articles couplés)

= Prob (concordance exacte | articles couplés) $\times \frac{n_L(k)}{N_L}$

où, $n_L(k)$ = fréquence de la valeur k parmi les couplages

N_L = nombre de couplages.

En l'absence d'erreur dans les données, on voit aisément que

Prob (concordance exacte ou valeur k | articles non couplés)

= Prob (valeur k pour le fichier A) × Prob (valeur k pour le fichier B)

En pratique, bien sûr, les fichiers renferment des erreurs. Toutefois, comme les erreurs se présentent dans les deux sens, on peut raisonnablement supposer qu'elles s'annulent mutuellement et que l'utilisation de $RF_A(k) \times RF_B(k)$ est raisonnable.

Si l'on substitue ces résultats dans (1), on obtient

$$W_K = \frac{RF_L(k) \times \text{Prop. (Concordance exacte parmi les couplages)}}{RF_A(k) \times RF_B(k)} \qquad (2)$$

où

k représente la valeur convenue (p. ex., si les règles concernaient la comparaison des noms de famille, alors le (2) donne la pondération en cas de concordance exacte quant à la valeur k, disons Philippe);

$RF_L(k)$ est la fréquence relative de la valeur k parmi les couplages réels et n'est généralement pas connue à l'avance;

$RF_A(k)$ et $RF_B(k)$ sont les fréquences relatives des deux fichiers à être couplés; et

Prop. (concordance exacte entre les couplages) représente la proportion de couplages réels, qui ont eu pour résultat la concordance exacte.

Eagen (1978) traite de cette situation d'une manière plus détaillée.

Par convention, le plus petit des deux fichiers en voie de couplage est appelé le fichier A. Dans beaucoup d'applications, particulièrement dans le domaine de l'épidémiologie, le fichier A est beaucoup plus petit que le fichier B. Comme la proportion d'enregistrements du fichier A qui seront couplés est beaucoup plus élevée que celle du fichier B, il est préférable d'estimer à partir du fichier A la répartition des fréquences des valeurs attribuables au couplage. Dans de nombreux cas, il est parfaitement raisonnable de supposer que $RF_L(k) = RF_A(k)$, ce qui donne

$$W = \frac{\text{Prop. (concordance exacte parmi couplages)}}{RF_B(k)} \qquad (3)$$

Il est bien simple de tirer du grand fichier la valeur $RF_B(k)$ pour tous les k. La proportion de concordance exacte entre couplages pour le domaine en question peut être estimée par itération à mesure que l'on converge vers les vrais couplages. À Statistique Canada, le système GIRLS nous donne cette possibilité. D'ordinaire, cette proportion sera proche de un et l'on peut en toute quiétude n'en tenir aucun compte dans les conditions suivantes:

- toutes les règles de couplage concernent une seule zone et n'ont pour résultat que la concordance exacte, la non-concordance et l'absence; et
- la proportion de concordance exacte entre les couplages réels ne varie pas appréciablement d'une zone à une autre.

Dans ces circonstances, les pondérations de la concordance exacte dépendent uniquement des fréquences au sein du fichier B, le plus grand des deux. C'est un résultat significatif, qui simplifie considérablement la pondération. Il est arrivé souvent, cependant, que les résultats soient mal appliqués.

L'approche directe

Cette approche directe estime les deux probabilités en (1) en déterminant les proportions de cas de concordance exacte pour les couplages et les non-couplages (non compte tenu de la valeur convenue) et en ajustant les pondérations suivant la fréquence de la valeur convenue. Les tenants de cette approche n'ont pas décrit leur méthode mathématiquement mais la formule (4) ci-dessous la décrit bien.

$$W_K = \frac{\text{Prop. (concordance exacte parmi les couplages)}}{\text{Prop. (concordance exacte parmi les non-couplages)}} \times AF(k) \qquad (4)$$

où

$$AF(k) = \text{Facteur de correction pour la valeur k}$$

$$= \frac{GF_B}{RF_B(k)}$$

où

$$GF_B = \text{Fréquence générale de concordance exacte}$$

$$= \sum_k RF_B(k)^2$$

Il serait plus juste d'utiliser

$$GF_{AB} = \sum_k RF_A(k) \times RF_B(k)$$

mais c'est beaucoup plus difficile en pratique et ne se fait pas communément.

On obtient la proportion de concordance exacte entre non-couplages en créant au hasard un fichier de comparaisons à partir des fichiers d'entrée. Étant donné que pratiquement aucune comparaison de l'échantillon ne donnera lieu à un couplage, cette

approche, dite celle des paires incouplables, est bonne. La proportion de concordance exacte entre couplages est estimée de façon itérative comme ci-dessus.

Observations

Dans ce cas, les deux méthodes sont théoriquement équivalentes. (Strictement, l'équivalence théorique n'est vraie que seulement si l'on utilise GF_{AB} à la place de GF_B pour calculer $AF(k)$. Sur le plan pratique, toutefois, les deux méthodes sont essentiellement équivalentes même si l'on se sert de GF_B.) On s'est souvent servi de l'approche théorique en pratique parce qu'elle est relativement simple à utiliser dans cette situation. L'approche directe ne présente aucun avantage pour la pondération des concordances exactes et nécessite en plus la création et l'analyse du fichier de paires incouplables.

3.2 Concordance Partielle

Les praticiens expérimentés savent que, même s'il n'y a pas concordance exacte, certains résultats appuient le couplage mieux que d'autres. Si l'on compare les prénoms, par exemple, les comparaisons de Phil et Philip, Phil et Philippe, et Phillip et Philippe justifient plus clairement un couplage qu'un résultat comme Phillip et George. Il arrive donc communément que l'on définisse des résultats de concordance partielle qui tiennent compte de situations de ce genre. Si l'on définit la concordance des quatre premiers caractères comme une concordance partielle, les trois premières comparaisons ci-dessus sont en fait des concordances partielles. Cette pratique est évidemment bonne, mais elle rend quand même la pondération plus difficile. Même si cette difficulté n'est pas extrême, beaucoup d'erreurs se commettent dans la pratique. Voyons ce que chaque approche peut offrir.

Approche théorique

Eagen (1978) a démontré que la formule de pondération convenant à cette situation est la suivante:

$$W_m = \frac{RF_L(m) \times \text{Prop. (Concordance partielle parmi les couplages)}}{RF_A(m) \times RF_B(m) - \sum_{k \in m} RF_A(k) \times RF_B(k)} \qquad (5)$$

où m est la valeur dont on a convenu partiellement (p. ex. Phil), k est une valeur qui renferme m (p. ex. Philip) et tous les autres éléments sont définis comme antérieurement.

Le terme de sommation au dénominateur sert à compenser l'absence de concordance exacte. (**Nota:** Si le nom Phil se trouve dans les deux fichiers en tant que prénom complet, la concordance est exacte).

Cette formule est plus problématique que la formule (2). En premier lieu, la proportion au numérateur n'est généralement pas proche de un et **doit** être estimée de façon itérative. Deuxièmement, quelle que soit l'hypothèse, la formule ne se simplifie pas par voie d'annulation. Troisièmement, il faut maintenant supposer que $RF_L = RF_A = RF_B$ si l'on doit recourir à un seul fichier pour effectuer les calculs. Ceci représente beaucoup plus que l'hypothèse $RF_L = RF_A$. Quatrièmement, le calcul de ces pondérations à l'échelle du système est plus complexe que le calcul des pondérations en cas de concordance exacte. Eagen (1978) donne des spécifications applicables à cette situation mais, à notre connaissance, celles-ci n'ont jamais été appliquées. Dans beaucoup de cas, on a appliqué la formule relative à la concordance exacte même si l'on sait qu'elle est fautive et,

comme Newcombe, Fair et Lalonde (1987) l'ont démontré, sous-estime la pondération (quelquefois gravement).

Nous notons que la formule ci-dessus s'applique également aux niveaux multiples de concordance partielle. Le calcul à l'échelle du système, cependant, se complique en cas de niveaux multiples de concordance partielle.

Approche directe

L'approche directe a pour grand avantage que la méthodologie ne dépend pas de la nature des résultats. La méthode de pondération des concordances partielles est la même que dans le cas des concordances intégrales.

La formule est la suivante:

$$\text{Pondération} = \frac{\text{Prop. (concordance partielle parmi les couplages)}}{\text{Prop. (concordance partielle parmi les non-couplages)}} \times AF(m) \quad (6)$$

où

$$AF(m) = \frac{GF_B}{RF_B(m)}$$

Il faut noter qu'en calculant les proportions, concordance partielle signifie bien partielle, et non pas concordance complète.

Cette approche donne des pondérations qui diffèrent dans une certaine mesure des pondérations théoriques. Il est facile de démontrer que les pondérations de (6) seront les mêmes, en moyenne, que celles qui résulteront de (5), même si la pondération pour toute valeur donnée, m, différera probablement. Nous ne croyons pas que ces différences importent tellement en pratique, mais on a effectué peu de recherche sur la question.

L'approche intuitive ou empirique

Ici, les pondérations pour les concordances partielles sont fixées quelque part entre la pondération globale ou moyenne de fréquence de concordance et la pondération de non-concordance. Cela semble tout à fait sensé et, intuitivement, est préférable à l'utilisation de règles de couplage plus simples, sans concordance partielle. Avec des intuitions comme les nôtres, cependant, cette approche peut être problématique. La gamme des choix est souvent vaste et il faut posséder une certaine connaissance des probabilités sous-jacentes, sinon on risque d'empirer une situation déjà peu claire. La sélection des pondérations peut être particulièrement problématique lorsqu'on se sert de niveaux multiples de concordance partielle. On obtient les meilleurs résultats en prenant les coefficients de rectification utilisés dans l'approche directe.

Observations

L'approche directe est évidemment la meilleure dans ce cas. Elle garantit un degré de rigueur qui manque à l'approche intuitive. En outre, elle est plus facile à utiliser que l'approche théorique. Le principal obstacle à son utilisation est l'absence de moyens aisément disponibles.

3.3 Règles plus complexes

Les praticiens expérimentés ont jugé valable d'élaborer des règles qui couvrent une vaste gamme de situations additionnelles. Les cas de concordance partielle ont été

répartis entre "concordance + élément manquant" (p. ex., Phil - Philip) et "concordance + non-concordance" (p. ex., Phillip - Philippe) car on estime que la situation "concordance + élément manquant" appuie mieux le couplage. Une autre technique commune est la comparaison croisée (p. ex., comparer le premier prénom du fichier A au deuxième prénom du fichier B). Un troisième type de règle consiste à construire une règle unique de couplage englobant plusieurs zones apparentées (p. ex., les données de l'adresse). Ce genre de comparaison est utile lorsque les codes postaux correspondent mais ne précise que la municipalité; il serait erroné d'assigner une pondération positive lorsque seulement la municipalité concorde. D'autre part, si les codes postaux ne concordent pas, la concordance au niveau de la municipalité peut très bien contribuer utilement à la décision de coupler ou non.

Beaucoup de ces règles sont parfaitement sensées et peuvent être très utiles dans la prise de la décision de coupler dans les cas rapprochés. D'autre part, si ces règles ne sont pas correctement pondérées, elles peuvent rendre pire une situation déjà confuse. Pis encore, elles peuvent transformer une situation claire en zone grise. (Comme, normalement, ces règles attribuent une pondération positive à une situation qui autrement ne serait pas reconnue, le principal danger consiste à donner à des non-couplages des pondérations totales équivalant à certains couplages réels). Étant donné que la résolution des zones grises est la principale raison de pondérer (tout algorithme peut régler les cas faciles), une stratégie juste de pondération est extrêmement importante pour ces situations (même si celles-ci sont relativement peu fréquentes).

D'ordinaire, l'approche théorique a peu à offrir dans ces situations puisqu'une formulation convenable des probabilités sous-jacentes ne peut généralement pas se faire sur une base théorique.

L'approche directe, au contraire, est très prometteuse. Comme elle se limite à observer la fréquence des résultats observés parmi les couplages et les non-couplages, la complexité de la règle n'influe aucunement sur le degré de difficulté d'établir la pondération.

L'approche intuitive ou empirique est applicable ici également par des praticiens qui connaissent **très bien** les données. Notre expérience de cette approche n'est pas encourageante et nous ne la recommandons pas.

La formule pour ces situations est la suivante:

$$\text{Pondération} = \frac{\text{Prop. (résultat parmi couplages)}}{\text{Prop. (résultat parmi non-couplages)}} \times \text{AF(valeur)} \qquad (7)$$

Il s'agit simplement d'une version généralisée de (4) et (6).

Dans certains cas, on laisse tomber le facteur de rectification et l'on se sert de la formule simplifiée ci-dessous:

$$\text{Pondération} = \frac{\text{Prop. (résultat parmi couplages)}}{\text{Prop. (résultat parmi non-couplages)}} \qquad (8)$$

Cette formule donne ce qu'on appelle d'ordinaire des pondérations globales ou moyennes. Les pondérations globales sont celles qui servent ordinairement pour le couplage initial. Une fois les non-couplages évidents éliminés, il est habituellement utile de substituer les pondérations rectifiées [(4), (6) et (7)]. Toutefois, si les valeurs possibles ont toutes des fréquences semblables (p. ex., mois de naissance), les facteurs de rectification seront aussi semblables et le praticien expérimenté s'en passera pour diminuer le coût. Dans d'autres cas, où des règles très complexes entrent en jeu, il serait utile d'incorporer les facteurs de rectification, mais il se peut en pratique qu'il soit trop

difficile de les calculer. La règle relative aux adresses mentionnées dans le premier alinéa de la présente section est un cas de ce genre.

3.4 Pondérations de non-concordance

Nous avons évité dans les sections précédentes toute mention des pondérations de non-concordance. Celles-ci sont ordinairement faciles à obtenir. Eagen a élaboré en 1978 des formules pour l'approche théorique. Si l'on se sert de l'approche directe, la non-concordance est un résultat comme un autre, qui a l'avantage additionnel de ne nécessiter aucun facteur de rectification.

4. SÉLECTION D'UNE STRATÉGIE GLOBALE

La situation idéale en matière de couplage d'enregistrements est bien sûr celle où l'on a beaucoup de données communes, alliées à l'expérience d'une situation semblable. En pareil cas, il est tout à fait réalisable de se servir uniquement de règles de concordance et de non-concordance, qui sont relativement faciles à pondérer. S'il faut utiliser des règles plus complexes, on peut déduire les pondérations à partir de l'expérience passée. Toutefois, la réalité n'est habituellement pas aussi simple.

Dans une situation plus plausible, on trouvera un certain nombre de simples règles de concordance et de non-concordance, un certain nombre de concordances partielles et un certain nombre de cas, plus complexe, nécessitant des comparaisons croisées ou la comparaison de zones multiples. Il nous arrive souvent, aussi, de commencer avec des règles simples pour se rendre compte plus tard que des règles plus complexes seraient utiles pour résoudre les cas douteux.

La discussion, à la section 2, de règles de couplage individuelles, peut porter le lecteur à conclure que les choix se font assez simplement. Il faut toutefois se rappeler que c'est la pondération totale de la comparaison qu'il faut utiliser pour décider du couplage. Comme on l'a indiqué plus tôt, la pondération totale est le produit des pondérations individuelles (ou la somme des logarithmes). De toute évidence, si l'on définit pour une règle des pondérations inexactes ou des pondérations qui ne sont pas cohérentes avec les autres, on peut détruire l'intégrité de tout le processus.

Il faut évidemment élaborer une stratégie globale qui assure le maintien de l'intégrité. C'est malheureusement très difficile. Nous avons noté particulièrement que les progrès des dernières années (c.-à-d. le perfectionnement des règles de couplage) ont grandement accru cette difficulté.

Nous avons vu des couplages (en fait, nous avons effectué des couplages) où les trois stratégies de pondération ont été appliquées à diverses règles. Prise individuellement, la stratégie de pondération pour chaque règle est bonne. En outre, chaque règle utilise manifestement de l'information additionnelle et devrait améliorer le couplage.

Pourtant, sans raison apparente, ça ne fonctionne habituellement pas ensemble. Il devient clair, au moment où le projet tire à sa fin, qu'une règle relativement sans importance compte plus qu'une autre, plus importante. Le praticien consciencieux rectifie les pondérations et bute sur encore deux ou trois règles (dont souvent celles qu'il vient de rectifier) qui lui posent toujours le même problème. On peut toujours continuer à rectifier mais, malheureusement, la convergence ordinairement nous échappe.

Nous croyons que la sélection d'une stratégie globale de pondération constitue l'élément le plus important, le plus difficile et souvent le plus négligé de la méthodologie de couplage des enregistrements.

Nous ne connaissons que deux stratégies qui permettent d'éviter ce dilemme.

1re stratégie: La simplicité

Cette stratégie, qui consiste essentiellement à éviter le tout dernier cri, a été utilisée avec grand succès dans beaucoup de cas. Elle convient particulièrement lorsqu'on a beaucoup de données à comparer ou encore lorsque les données sont de très grande qualité et peu aptes à changer. En d'autres circonstances, cette approche demeure valable lorsque le coût des décisions erronées (c.-à-d. les faux couplages ou les couplages manqués) est peu élevé.

Ordinairement, ce genre de situation nécessite le recours à des règles simples n'englobant qu'une zone et ayant pour résultat la concordance, la non-concordance ou l'absence. On peut utiliser seulement des pondérations globales ou faire appel à des pondérations spécifiques fondées sur la formule (2) ou (3) (ou une approche équivalente).

Dans beaucoup d'applications (surtout en épidémiologie), cependant, il est essentiel que pratiquement toutes les décisions de couplage soient justes. La quantité de données disponibles est souvent limitée, et celles-ci sont souvent de qualité douteuse. En pareil cas, les règles complexes de couplage sont inévitables et il faut recourir à la deuxième stratégie.

2e stratégie: Pondération directe de toutes les règles

La méthode directe est la seule qu'on puisse appliquer à toutes les règles de couplage. Par conséquent, pour les applications exigeant des règles plus complexes, il faut s'en servir pour toutes les règles afin de garantir l'intégrité du processus de pondération.

5. LA PONDÉRATION DIRECTE

La pondération directe est toutefois bien peu commune et comporte ses propres difficultés. En fait, nous ne connaissons qu'un organisme, la Division de la santé de Statistique Canada, qui aie l'expérience de cette approche (Newcombe, Fair, Lalonde (1987)). En outre, nous ne connaissons qu'un système informatique, le GIRLS de Statistique Canada (Hill, T. et Pringmill, F. (1985)) qui puisse servir à cette fin. Pour que cette approche devienne plus accessible, il faut respecter les conditions suivantes:

- il faut élaborer des outils informatiques généralisés pour faciliter le calcul des proportions et des facteurs de rectification nécessaires. Ce n'est pas compliqué à notre avis mais, même à Statistique Canada, ces outils n'existent pas à l'heure actuelle;
- comme l'estimation du numérateur des formules de pondération (c.-à-d. Prop. (résultat parmi les couplages)) doit souvent se faire de façon itérative, la réussite et le coût de la méthode dépendent de bonnes estimations initiales. Les praticiens expérimentés doivent rendre les résultats de leur travail accesssibles et les "nouvelles" règles doivent d'abord être testées dans des situations expérimentales pour que de bonnes estimations initiales soient disponibles;
- à court terme, les organismes qui effectuent des couplages complexes peuvent s'attendre à des résultats fiables seulement en adjoignant des experts en couplage à leurs équipes d'application;
- nous recommandons que les organismes tels que Statistique Canada, qui font du couplage d'enregistrements, créent un service de consultation en la matière afin d'en garantir l'application cohérente et juste dans tout l'organisme; et
- il faut rédiger des directives détaillées afin que les débutants puissent appliquer cette méthodologie correctement. Newcombe (1986) présente actuellement la

meilleure source d'information, même s'il met l'accent sur la stratégie plutôt que sur la mise en oeuvre; c'est inévitable étant donné que les outils nécessaires ne sont pas encore aisément disponibles.

6. RECHERCHE FUTURE

Il y a quatre grands domaines où l'on a besoin de recherches additionnelles. Les deux premiers sont prioritaires.

6.1 Rentabilité

Nous ne connaissons aucune analyse de la rentabilité des règles perfectionnées de couplage qui ont été adoptées ces dernières années. Il est évident que des règles plus perfectionnées, judicieusement pondérées, devraient améliorer la prise des décisions en matière de couplage mais il faut aussi se souvenir que:

- la plupart des décisions de couplage se font aisément au moyen de n'importe quel ensemble raisonnable de règles; et
- les perfectionnements augmentent sensiblement le coût.

Pour faire un choix raisonné entre les première et deuxième stratégies de la section 5, il importe d'avoir une idée du coût et des avantages de chacune.

6.2 Pondérations provenant d'autres applications

Comme on l'a noté à la section 5, la deuxième stratégie nécessite le recours à l'itération. L'itération a le plus de chances de donner lieu à la convergence si l'on a un bon point de départ; meilleur est le point de départ, et moins d'itérations il faut effectuer, ce qui réduit les frais et la durée du travail. Comme il faut souvent prendre son point de départ à l'aide de l'expérience du passé, les praticiens expérimentés rendraient un grand service aux autres en publiant leurs résultats. Nous avons constaté qu'il est relativement facile, mais pas particulièrement utile, d'obtenir des renseignements sur les pondérations nettes (voir la formule (7)) utilisées pour d'autres applications. Considérant qu'il faut un point de départ seulement pour la proportion du résultat parmi les couplages, c'est là le résultat qui serait le plus utile. Il serait avantageux également d'obtenir à partir de ces applications passées:

- la proportion du résultat parmi les non-couplages. D'une manière générale, il faut la calculer pour chaque nouvelle application mais il est utile de disposer de ces résultats d'applications passées en tant que reflet de la similitude des deux applications. S'il y a de grandes différences, peut-être faut-il recourir à une autre règle de couplage; et
- la gamme des facteurs de rectification. L'utilité principale de cette information est qu'elle précise l'importance de ces facteurs pour une règle donnée.

Bien sûr, si seulement la pondération nette ou quelque autre ventilation de la pondération est disponible, la valeur de l'expérience passée sera limitée.

6.3 Élaboration de nouvelles règles de couplage

Idéalement, toute nouvelle règle de couplage devrait être élaborée en milieu expérimental, dans une situation où les couplages et les non-couplages sont déjà connus parce qu'il est relativement facile de perfectionner la règle en pareille situation.

Cependant, la créativité naît de l'adversité et la plupart des nouvelles règles sont en fait élaborées en situation de production. La publication des résultats de la manière décrite ci-dessus importe pour que se poursuive la mise au point de la méthodologie du couplage des enregistrements.

6.4 Échantillonnage des paires non couplables

La manière usuelle d'échantillonner les paires non couplables, élaborée par Pierre Lalonde de Statistique Canada (Newcombe, Fair, Lalonde (1987)) comprend la sélection d'un échantillonnage aléatoire parmi toutes les comparaisons possibles. Étant donné que le nombre de comparaisons possibles est le produit du nombre d'enregistrements dans les deux fichiers et que le nombre possible de couplages équivaut au plus au nombre d'enregistrements dans le plus petit des deux fichiers, il s'ensuit que la probabilité d'inclure des couplages réels dans l'échantillon est très faible. Cette approche est tout à fait valable.

En réfléchissant, par contre, on se rend compte que la plupart des comparaisons possibles n'atteignent jamais l'étape de la pondération parce qu'il est absolument évident que ce sont des non-couplages. Ces comparaisons ne sont donc jamais effectuées (grâce aux stratégies de blocage) ou sont éliminées tôt dans le processus par les algorithmes initiaux de couplage. Il est donc possible que seulement les non-couplages qui répondent à quelque norme minimale de concordance pourraient servir au calcul des pondérations. Considérant que l'objectif de la pondération est de distinguer les couplages réels des non-couplages qui, d'une manière ou d'une autre, en ont l'apparence, peut-être la pondération globale appropriée est-elle:

$$\text{Pondération} = \frac{\text{Prop. (résultat parmi les couplages)}}{\text{Prop. (résultat parmi les non-couplages "proche")}}$$

Cette approche semble prometteuse, même si la recherche préliminaire effectuée par la Division de la santé de Statistique Canada indique le contraire.

7. CONCLUSIONS

Nos conclusions sont les suivantes:

- Le perfectionnement des règles de couplage, tout en étant évidemment bénéfique, a compliqué la pondération et nécessite le recours à une méthodologie intégrée qui garantit l'intégrité des décisions de couplage;
- la seule méthodologie possible est difficile à appliquer et n'est pas actuellement à la disposition de la plupart des praticiens. La seule solution consiste à élaborer de nouveaux outils généralisés;
- à court terme, il faut posséder une connaissance appréciable en la matière pour effectuer des couplages complexes. Il faut donc créer un service de consultation en matière de couplage des enregistrements; et
- à l'heure actuelle, en recherche, on accorde la priorité à l'application plutôt qu'à l'élaboration de nouvelles méthodologies car les applications actuelles n'approchent pas souvent des possibilités qu'offre la méthodologie existante. Il faut donc prendre soin de s'assurer que l'information découlant de l'expérience soit rendue disponible d'une manière utile.

BIBLIOGRAPHIE

Eagen, M. (1978). "Specification for the Calculation of Weights for GIRLS". Technical Report. Institutional and Agriculture Survey Methods Division, Statistics Canada.

Fellegi, I.P., and Sunter, A.B. (1969), "A Theory for Record Linkage". *Journal of the American Statistical Association*, pp. 1182-1210.

Hill, T. and Pringmill, F. (1985). "A Generalized Iterative Record Linkage System". Proceedings of the Workshop on Exact Matching Methodologies. Arlington, Virginia, May 9-10, 1985.

Newcombe, H.B. (1986). "Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business". Publication pending by Oxford University Press.

Newcombe, H.B., Fair, M.E., and Lalonde, P. (1987). "Concepts and Practices that Improve Probabilistic Linkage". International Symposium on STATISTICAL USES OF ADMINISTRATIVE DATA, sponsored by Statistics Canada, Ottawa, Canada, 1987 November 23-25.

SESSION III: COMMUNICATIONS SOLLICITÉES

APPROCHE INTÉGRÉES À L'ÉLABORATION

Président: B. Petrie, Statistics Canada

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

UTILISATION DES DONNÉES ADMINISTRATIVES DANS LE CONTEXTE DU PROJET DE REMANIEMENT DES ENQUÊTES-ENTREPRISES

MICHAEL COLLEDGE[1]

RÉSUMÉ

Le projet de remaniement des enquêtes-entreprises, en cours à Statistique Canada, est une tâche importante qui bouleversera le programme de la statistique économique. Le projet consiste essentiellement à remanier les notions, les procédures et les systèmes en vue de fournir des bases de sondage et d'utiliser les données fiscales. En outre, le projet permet de passer en revue la gamme complète des enquêtes économiques et d'élaborer des procédures et des systèmes génériques pour favoriser l'intégration des données et réduire les frais et le fardeau de réponse. L'exploitation des données administratives, pour constituer des bases de sondage ou pour remplacer les données faisant l'objet d'une collecte, est l'un des éléments essentiels de l'approche. Le présent document résume la stratégie et la mise en oeuvre du projet en particulier en ce qui a trait aux données administratives.

1. INTRODUCTION

Le présent document est divisé en cinq parties. La première section renferme un aperçu du projet de remaniement des enquêtes-entreprises, notamment les objectifs, des moyens de les atteindre et les sources administratives pertinentes. Les trois sections suivantes portent essentiellement sur les principales applications des données administratives dans le contexte du projet. La deuxième section traite de l'utilisation de ces données en vue de définir et de maintenir la base de sondage des entreprises. La troisième section décrit le remplacement partiel des méthodes de collecte annuelles des données d'enquêtes par des procédures fondées sur les déclarations fiscales. La quatrième section traite des applications des données administratives dans le contexte du remaniement d'autres tâches relatives aux enquêtes, et souligne les efforts déployés en l'occurrence pour élaborer des méthodes et des systèmes généralisés.

On donne des exemples des restrictions inhérentes aux données administratives, découlant du fait qu'elles sont axées sur des besoins administratifs plutôt que statistiques, et on décrit les méthodes visant à contrer ces restrictions. On ne tient pas compte dans le présent document des questions de principe touchant, par exemple, l'échange, la protection et la confidentialité des données. Pour plus de détails à ce sujet voir, notamment, Brackstone (1987). Il est fait mention des problèmes de qualité qui ne font toutefois pas l'objet de discussions approfondies (voir Lussier et Colledge, 1987). Viennent enfin un résumé et certaines observations générales au sujet du rôle important des données administratives dans les enquêtes-entreprises.

[1] Michael Colledge, Assistant Directeur, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11ième étage, Edifice R.H. Coats, Ottawa, Ontario. K1A 0T6

1.1 Projet de remaniement des enquêtes-entreprises

Le programme de la statistique économique de Statistique Canada comprend environ 300 enquêtes (le nombre variant selon la définition précise du terme enquête), mensuelles, trimestrielles, annuelles et occasionnelles, sur les statistiques financières, industrielles, fiscales, relatives aux machandises, à l'emploi, et aux dépenses en capital. De ce nombre, approximativement 125 sont des enquêtes infra-annuelles et le reste, des enquêtes annuelles ou occasionnelles.

En 1985, Statistique Canada a mis sur pied le programme de remaniement des enquêtes-entreprises (PREE), dont le principal objectif est d'uniformiser et d'intégrer tous les systèmes et les données de sortie du programme économique. Pour atteindre cet objectif, il faut absolument se servir d'une même base centrale et adopter une méthodologie généralisée d'enquête, ce qui non seulement favorise l'intégration des statistiques et la rationalisation des opérations, mais réduit le coût de la conception, du remaniement et de l'amélioration des enquêtes, au fur et à mesure que les méthodologies spécifiques et les systèmes sur mesure sont remplacés par des méthodologies et des systèmes de nature générale.

Il faut signaler sans plus tarder que le terme "entreprise", qui apparaît dans le titre du projet, est utilisé à Statistique Canada pour désigner les unités de production économique appartenant non seulement aux secteurs du commerce et des services marchands mais également à ceux de la fabrication, de la construction, des transports, des activités professionnelles, etc.

Grâce au PREE on s'efforce de garantir la qualité du programme économique. Parmi les avantages qu'on espère en tirer, on compte:

(a) la normalisation des notions, définitions, plans de classification et procédures d'enquête;

(b) l'élargissement de la couverture de la base, la définition plus précise des unités économiques et des ententes relatives aux déclarations et la fiabilité accrue de la classification industrielle

(c) l'utilisation accrue des données administratives, la réduction du fardeau de réponse et l'amélioration des rapports avec les répondants;

(d) la réduction de l'ensemble des frais se rapportant à l'entretien de la base, aux envois par la poste et à la collecte des données;

(e) l'amélioration des installations relatives à l'intégration des données et l'élargissement de la portée de la vérification;

(f) l'élaboration de systèmes généralisés pour une gamme de tâches se rapportant aux enquêtes, notamment le codage automatisé et informatisé des branches d'activité, la détermination de la taille de l'échantillon, la répartition et la sélection, la correction, l'imputation, etc.

La statégie visant à réaliser ces objectifs comporte deux volets importants. Le premier est celui de l'uniformisation, c'est-à-dire l'introduction de notions uniformisées, de méthodes génériques et de systèmes généralisés et, en particulier, d'une nouvelle base de données centrale. Le second volet est axé sur l'utilisation des données administratives, à la fois pour définir et maintenir la base de sondage et pour remplacer la collecte directe des données. Les données administratives sont donc l'une des pierres angulaires du PREE. Tous les détails au sujet des objectifs et de la stratégie sont fournis par Cain et al (1984), et ont été présentés sous forme abrégée par Colledge et Lussier (1985, 1987) et Colledge (1987).

1.2 Sources de données administratives

Il existe de nombreuses sources de données administratives, qu'elles soient gouvernementales (administrations fédérale, provinciales et municipales) ou commerciales. En voici quelques exemples: les retenues sur la paie, l'impôt sur le revenu des sociétés, l'impôt sur le revenu personnel, l'assurance-chômage, les importations, les exportations, les banques à charte, les sociétés de fiducie enregistrées, les compagnies aériennes, les navires, les commissions de transport, les services publics (électricité, eau, téléphone), les associations professionnelles, les offices de producteurs, etc. Au tout début du projet, toutes les sources éventuelles ont été répertoriées et évaluées en termes de leur utilité dans le contexte des enquêtes-entreprises (Bankier et al, 1985). On a alors décidé d'effectuer le remaniement en se fondant essentiellement sur trois sources principales, provenant de Revenu Canada, Impôt, c'est-à-dire les retenues sur la paie, l'impôt sur le revenu des sociétés et l'impôt sur le revenu personnel. (Si on décidait d'imposer une taxe sur les opérations commerciales ou à la valeur ajoutée, il s'agirait là d'une quatrième source très importante. Pour l'instant, les plans relatifs à l'imposition de cette taxe sont tellement vagues qu'il n'est pas question d'en tenir compte dans la stratégie.) Dans les sections suivantes, on discutera en détail du rôle de ces sources.

2. DONNÉES ADMINISTRATIVES VISANT À DÉFINIR ET À MAINTENIR LA BASE

2.1 Système actuel et systèmes proposés

On trouvera à la figure 1A une illustration du système actuel visant à définir et à maintenir les bases des enquêtes-entreprises. Il s'agit d'un système lourd et quelque peu morcelé dont l'élément principal est le registre des entreprises qui fournit des renseignements de base à partir desquels on élabore des bases distinctes pour répondre aux besoins de chaque enquête. Le système de retenues sur la paie de Revenu Canada est la principale source de données permettant de mettre à jour le registre des entreprises. On maintient deux bases de données indépendantes ayant trait aux déclarations sur le revenu des sociétés et aux déclarations sur le revenu des particuliers, fournies également par Revenu Canada. Certaines enquêtes sont fondées sur des bases élaborées à partir de ces sources et non pas du registre des entreprises.

On peut résumer comme suit les problèmes caractérisant le système actuel. Premièrement, il y a le manque d'efficacité compte tenu du sous-dénombrement et du dédoublement ainsi que de la difficulté d'intégrer les données découlant des enquêtes. Deuxièmement, il y a le manque d'efficience découlant du dédoublement au niveau du maintien des bases d'enquête individuelles.

En principe, le système figure 1A sera remplacé par celui de la figure 1B. Les principaux éléments stratégiques qui permettront d'en venir à cette solution sont les suivants:

(a) l'élaboration de notions uniformes et d'un modèle d'information;

(b) la mise en oeuvre d'un nouveau système de stockage des données de la base centrale fondé sur les notions uniformes et le modèle d'information;

(c) l'utilisation des données administratives et l'introduction d'un programme "d'établissement de profils" pour créer et maintenir les données de la base.

Dans la sous-section suivante, ces éléments sont envisagés sous l'angle d'une séquence de problèmes et des solutions correspondantes.

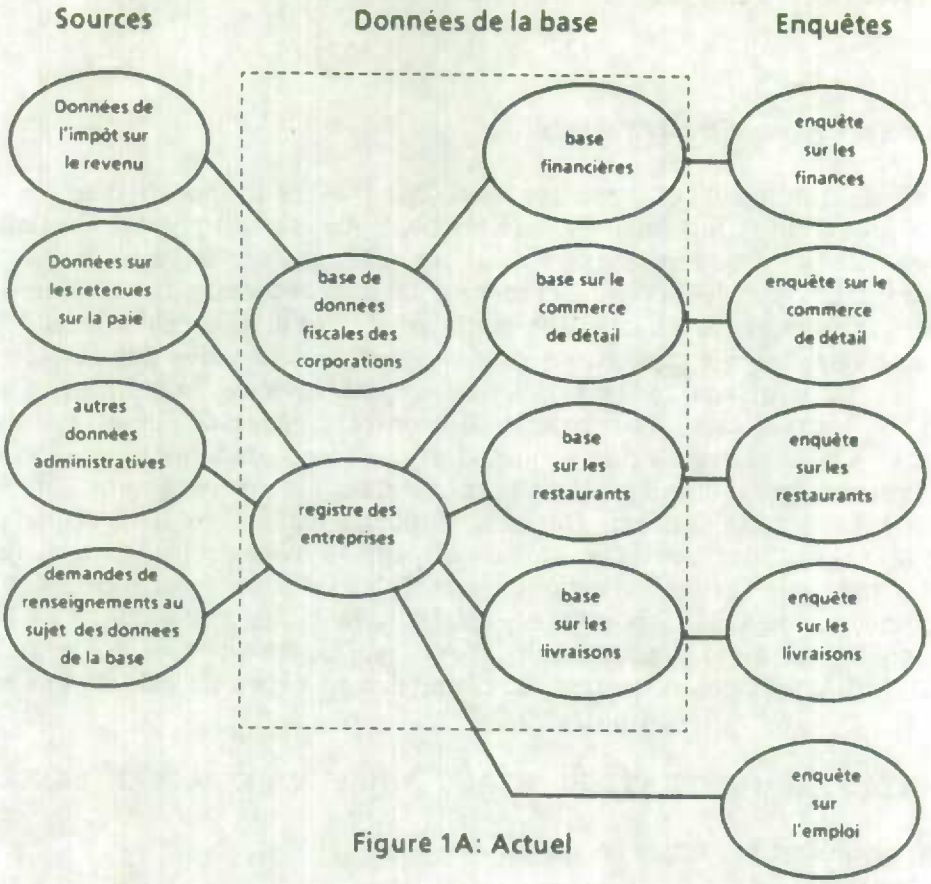


Figure 1A: Actuel

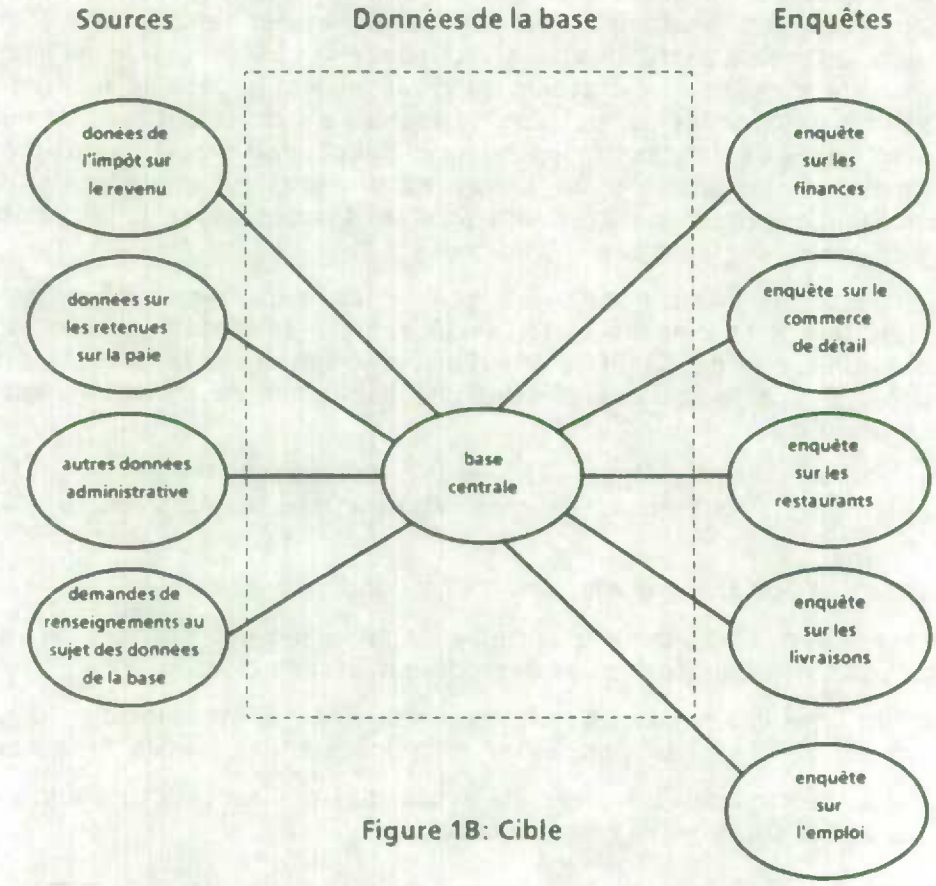


Figure 1B: Cible

Figure 1: Systèmes visant à fournir les données de la base (simplifiés)

2.2 Création de la base : problèmes et solutions

Problème. On détermine l'unité cible appropriée en fonction des exigences des données d'enquête. En d'autres termes, il n'existe pas d'unité pour laquelle on peut établir une population cible pour toutes les enquêtes-entreprises. Prenons l'exemple d'une entreprise possédant trois points de vente au détail dans deux provinces, deux filiales de vente qui s'occupent des points de vente, une filiale de vente en gros à laquelle sont rattachés deux points de vente au détail, et un bureau central. Pour l'ensemble de l'entreprise, il est possible qu'on ne puisse obtenir de données relatives aux ventes que pour les points de vente individuels, de comptes de profits et pertes qu'au seul niveau des filiales et un bilan consolidé qu'en ce qui a trait au bureau central.

Solution. Il s'agit de déterminer explicitement la nécessité d'élaborer divers types d'unités statistiques. On a élaboré une hiérarchie à quatre niveaux d'unités statistiques cibles, définissant ainsi une unité pertinente pour chaque type d'enquête. Cette classification est illustrée au tableau 2 où l'on voit une "entreprise-statistique" pouvant déclarer des données consolidées. L'entreprise englobe deux "compagnies statistiques", qui fournissent des états financiers non consolidés, trois "établissements statistiques" qui fournissent des données permettant d'établir la valeur ajoutée, et cinq "emplacements statistiques" qui fournissent des données relatives aux ventes.

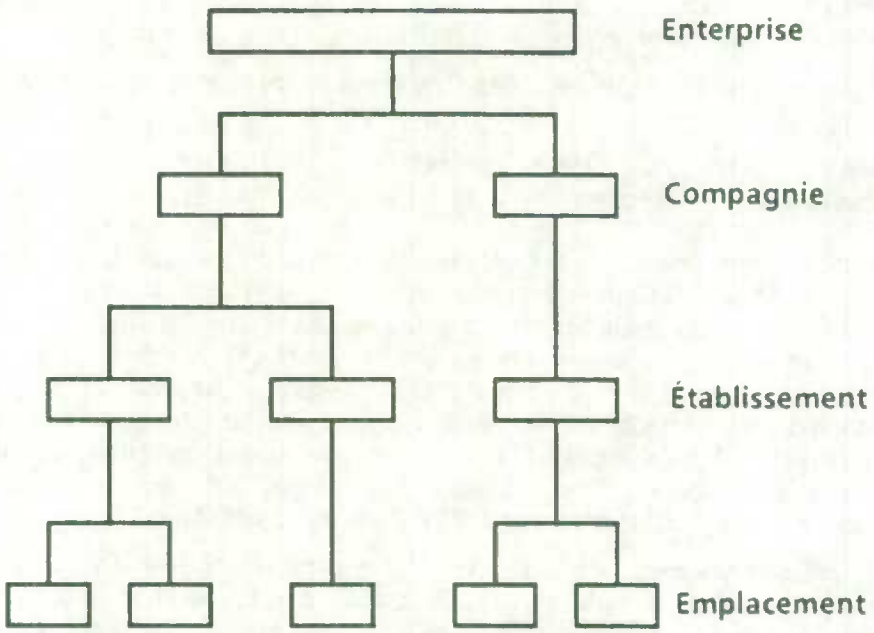


Figure 2: Représentation statistique d'une entreprise sour forme d'une Hiérarchie à quatre niveaux

Problème. A priori, il n'existe pas de liste d'unités statistiques pouvant servir dans le contexte des enquêtes-entreprises. Statistique Canada doit dresser et tenir ces listes.

Solution. L'approche adoptée consiste à dresser des listes des unités statistiques fondées sur des données administratives. Il ne s'agit pas là d'une nouvelle approche. En fait, à part une exception, chaque base d'enquêtes-entreprises est établie à partir d'une liste administrative, ce qui prouve donc que les enquêtes-entreprises dépendent essentiellement des sources administratives. La seule exception est la base aréolaire élaborée par Statistique Canada qui sert, à l'heure actuelle, à compléter la couverture de la base de l'enquête mensuelle sur le commerce de détail.

Même ce dernier vestige du registre des entreprises qui ne provient pas d'une source administrative disparaîtra, au moins à titre expérimental, au moment de l'introduction de la nouvelle enquête remaniée sur le commerce de détail.

Problème. On ne peut s'attendre à ce que les unités administratives et statistisques coïncident, étant donné qu'elles répondent à des besoins différents. Plus précisément, une unité statistique cible est divisée en quatre niveaux, en fonction des données d'enquête recueillies, dont l'un tout au plus pourrait être apparié aux unités fournies par une source administrative déterminée. Donc, en général, les listes administratives ne peuvent servir de base d'enquête.

Solution. La première étape consiste à définir un modèle d'information (Statistique Canada 1985) qui détermine explicitement les divers types d'unités du monde des entreprises et qui établit un lien entre ces unités et les cibles statistiques. Ce modèle comprend cinq types distincts d'unités:

(a) juridique - par exemple, les sociétés incorporées sous une charte fédérale ou provinciale;

(b) administrative - par exemple, les titulaires de comptes de retenues sur la paie, les déclarants fiscaux;

(c) opérationnelle - par exemple, les divisions, les centres de profit, etc., correspondant à la façon dont l'entreprise se structure et tient ses comptes d'exploitation; ensemble, les unités juridiques, administratives et opérationnelles définissent l'image que l'entreprise a d'elle-même;

(d) statistique - les unités cibles à des fins de mesure statistique, c'est-à-dire l'image que l'organisme statistique à de l'entreprise;

(e) déclarante - l'unité qui permet de relier les unités cibles statistiques et les unités d'exploitation de l'entreprise.

La seconde étape consiste à définir les procédures qui permettront d'obtenir des ensembles d'unités statistiques à partir de listes administratives. Ces procédures, qu'on qualifie globalement de procédures "d'établissement de profils", sont les suivantes: la délimitation des entreprises est fondée sur les données administratives, par exemple les listes des sociétés déclarantes, et sur les renseignements relatifs à l'appartenance et au contrôle fournis par le programme de la Loi sur les déclarations des corporations et des syndicats ouvriers. Grâce à une interview ou à un questionnaire, on peut alors définir les structures d'exploitation de l'entreprise. Finalement, les structures statistiques découlent automatiquement des renseignements relatifs à l'exploitation.

La troisième étape consiste à mettre en oeuvre le modèle d'information sous la forme d'une base de données, appelée base de données du registre central (BDRC), ayant la capacité de stocker et de manipuler tous les renseignements découlant de l'établissement de profils.

Problème. L'établissement de profils coûte cher. Il serait beaucoup trop coûteux "d'établir un profil" pour chaque entreprise au pays.

Solution. Deux caractéristiques importantes des entreprises permettent d'en arriver à une solution. Premièrement, l'économie est dominée par un nombre relativement peu élevé de grandes entreprises; deuxièmement, les unités administratives et statistiques coïncident souvent (aux quatre niveaux) dans le cas des petites entreprises. Par conséquent, l'approche adoptée consiste à établir le profil des entreprises importantes afin de déterminer les unités statistiques correspondantes et, dans le cas des petites entreprises, à se servir des listes administratives comme bases. La base de données du registre central est donc divisée en parties distinctes, comme on le voit au tableau 3A. La "partie intégrée" (PI) constitue une liste unique, pour laquelle il n'existe pas de double, des unités

statistiques regroupant toutes les entreprises importantes, de même que les renseignements "structurels" qui s'y rapportent, c'est-à-dire les données administratives, juridiques, opérationnelles et de déclaration. Elle découle de l'établissement de profils et nécessite le raccordement intégral et le non-dédoublement de toutes les sources administratives et autres entrées. Les unités statistiques, découlant de la structure d'exploitation, sont classées selon la branche d'activité, le secteur géographique et la taille. La "partie non intégrée" (PNI) renferme les petites unités qui viennent compléter la base. Elle est constituée directement à partir des données administratives en supposant que chacune des unités statistiques correspond à une unité administrative. Les très petites unités sont définies comme étant hors champ aux fins de l'enquête.

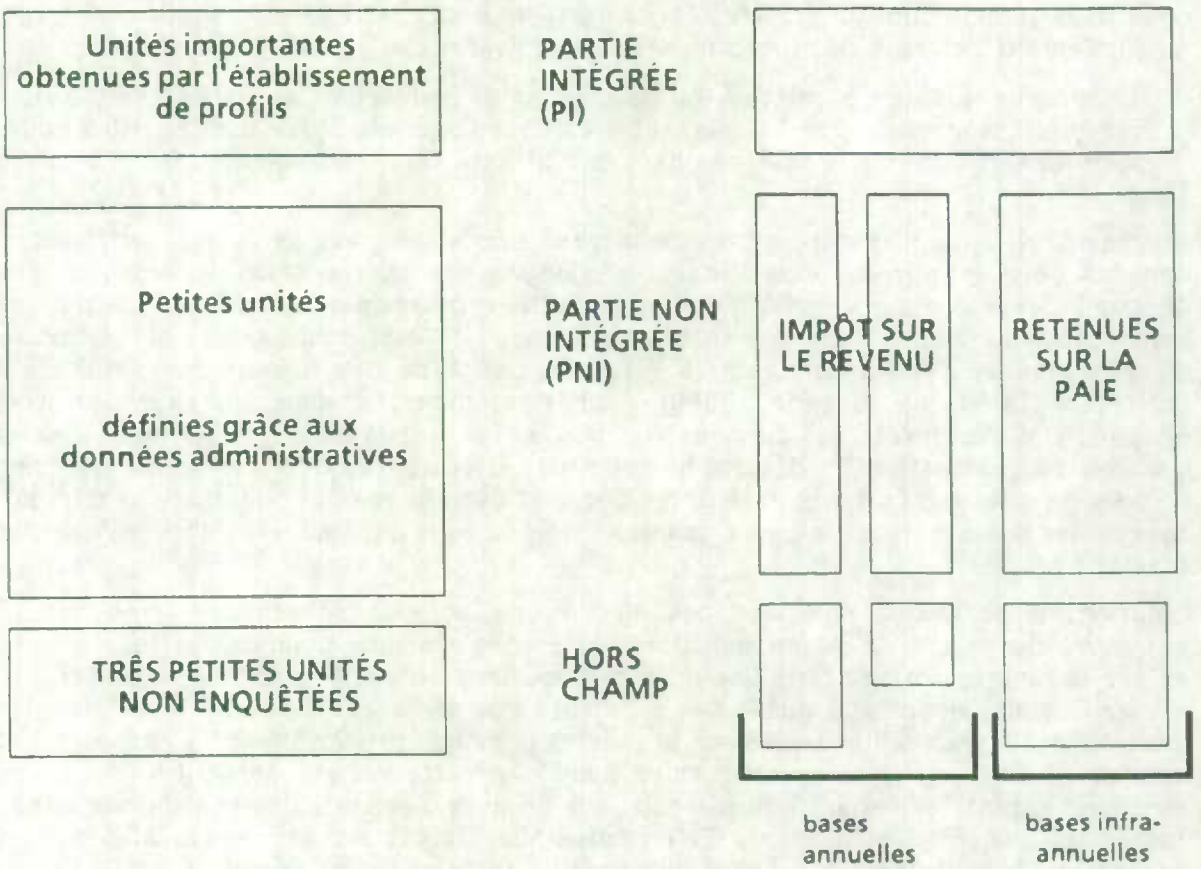


Figure 3A: concept de base

Figure 3B: bases alternatives

Figure 3: Base de données du registre central

Problème. Il n'existe pas de source administrative unique qui peut produire un ensemble complet de petites unitées pour la PNI. Les données fiscales servent à constituer des bases pour des enquêtes annuelles relatives à la production économique et aux finances mais les délais sont trop importants pour qu'on puisse s'en servir dans le contexte des enquêtes infra-annuelles. Les retenues sur la paie fournissent des données plus actuelles, qui conviennent aux enquêtes infra-annuelles, mais elles ne tiennent pas compte des non-employeurs.

Solution. Il s'agit d'utiliser ces deux sources pour établir la PNI.

Problème. De nombreuses entreprises étant inscrites à la fois sur les listes de retenues sur la paie et d'impôt sur le revenu, il est donc possible que de nombreux dédoublements se produisent si les deux listes sont utilisées.

Solution. On pourrait se servir de techniques se rapportant à une base multiple. Cela rendrait les estimations beaucoup plus complexes ce qui n'est pas souhaitable compte tenu du grand nombre d'enquêtes dont il est question.

Une deuxième approche consisterait à apparier les deux listes, en éliminant le double emploi. Toutefois, bien que ces deux listes proviennent de Revenu Canada, elles ne sont pas identifiées de la même manière, elles ne sont pas raccordées à l'heure actuelle et il est impossible de procéder aisément à leur appariement et à l'élimination des doubles comptes. Si on fonde le raccordement informatisé sur les noms et les adresses, on court le risque de ne pas effectuer de très nombreux raccordements car il est possible que les renseignements relatifs au nom et à l'adresse fournis par le système de retenues sur la paie diffèrent souvent beaucoup de ceux qui figurent sur les déclarations d'impôt, et ce pour une même unité. Cette façon de procéder obligerait également à traiter manuellement un grand nombre d'appariements éventuels, ce qui coûte cher.

L'approche adoptée consiste à ne pas tenter de raccorder les listes, c'est-à-dire à les laisser "non intégrées" (d'ou le nom), et à utiliser l'une **ou** l'autre d'entre elles pour servir de complément à la partie intégrée dans le contexte de l'établissement de la base. Voir la figure 3B.

Problème. Ni l'une ni l'autre des deux sources administratives ne fournit suffisamment de données pour établir les classifications précises des unités selon la branche d'activité économique, le secteur géographique et la taille, requises à des fins de stratification et de sélection judicieuse d'un échantillon. Ainsi, il est souhaitable de stratifier et d'échantillonner des unités selon la classification type des industries de 1980 à quatre chiffres (Statistique Canada, 1980) pour des enquêtes annuelles sur la production économique. Toutefois, les données figurant sur la déclaration d'impôt sur le revenu des sociétés ne permettent de déterminer une classification à quatre chiffres que pour 74 % de tous les déclarants. Le chiffre correspondant dans le cas des déclarations d'impôt sur le revenu des déclarants qui ne sont pas des sociétés mais qui font état d'un revenu d'affaires atteint 50 %.

Solution. Auparavant, dans les cas où l'on ne pouvait déterminer avec précision la catégorie de la classification industrielle des déclarations d'impôt, on leur attribuait la valeur la plus vraisemblable. Les erreurs découlant de cette façon de procéder ont laissé planer le doute quant à la qualité et à l'utilité des déclarations d'impôt. En ce qui a trait aux données tirées des retenues sur la paie et des déclarations d'impôt, l'approche adoptée consiste donc toujours à communiquer avec une unité en cas de doute. En outre, pour réduire d'autant les frais, on aura recours à un plan d'échantillonnage à deux phases pour toutes les enquêtes utilisant la PNI fondée sur l'impôt sur le revenu afin qu'il ne soit nécessaire de maintenir des codes industriels précis que pour l'échantillon de la première étape (Colledge, Estevao et Foy, 1987).

Problème. Lorsqu'on sélectionne un échantillon à partir de la PNI, il se peut qu'il contienne certaines unités administratives qui ne correspondent pas de façon biunivoque avec les unités statistiques pertinentes, c'est-à-dire celles qui auraient été définies si l'on avait procédé à l'établissement du profil de l'unité. Par exemple, une unité administrative unique peut correspondre à trois entreprises -prenons le cas d'une déclaration d'impôt remplie par un déclarant qui n'est pas une société, c'est-à-dire qu'il s'agit d'un médecin qui a un revenu professionel, qui est également associé dans l'exploitation d'un club d'athlétisme et qui possède une ferme d'agrément. Dans ce cas, il y aurait idéalement trois unités statistiques au lieu d'une seule. La situation inverse peut également se produire. Une unité fiscale faisant partie de l'échantillon peut représenter un déclarant associé à quatre autres personnes, c'est-à-dire qu'il s'agit idéalement d'une unité statistique mais, en fait, de cinq unités fiscales. Dans le cas des échantillons tirés de la PNI fondée sur les retenues sur la paie, il se peut également que les unités administratives ne correspondent pas aux unités statistiques pertinentes.

Solution. L'approche adoptée dans le cas des unités fiscales sélectionnées pour faire partie de l'échantillon et faisant état d'entreprises multiples consiste à définir une unité statistique distincte pour chaque entreprise individuelle, chacune d'entre elle fournissant ses propres états financiers annexés aux déclarations d'impôt. La pondération de l'échantillonage de l'unité administrative est reportée dans le cas de chaque unité statistique.

Le cas des déclarants associés entraîne le dédoublement de la base fiscale. Il s'agit alors de multiplier la pondération de chaque unité sélectionnée pour faire partie de l'échantillon par la part que détient l'unité dans la société. On procède de la même façon en ce qui a trait à l'utilisation des comptes de retenues sur la paie se rapportant à une entreprise unique.

Problème. La base de données du registre central doit être mise à jour afin de tenir compte, dans la mesure du possible, de tous les changements pertinents au niveau de la structure et de la classification dans le monde des affaires. Il s'agit là d'un processus complexe. Diverses sources peuvent fournir les renseignements permettant de procéder à cette mise à jour. Les données émanant de l'une ou l'autre de ces sources peuvent être imcomplètes et elles peuvent même contredire en partie les renseignements d'autres sources. Il existe un nombre effarant d'autres façons de mettre à jour les ensembles d'unités juridiques, administratives, opérationnelles et statistiques, ainsi que les rapports qu'elles entretiennent entre elles, et il faut éviter certaines embûches. Par exemple, les modifications de la structure légale, notamment les absorbtions, les fusions, les prises de contrôle, les créations de filiales, etc. n'entraînent pas nécessairement des changements au niveau des structures correspondantes d'exploitation ou statistiques. Par conséquent, si l'on procédait à la mise à jour automatique des ensembles d'unités statistiques en se fondant sur les renseignements transmis par des sources administratives ou juridiques, on pourrait se retrouver avec un nombre illusoirement élevé de "créations" et de "disparitions" apparentes d'unités statistiques et courir le risque correspondant d'avoir un champ d'observation incomplet ou dont certaines parties se recoupent.

Solution. Pour faire face à cette situation, on a défini un ensemble d'environ 50 "événements types" pouvant se produire dans le monde des affaires (Armstrong et al, 1986). Toute indication d'un changement est considérée comme un "indice" provoquant l'examen des unités correspondantes, et, au besoin, l'établissement de nouveaux profils, pour déterminer si l'un ou l'autre des 50 événements types s'est produit. On procède toujours à la mise à jour de la base de données du registre central en fonction de ces événements. Les définitions précises des créations, disparitions et modifications relatives aux entités statistiques sont incorporées dans les critères de détermination des événements types et de production d'entités statistiques. Il existe un grand nombre d'indices émanant essentiellement de trois sortes de sources : résultats d'enquêtes, processus administratifs et établissement de nouveaux profils dans le contexte des opérations de routine de la base de données du registre central. Ces signaux sont placés sur le "bloc de montage" de la base de données du registre central, où ils sont triés en fonction de la complexité, groupés en une unité de travail et automatiquement attribués aux membres du personnel chargés du maintien de la base de données du registre central lorsqu'ils demandent du travail. (Voir figure 4).

Conformément à la ligne de pensée générale voulant que l'on traite différemment les entreprises, selon qu'elles soient petites ou grosses, le maintien des unités de la PNI consiste en la mise à jour informatisée fondée sur les enregistrements administratifs, à laquelle s'ajoutent les résutats d'enquête et, au besoin, les contacts directs.

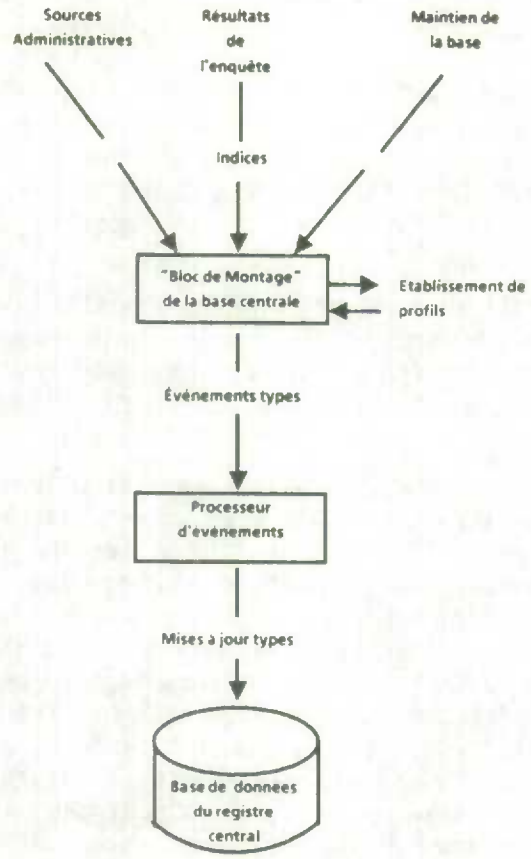


Figure 4: Mise à jour de la base de données du registre central

3. UTILISATION DES DONNÉES ADMINISTRATIVES POUR COMPLÉTER LES DONNÉES D'ENQUÊTE

L'un des objectifs principaux du projet de remaniement des enquêtes-entreprises est de réduire les frais d'opération et le fardeau des répondants grâce à l'utilisation des données administratives. Dans ce contexte, l'application la plus importante est le remplacement partiel des données d'enquêtes annuelles par des données fiscales. Par conséquent, les éléments essentiels de la stratégie sont les suivants:

(a) dans le cas des petites entreprises, l'utilisation des données financières tirées des déclarations d'impôt au lieu des données recueillies dans le contexte d'une enquête;

(b) La coordination des procédures d'échantillonnage, d'acquisition et de traitement des données fiscales pour répondre aux besoins collectifs de toutes les enquêtes.

La figure 5 illustre les procédures générales grâce auxquelles on obtient de Revenu Canada les données fiscales. Cet organisme saisit toutes les données fiscales et certaines données financières de chaque déclaration d'impôt. Ces données sont présentées sous une forme ordinolingue. Toutefois, elles ne renferment pas tous les renseignements requis aux fins des enquêtes-entreprises. Statistique Canada doit donc demander qu'on lui fournisse un échantillon des déclarations d'impôt, identifié par les numéros de l'unité individuelle ou par l'algorithme d'échantillonnage fondé sur les catégories d'unités. Les déclarations d'impôt ainsi spécifiées sont identifiées durant le traitement à Revenu Canada, photocopiées puis, envoyées à Statistique Canada. Les informations additionnelles sont alors extraites des annexes et des états financiers joints aux déclarations. On trouvera dans les sous-sections suivantes une description de l'utilisation de ces données dans le contexte des enquêtes annuelles sur la production économique et de l'enquête annuelle sur les statistiques financières et fiscales.

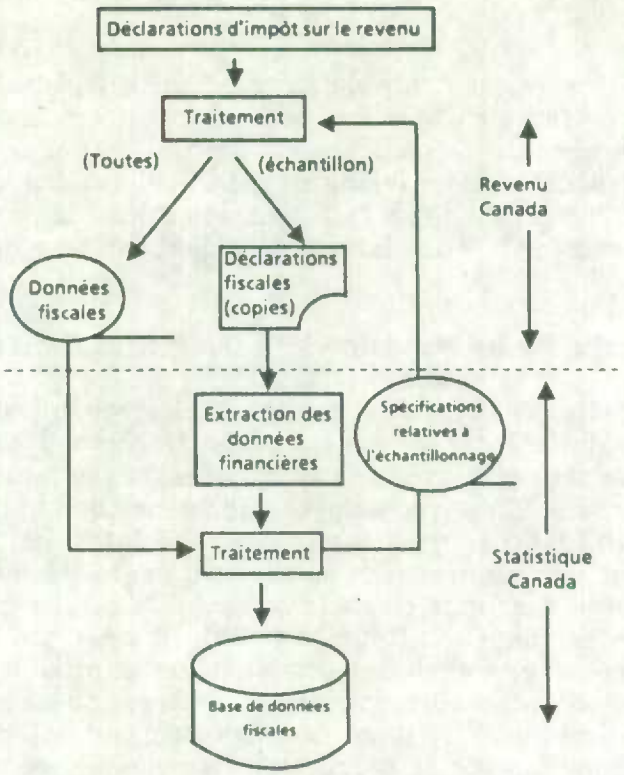


Figure 5: Acquisition des données fiscales

3.1 Enquête annuelle sur la production économique

À l'heure actuelle on recueille les données sur la production annuelle grâce à un certain nombre d'enquêtes indépendantes portant sur des groupes d'activités distinctes, notamment le recensement annuel des manufactures, le recensement de la construction, l'enquête sur le commerce de gros, l'enquête sur le commerce de détail, le programme des enquêtes annuelles, etc. L'objectif stratégique du projet de remaniement des enquêtes-entreprises est d'uniformiser les notions et de coordonner les procédures se rapportant à toutes ces enquêtes afin de pouvoir les regrouper en une seule "enquête annuelle sur la production économique". La description de certaines notions types figure aux paragraphes suivants.

L'établissement constitue l'unité statistique cible. On recueille au moins un nombre de données suffisant pour calculer la "valeur ajoutée recensée" (Statistique Canada, 1980) au niveau géographique et industriel le plus petit qu'on puisse atteindre sur une base annuelle.

La base est tirée de la PI et de la PNI fondée sur le revenu. Toutes les unités dont l'année financière se termine pendant la période allant du 1er avril, A au 31 mars, A+1 sont définies comme faisant partie de l'année de référence A.

Les données sont obtenues des unités de la PI grâce au contact direct au moyen d'un questionnaire personnalisé. Pour réduire le fardeau du répondant et les erreurs de déclaration, on demande aux répondants de fournir des données correspondant à leur propre année financière (plutôt qu'à l'année civile). Dans les cas où les unités statistiques et fiscales coïncident, on peut utiliser les déclarations d'impôt comme sources de données financières plutôt que de procéder à la collecte de données.

Dans le cas des unités de la PNI, toutes les données financières proviennent des déclarations d'impôt sur le revenu. Pour compléter ces données, on procède à des enquêtes portant sur d'autres caractéristiques, notamment l'emploi, les marchandises produites ou consommées, les types de services fournis, etc. Étant donné qu'on ne peut pas vraiment classer les unités fiscales de la base dans les catégories à quatre chiffres de la CTI de 1980, on se sert d'un plan de sondage à deux phases, comme il a été indiqué précédemment.

D'une année à l'autre, on tente de faire en sorte que les échantillons se recoupent le plus possible afin de fournir une bonne base pour les comparaisons entre les années. Dans ce contexte, le fardeau du répondant n'est pas un facteur important compte tenu du fait que la plupart des données proviennent des déclarations d'impôt. La surveillance du recoupement de l'échantillon se fait par sélection au moyen d'un nombre aléatoire reproductible obtenu par "manipulation" de l'identificateur de l'unité fiscale (voir Sunter, 1986).

3.2 Enquête annuelle sur les statistiques financières et fiscales

À l'heure actuelle, l'enquête est fondée exclusivement sur les données fiscales. Les données financières de l'univers des sociétés transmises par Revenu Canada sous une forme ordinolingue sont fusionnées à des données plus détaillées tirées d'un échantillon de déclarations d'impôt. Les problèmes découlant de cette approche sont triples. Premièrement, les données relatives à des sociétés qui appartiennent à un même propriétaire ou qui sont centralisées ne sont pas regroupées et il est donc difficile de les rapprocher des renseignements financiers recueillis par l'enquête financière trimestrielle auprès des unités regroupées. Deuxièmement, il n'est pas facile d'apparier, au micro-niveau, ces données avec celles des enquêtes sur la production économique étant donné que les ensembles d'unités correspondants n'entretiennent pas de rapports bien définis entre eux. Troisièmement, l'échantillon des déclarations d'impôt ne correspond pas à l'échantillon des enquêtes sur la production économique, ce qui entraîne le dédoublement des efforts et la perte de renseignements.

Les éléments de la stratégie sont les suivants:

(a) définir l'unité cible comme étant l'entreprise statistique définie pour les besoins de la base de données du registre central et non l'unité d'impôt sur le revenu des sociétés, et établir la base à partir de la base de données du registre central plutôt que du fichier des sociétés de Revenu Canada;

(b) dans le cas des unités importantes, obtenir les données financières requises en se basant sur quatre trimestres de l'enquête trimestrielle, permettant ainsi de laisser de côter l'utilisation des données administratives, afin d'obtenir des renseignements regroupés;

(c) se servire d'un échantillon semblable à celui de l'enquête annuelle sur la production économique, comportant notamment deux étapes pour les unités de la PNI.

4. AUTRES UTILISATIONS DES DONNÉES ADMINISTRATIVES

Cette section renferme la description d'autres utilisations des données administratives dans le contexte du remaniement des enquêtes-entreprises.

Les lacunes actuelles des systèmes et des procédures des enquêtes-entreprises de Statistique Canada s'expliquent surtout par le fait qu'il existe de nombreuses enquêtes qui ont été élaborées indépendamment l'une de l'autre. Il est difficile d'intégrer les résultats de ces enquêtes et il en coûte cher pour maintenir ou remanier les systèmes faits sur mesure. Les éléments de la stratégie du PREE axée sur la résolution de ces problèmes sont, en résumé, les suivants.

Premièrement, il faut réviser et rationnaliser les objectifs de l'ensemble du programme des enquêtes-entreprises. Règle générale, les enquêtes infra-annuelles doivent essentiellement fournir des estimations du changement et des données actuelles plutôt que des données détaillées qui sont fournies par les enquêtes annuelles. Deuxièmement, l'uniformisation des notions et l'élaboration de méthodes génériques et de

sytèmes généralisés permettra de réduire le fardeau du répondant, de favoriser l'intégration des données et de réduire les frais relatifs au remaniement et au maintien. Plus précisément, le programme actuel d'enquête sera réorienté pour tenir compte des nouvelles procédures de prestation des données de la base, d'acquisition et d'utilisation des données fiscales mentionnées dans les sections précédentes.

Dans ce contexte, les données administratives ont une importance considérable mais, du point de vue du responsable d'une enquête, leur rôle est essentiellement transparent. Par exemple, bien que les sources administratives soient un élément clé de l'établissement et du maintien des données de la base, chaque responsable d'enquête reçoit simplement une base de données du registre central et il n'a pas la chance de consulter ou de remettre en question les sources (administratives) de cette base. Parallèlement, le service de traitement central des données fiscales transmet au responsable de l'enquête les données fiscales, au niveau de l'unité et de l'agrégat, de la même façon que le service central de saisie des données transmet les données des questionnaires d'enquête. Encore une fois, les responsables n'entretiennent aucun rapport direct avec la source administrative.

4.1 Fonction générique: Utilisation des données administratives

La figure 6 présente un résumé des fonctions génériques des bases de données associées à une enquête-entreprise. La ventilation du processus global d'enquête en dix fonctions génériques est quelque peu arbitraire mais elle est pratique à des fins de description. Voir Colledge et Lussier (1987) pour plus de détails. Dans les paragraphes qui suivent, on explique le rôle des données administratives en abordant les fonctions l'une après l'autre.

On a discuté longuement (à la section 2) de l'utilisation des données administratives pour créer et maintenir la base. Le fait que les données proviennent d'un questionnaire d'enquête ou d'une déclaration d'impôt ne change rien à la production des unités déclarantes et des documents servant à établir des contacts (questionnaires, listes de contrôle, etc.), sauf qu'on envoie à Revenu Canada les listes des unités déclarantes, plutôt que les questionnaires, pour que le ministère intercepte les déclarations d'impôt correspondantes. En ce qui a trait à la stratification et à la répartition de l'échantillon, on peut également se servir des données administratives pour remplacer les données d'enquête pertinentes si celles-ci ne sont pas disponibles, par exemple dans les cas où il s'agit de concevoir une enquête entièrement nouvelle. La répartition et la sélection de l'échantillon n'ont essentiellement rien à voir avec le fait que les données sont tirées d'une enquête ou des déclarations d'impôt sur le revenu. De la même façon, les procédures relatives à la saisie des données, à la correction et au suivi sont fidèles au même principe, quelle que soit la source de données même si les demandes de suivi en ce qui a trait aux données fiscales sont transmises à Revenu Canada plutôt qu'aux entreprises.

La procédure de "transformation" est fondée sur une notion relativement nouvelle. Elle vise à convertir les données fournies **par** les unités déclarantes **en** données destinées aux unités statistiques cibles, favorisant ainsi l'interprétation des données d'une enquête à l'autre. Cette procédure va à l'encontre de celles qui existent déjà en vertu desquelles les données de l'unité déclarante particulière à l'enquête constituent le produit final. Dans le cas des données fiscales, cette transformation se produit au moment où l'on établit une distinction entre les unités statistiques liées à l'échantillon fiscal de la première étape et les déclarations fiscales sélectionnées.

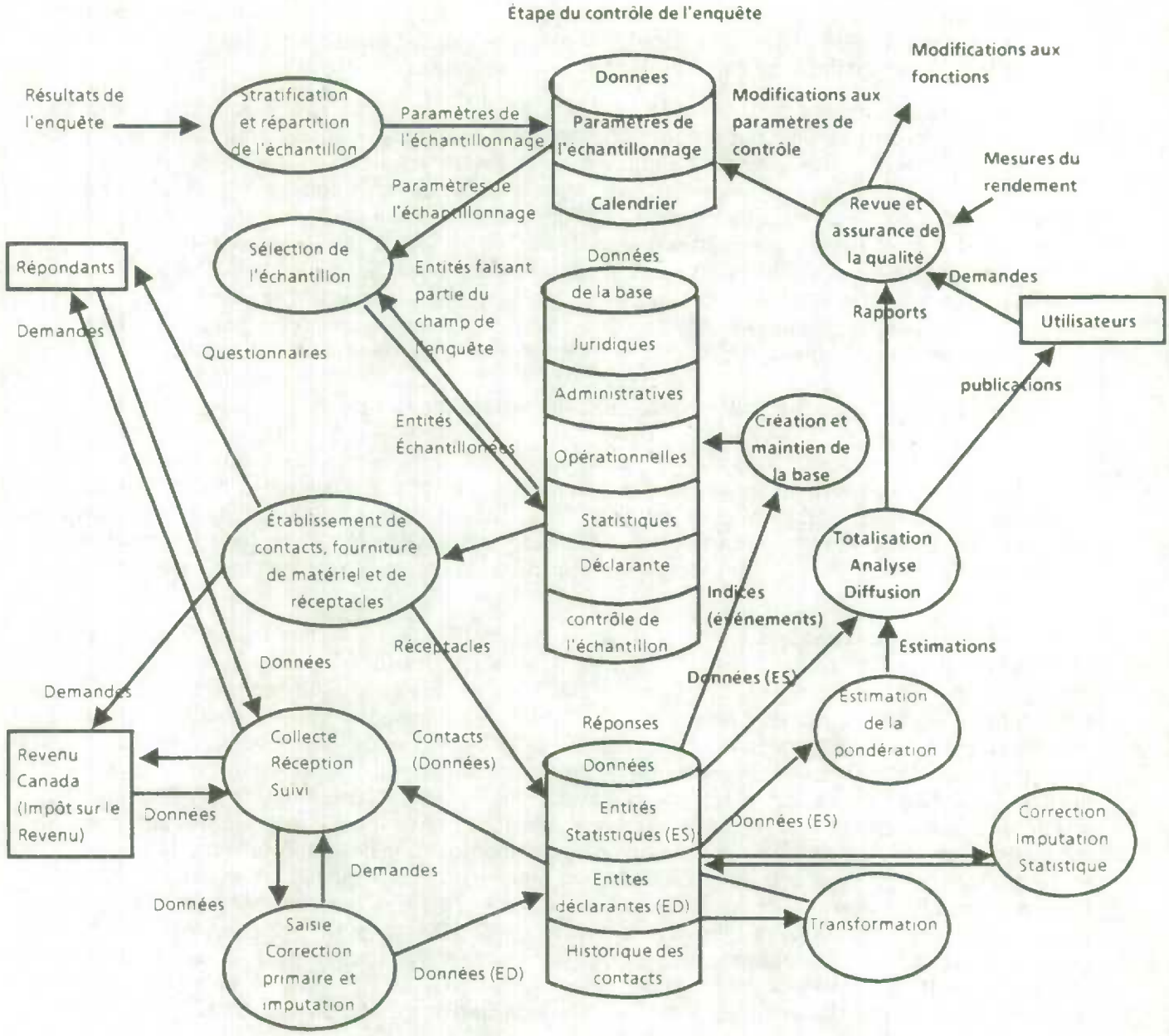


Figure 6: Fonctions génériques, flux et stockage des données
(version simplifiée excluant les détails relatifs à la création et au maintien de la base)

En ce qui a trait à la correction et à l'imputation, on utilise les données administratives fournies par les enregistrements sous une forme ordinolingue ou provenant de l'échantillon fiscal pour procéder à des imputations des données d'enquêtes manquantes. Les données administratives peuvent également jouer le rôle de données auxiliaires à l'étape de la pondération et de l'estimation de l'échantillon, par exemple en fournissant des totaux qui servent de points de repère pour le Programme de revenu du travail. On envisage également la possibilité d'utiliser les données se rapportant au paiement des retenues sur la paie pour l'enquête mensuelle sur l'emploi, la rémunération et les heures de travail (Cotton, 1987).

Les données administratives sont précieuses à des fins d'analyse, d'évaluation et de vérification. Pour ce qui est des totalisations et de la diffusion, il faut tenir compte des restrictions imposées par les organismes administratifs au sujet de l'utilisation de leurs données. Ceci peut restreindre la diffusion qui, autrement, serait possible. Par exemple, on peut communiquer aux administrations provinciales les micro-données d'enquête mais non pas les données fiscales se rapportant aux unités.

En ce qui a trait à la qualité, Statistique Canada ne peut exercer le même contrôle qualitatif sur les procédures administratives que sur le traitement des données effectué par l'organisme lui-même. Il faut donc évaluer soigneusement la qualité des données transmises et aviser immédiatement la source pertinente en cas de lacune. Il faut également superviser les procédures administratives pour garantir que les changements sont effectués en toute connaissance des conséquences qu'ils auront pour le programme statistique. Il est indispensable d'assurer la liaison permamente avec les organismes administratifs (Gates, 1987).

5. CONCLUSION

Le présent document constitue un résumé du projet de remaniement des enquêtes-entreprises en cours à Statistique Canada du point de vue de l'utilisation des données administratives. Les applications des enregistrements administratifs ont été décrits et classés en trois catégories: définition et maintien de la base, remplacement partiel de la collecte de données d'enquêtes annuelles par des données fiscales et autres utilisations dans le contexte du processus d'enquête. On a souligné en particulier les deux sources principales, c'est-à-dire les retenues sur la paie et l'impôt sur le revenu. Voici un résumé des conclusions.

Les données administratives jouent un rôle crucial dans la définition de la base de données du registre central devant desservir l'ensemble du programme des enquêtes-entreprises. Ces données constituent le fondement de **toutes** les bases d'enquête. Sans elles, il faudrait investir des sommes astronomiques dans l'établissement de bases. Dans le cas de l'échantillonnage des petites entreprises, les unités administratives servent d'unités statistiques. Cette approche ne convient toutefois pas pour les entreprises importantes. Par conséquent, les procédures diffèrent selon la taille de l'unité. Dans le cas des entreprises importantes, les données administratives constituent le point de départ du processus ("établissement de profils") de délimitation des unités statistiques. Aucune source administrative ne fournit une couverture suffisante pour compléter l'échantillon des petites entreprises, ce qui explique alors que la partie non intégrée (PNI) de la base de données du registre central contient des bases alternatives fondées sur deux sources administratives distinctes, les retenues sur la paie et l'impôt sur le revenu. Dans le cas de certaines unités administratives, il est impossible d'obtenir les renseignements relatifs à la classification qui sont nécessaires pour stratifier la PNI. Pour les obtenir, il faut communiquer directement avec les unités. Compte tenu des frais que cela entraîne, on utilise un échantillonnage à deux phases de la PNI fondée sur l'impôt sur le revenu afin d'obtenir tous les renseignements nécessaires à la classification des seules unités de l'échantillon de la première étape. On ne tient compte des distinctions qui, idéalement, devraient exister entre les unités administratives et les unités statistiques de la PNI que dans le cas des unités sélectionnées dans l'échantillon fiscal de première phase ou d'autres échantillons d'enquête.

Le traitement des enregistrements administratifs et des résultats d'enquête, de même que des renseignements découlant de la revue de la base proprement dite, permettent d'assurer le maintien de la base de données du registre central. On prend note des "indices" qui indiquent un changement et on les classe dans l'une ou plusieurs catégories "d'événements types" en fonction desquelles on procède à la mise à jour de la base de

données du registre central. Dans le cas des unités de la PNI, la conversion des signaux administratifs en événements types est informatisée mais il faut que les employés examinent les unités de la PI. On a envisagé d'automatiser les procédures de traitement des indices administratifs mais, jusqu'à présent, on n'a pas encore atteint cet objectif. L'informatisation permettrait de faire des économies appréciables au niveau des employés de bureau.

Dans le contexte du remplacement de la collecte des données d'enquête par des données administratives et de la réduction subséquente du fardeau du répondant et des frais de fonctionnement, il faut souligner l'utilisation des données fiscales. En vertu de la stratégie de la BDRC, les données financières sont tirées des déclarations fiscales pour toutes les petites unités faisant partie de l'enquête annuelle sur la production économique. Il arrive qu'on entre en communication avec les petites unités sélectionnées mais seulement pour recueillir des renseignements additionnels au sujet de la classification industrielle (dans les cas où les données fiscales sont insuffisantes), et d'autres caractéristiques non financières. En ce qui a trait à l'enquête annuelle sur les statistiques financières et fiscales des sociétés, les déclarations d'impôt sur le revenu sont l'unique source de toutes les données, à l'exception des données financières dans le cas des très grandes sociétés, où l'on ne peut se servir de l'unité administrative et où il faut avoir recours aux données de l'enquête financière trimestrielle. C'est le seul cas au cours des dernières années où les données administratives n'ont pas été utilisées autant que prévu.

Il est difficile d'évaluer la qualité des données fiscales qui remplacent les données d'enquête. Des comparaisons de la valeur des données tirées des déclarations d'impôt sur le revenu et des questionnaires d'enquêtes annuelles révèlent des écarts appréciables entre les deux sources, mais il n'est pas facile de déterminer laquelle fournit des données plus "précises" compte tenu de la réticence des répondants à répondre à des questions au sujet de leurs déclarations d'impôt sur le revenu. On est d'avis que les valeurs qui figurent sur les états financiers joints aux déclarations d'impôt sur le revenu sont vraisemblablement plus précises que les données fournies en réponse à une enquête, mais que les données financières ne permettent pas toujours d'effectuer la ventilation précise des éléments financiers requis.

Outre ces importantes applications, les données administratives servent à établir des imputations dans les cas où il y a des données d'enquête qui manquent, à établir des totaux servant de points de repère pour la répartition ou l'estimation de l'échantillon et à favoriser l'évaluation et la vérification des enquêtes. En résumé, les données administratives jouent un rôle très important au sein du programme d'enquêtes-entreprises, rôle qui prendra sans doute de l'ampleur.

BIBLIOGRAPHIE

Armstrong, G., Monty, A., Woods (1986). "Definitions of Standard Events", Business Survey Redesign Project Working Paper, Statistique Canada, Ottawa.

Bankier, M., Bordeleau, C., Carruthers, I., Demmons, P., Finlay, M., et Leduc, J., (1985). "Preliminary Report on Documentation of Administrative Data Sources", Business Survey Redesign Project Working Paper, Statistique Canada, Ottawa

Beckstead, D., et al (1985). "PD Processing Time Lag Study", Business Register Working Paper, Statistique Canada, Ottawa.

Brackstone, G.J. (1987). Issues in the Use of Administrative Records for Statistical Purposes, Presented at the Statistical Society of Canada Annual Meeting, 1987, Québec City.

Cain J., et al (1984). "Infrastructure Development, Objectives, Policy and Strategy", Business Survey Redesign Project Working Paper, Statistique Canada, Ottawa.

Cochran, W.G. (1977). "Sampling Techniques", Wiley, New York.

Clark, C., et Lussier, R. (1987). "The Use of Administrative Data for Initial and Subsequent Profiles of Economic Entities", *Proceedings of the Section on Survey Methods*, 1987, American Statistical Association, Washington.

Colledge, M. (1987). "The Business Survey Redesign Project - Implementation of a New Strategy of Statistique Canada", presented at the Bureau of the Census Third Annual Research Conference, Washington.

Colledge, M., et Lussier, R. (1985). "A Strategy for the Provision of Frame Data and Use of Tax Data for Economic Surveys", *Proceedings of the Section on Survey Methods, 1985*, American Statistical Association, Washington.

Colledge, M., Estevao, V., et Foy, P. (1987). "Experiences in Coding and Sampling Administrative Data", Proceedings of the section on Survey Methods 1987 (to be published). American Statistical Association, Washington.

Estevao, V., Ambroise, P., et Colledge, M. (1983) "A study on the Quality of Certain Fields of the Business Register Master File", Business Register Working Paper, July 1983, Statistique Canada, Ottawa.

Estevao, V., et Tremblay, J. (1985). "A Report on the Quality of the Data in the BRMF - SARUS Study" - 1984/85.

Estevao, V., et Tremblay, J. (1986a). "An Evaluation of the Assignment of Standard Industrial Codes from PD-20 Data", Business Survey Redesign Project Working Paper, Statistique Canada Ottawa.

Estevao, V., et Surman, P. (1987b). "A Study on the Use of Research Information to Obtain Complete SIC Codes for Incorporated Businesses", Business Survey Redesign Project Working Paper, Mars 1987. Statistique Canada, Ottawa.

Estevao, V., et Tremblay J. (1986b). "An Evaluation of the Assignment of Standard Industrial Codes from T2 Tax Data", Business Survey Redesign Project Working Paper, Novembre 1986, Statistique Canada, Ottawa.

Foy, P. (1987). "Two Phase Sample Design for Estimation from Tax Data for Annual Surveys of Economic Production", Business Survey Redesign Working Paper, Septembre 1987, Statistique Canada, Ottawa.

Foy, P., et Corriveau, P. (1986). "Evaluation preliminaire de l'emploi d'un échantillon maître des comptes PD dans la partie non-intégrée du CFDB" Business Survey Redesign Project Working Paper, Mars 1986, Statistique Canada, Ottawa.

Foy, P. (1987). "Development of the OC Capability for the Annual Surveys of Economic Production", Juillet 1987, Business Survey Redesign Project Working Paper, Statistique Canada, Ottawa.

Hostetter, S.C. (1983). "The Verification Method on a Solution to the Industry Coding Problem", *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Washington.

Konschnik, C., Monsour, N., et Detlefsen, R. (1985). "Constructing and Maintaining Frames and Samples for Business Surveys". *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington.

Lussier, R., et Colledge, M. (1987). "Business Survey Redesign Project Quality Assurance Strategy", Business Survey Redesign Project Working Paper, Juillet 1987, Statistique Canada, Ottawa.

Statistique Canada (1970). "Standard Industrial Classification 1970, Catalogue 12-501E, Statistique Canada. Ottawa.

Statistique Canada. (1980). "Standard Industrial Classification 1980", Catalogue 12-501E, Statistique Canada, Ottawa.

Sunter, A. (1986). "Implicit Longitudinal Sampling from Administrative Files: A Useful Technique", *Journal of Official* Statistique, Vol. 2, No.2, 161-168.

Sunter, A. (1987). "A Note on the Tax Universe Master Sample", Business Survey Redesign Project Working Paper, Juin 1987, Statistique Canada, Ottawa.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

ESTIMATION DE L'EMPLOI PAR PETITE RÉGION ET SELON LES HEURES DE TRAVAIL

SIXTEN LUNDSTRÖM[1]

RÉSUMÉ

Jusqu'en 1980, la formule du recensement de la population et du logement en Suède comportait des questions sur l'emploi mais, depuis 1985, l'intention est de relever les données sur l'emploi principalement à partir des registres. Une lacune de cette nouvelle source de données est qu'elle ne renferme aucune information sur les heures de travail. C'est pourquoi nous avons tenté de mettre au point un estimateur pour les petites régions, qui réunisse les données de l'enquête sur la main-d'oeuvre et celles du registre fiscal. L'auteur décrit les simulations, entre autres, d'un estimateur de modèle logit et d'un estimateur synthétique, ainsi que les améliorations apportées à l'estimateur sélectionné.

1. INTRODUCTION

Nombreux utilisateurs, tant du secteur privé que du secteur public, ont besoin de renseignements annuels sur les heures de travail des personnes économiquement actives. Les autorités municipales ont un grand besoin de ces données pour planifier les services de garderie, évaluer les besoins en transport en commun et établir des prévisions de l'emploi. Antérieurement, le recensement suédois de la population et du logement constituait une source quinquennale de données à ce sujet mais, à compter du recensement de 1985, on se renseignera sur l'emploi au moyen d'un système de registres. Ces registres, toutefois, ne comportent aucune indication des heures de travail. Les domaines sont tellement petits que les estimateurs conventionnels, fondés sur des échantillons à l'échelle nationale, ne produisent pas d'estimations fiables; nous nous sommes donc servis d'information auxiliaire pour élaborer un estimateur à modèle, appelé quelquefois estimateur synthétique. Nous n'aurons probablement pas l'occasion d'effecturer des évaluations approfondies à l'avenir; c'est pourquoi il importe d'élaborer un estimateur qui donnera de bons résultats sur une longue période. L'estimation à modèle est encore relativement peu répandue à Statistique Suède. Nous préférons, pour éviter toute confusion, un estimateur qui soit le moins compliqué possible.

2. SOURCES DE DONNÉES

Les données que nous voudrions utiliser pour effectuer notre estimation viendront principalement de deux sources: la statistique sur l'emploi régional (SER) et l'enquête sur la population active (EPA). Il se peut que nous utilisions aussi des données du recensement de 1980. La SER est basée sur un ensemble de plusieurs registres dont le plus important est le registre fiscal. Chaque année, les employeurs déclarent au fisc le revenu de tous leurs employés. Statistique Suède reçoit une copie de ce registre et le fusionne avec le

[1] Sixten Lunström, Statistique Suède, U/STM-0, 701 89 Örebro, Suède.

registre des entreprises et celui de la population totale. Ainsi, nous pouvons produire des statistiques sur la population économiquement active. Ce processus comporte toutefois une lacune: nous ne savons rien des heures de travail. L'EPA, par contre, nous éclaire à ce sujet: chaque mois, l'échantillon englobe environ 12,000 personnes économiquement actives. L'EPA comporte des variables qu'on retrouve également dans la SER. Les variables les plus importantes, en ce qui a trait à l'élaboration d'un estimateur des heures de travail, sont le sexe, l'âge, le revenu et l'industrie.

Le recensement mesure la situation d'emploi pour une semaine donnée, en novembre, tous les cinq ans. Les utilisateurs aimeraient une série ininterrompue de données, même si l'on adopte une nouvelle méthode de collecte des données. Il est difficile de répondre à leur désir. Relativement à la variable des "heures de travail par semaine", nous allons présenter certains résultats d'une comparaison de données du recensement et de l'EPA. La SER présente une difficulté lorsqu'il s'agit d'isoler la population économiquement active, mais nous n'en traitons pas ici.

3. LES ESTIMATEURS

Il y a beaucoup d'estimateurs pour petites régions, et l'on ne peut pas les comparer tous dans le cadre d'une seule et même étude. La présente est donc limitée aux estimateurs qui nous inspirent le plus de confiance et, à des fins de comparaison, il y a aussi un estimateur conventionnel.

Supposons qu'une population de taille N se compose de Q petites régions exhaustives et mutuellement exclusives, désignées $q = 1, \ldots, Q$. Supposons en outre que chaque région peut à son tour être répartie en H classes exhaustives et mutuellement exclusives, au moyen d'une partie ou de la totalité des variables sexe, âge, revenu et de l'industrie, appelées $h = 1, \ldots, H$. La désignation donne une classification recoupée en HQ cellules comptant N_{hp} membres dans la hq^e cellule, avec un échantillon correspondant de membres dans un échantillon aléatoire simple de taille n. En vue d'effectuer la sommation pour l'indice inférieur, nous remplaçons cet indice par '.'.

Nous voulons estimer le pourcentage de personnes économiquement actives pour une catégorie d'heures de travail:

$$T_q = \frac{100}{N_{.q}} \sum_{i=1}^{N_{.q}} y_i, \qquad (1)$$

où

$$y_i = \begin{cases} 1 \text{ si la } i^e \text{ personne fait partie de la catégorie donnée d'heures de travail} \\ 0 \text{ autrement} \end{cases}$$

L'estimateur synthétique SYNT présuppose que dans le h^e sous-groupe, les moyennes des Q petites régions sont à peu près égales. Le pourcentage prévu pour la région q est donc calculé au moyen de la formule suivante:

$$SYNT = \frac{100}{N_{.q}} \sum_{h=1}^{H} N_{hq} \hat{\bar{Y}}_{h.}, \qquad (2)$$

où

$$\hat{\bar{Y}}_{h.} = \sum_{i=1}^{n_{h.}} y_i / n_{h.}$$

Contrairement à l'estimateur SYNT, l'estimateur ci-dessus est basé sur des totaux des revenus plutôt que sur des comptes:

$$\text{RAPPORT DU SYNT} = \frac{100}{N_{.q}} \sum_{h=1}^{H} X_{hq} \sum_{i}^{n_{h.}} y_i / \sum_{i}^{n_{h.}} x_i, \qquad (3)$$

où

x_i est le revenu de la i^e personne

et

$$X_{hq} = \sum_{i=1}^{N_{hq}} x_i$$

Il semble plausible de supposer que la probabilité de travailler plus qu'un nombre donné d'heures par semaine augmente avec le revenu. Il semble plausible aussi que cette probabilité s'accroisse plus au milieu de l'échelle des revenus qu'au bas ou au sommet de celle-ci. Le modèle logit est donc assez évidement celui qu'il faut retenir.

Nous avons estimé les coefficients de régression suivants en nous servant des moindres carrés pondérés.

$$u_h = \alpha + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon ,$$

où

$$u_h = {}^{e}\log \frac{p_h}{1-p_h} ,$$

et

p_h = la probabilité qu'un membre du sous-groupe h fasse partie d'une catégorie donnée d'heures de travail.

La matrice $\underset{\sim}{x}' = (x_1, \ldots, x_m)$ dénote un ensemble de variables factices pour les variables catégorielles comme le sexe, l'âge (à l'intérieur des catégories), la branche d'activité, et une variable pour le revenu moyen de chaque catégorie. Lorsqu'on estime les coefficients, la proportion observée pour une catégorie donnée d'heures de travail dans le sous-groupe h sert d'estimation de P_h.

L'estimateur LOGIT est donné par la formule suivante:

$$\text{LOGIT} = \frac{100}{N_{.q}} \sum_{h=1}^{H} N_{hq} \, p_h^{*} , \qquad (4)$$

où P_h^{*} est la proportion prédite dans le sous-groupe h et est calculée à partir de

$$p_h^{*} = e^{u_h^{*}}/(1+e^{u_h^{*}}),$$

et

$$u_h^* = \hat{\alpha} + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_m x_m.$$

Remarque: Dans l'étude de la méthodologie, à la section 4, on se sert de l'estimateur LOGIT pour diverses catégories d'heures de travail. Tout au long de la présente étude, il est jugé que les heures de travail se répartissement en deux catégories — la catégorie donnée, d'une part, et toutes les autres catégories, d'autres part.

Antérieurement, Statistique Suède s'est penché sur un estimateur à modèle pour calculer le nombre de personnes économiquement actives dans chaque petite région durant la période intercensitaire. Cet estimateur est appelé l'estimateur SPINK, et il est défini comme suit:

On suppose que la population a été triée suivant la variable x (revenu). L'estimateur prend donc la forme suivante:

$$SPINK = \frac{100}{N_{.q}} \sum_{h=1}^{H} \sum_{i=1}^{N_{hq}} z_{hi}, \qquad (5)$$

où

$$z_{hi} = \begin{cases} 1 & \text{si la } i^e \text{ personne fait partie du sous-groupe } h \\ & \text{et a un revenu qui se situe entre } [\hat{G}_{1h}, \hat{G}_{2h}] \\ 0 & \text{autrement} \end{cases}$$

G_{1h} G_{2h} indiquent les limites inférieure et supérieure de la valeur qui représente un revenu "raisonnable" pour un membre d'une catégorie donnée d'heures de travail. La manière dont ces limites sont estimées est expliquée dans le texte qui suit et illustrée à la figure 1.

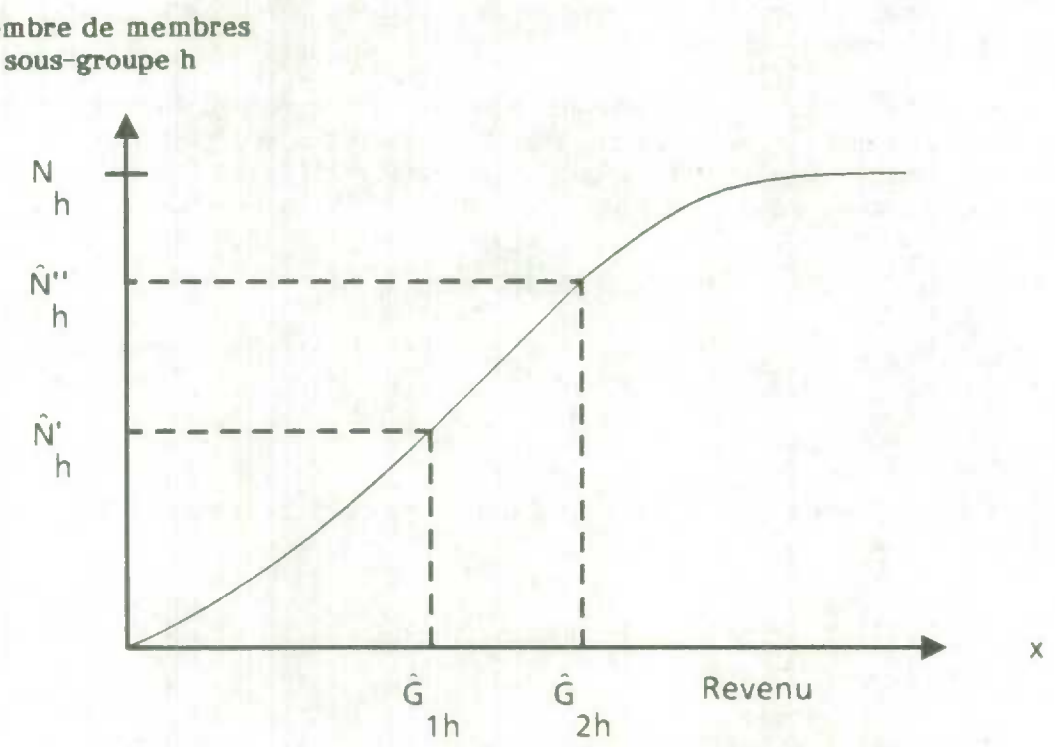


Figure 1 Nombre de membres du sous-groupe h dont le revenu est inférieur à x.

$\hat{N}'_h$ dénote le nombre estimatif de membres de la population qui se retrouvent dans les catégories d'heures de travail inférieures à la catégorie donnée. Le nombre de membres de la catégorie donnée, N_{Gh}, est calculé au moyen de la formule suivante:

$$\hat{N}_{Gh} = N_{h.} \sum_{i=1}^{n_{h.}} y_i / n_{h.}$$

et

$$\hat{N}''_h = \hat{N}'_h + \hat{N}_{Gh}.$$

Les limites $\hat{G}_{1h}$ et $\hat{G}_{2h}$ du revenu sont alors estimées par projection sur l'axe des revenus.

Särndal (1984) a élaboré, en se servant de la méthode de la régression généralisée, un estimateur asymptotiquement sans biais. Celui-ci comporte deux parties, soit l'estimateur synthétique (SYNT) moins une estimation du biais de l'estimateur synthétique. Le biais est souvent modifié par une erreur d'échantillonnage élevée. Cassel (1984) suggère, pour minimiser l'erreur quadratique moyenne (sous réserve de certaines hypothèses) de multiplier le biais par une constante α.

L'estimateur synthétique optimal corrigé (KORRSYNT) a la forme suivante:

$$\text{KORRSYNT} = 100 \left(\sum_{h=1}^{H} N_{hq} \hat{\bar{Y}}_{h.} - \hat{\alpha} \frac{N}{n} \sum_{i=1}^{n_{.q}} e_i \right) / N_{.q}, \qquad (6)$$

où

$$e_i = (y_i - \hat{\bar{Y}}_{h.}) \text{ pour les membres du sous-groupe } h,$$

et

$$\hat{\alpha} = 1 - \text{var}\,(\hat{A}_C)/\hat{A}_C^2,$$

où

$$\hat{A}_C = \frac{N}{nN_{.q}} \sum_{i=1}^{n_{.q}} e_i.$$

Les estimateurs susmentionnés sont à modèle et par conséquent biaisés (l'estimateur KORRSYNT est asymptotiquement sans biais à dessein). Nous avons voulu comparer nos estimateurs à modèle à l'estimateur direct conventionnel sans biais (DIR). Celui-ci utilise la moyenne de l'échantillon dans la région q, sans information supplémentaire.

$$\text{DIR} = 100 \sum_{i=1}^{n_{.q}} y_i / n_{.q}. \qquad (7)$$

4. DESCRIPTION DE LA SIMULATION

Toute évaluation d'estimateurs à modèle pour petites régions est problématique. Il faut des valeurs réelles pour estimer des quantités telles que le biais et le carré de l'erreur quadratique moyenne. Le plus souvent, ce genre d'estimateur est tellement compliqué qu'il faut recourir à des simulations. À cause des contraintes financières, le nombre et les genres de petites régions à inclure dans l'étude et la taille de l'échantillon sont plus petits qu'on souhaiterait. L'étude se ressent de ces difficultés générales, de même que des difficultés plus particulières qui se rattachent aux enquêtes dont nos données proviennent.

Nous nous attendons de trouver une corrélation entre le nombre d'heures de travail par semaine en novembre et le revenu annuel, et nous espérons nous servir de cette corrélation pour diminuer le biais d'estimation. Certaines personnes ne travaillent qu'une partie de l'année et, par conséquent, gagnent peu. Le registre fiscal renferme certaines données sur les heures de travail, mais elles sont de piètre qualité. Nous disposerons de données de meilleure qualité plus tard, lorsque la collecte des données aura été améliorée.

4.1 Mesures d'erreurs

L'erreur quadratique moyenne (EQM) englobe à la fois la variabilité d'échantillonage et le biais de l'estimateur, et il a la forme suivante:

$$EQM(\hat{T}_q) = Var(\hat{T}_q) + B^2(\hat{T}_q), \tag{8}$$

où

$Var(\hat{T}_q)$ est la variance et $B(\hat{T}_q)$ est le biais de l'estimateur $\hat{T}_q$.

Dans le cadre de la présente étude, nous avons estimé que $REQM(\hat{T}_q) = (EQM(\hat{T}_q)^{\frac{1}{2}}$, $Var(\hat{T}_q)$ et $B(\hat{T}_q)$. Nous avons estimé aussi le biais relatif $rel - |B|$, qui découle de:

$$rel - |B| = 100 \cdot |B(\hat{T}_q)|/REQM(\hat{T}_q). \tag{9}$$

Pour en arriver à une mesure sommaire, nous avons aussi calculé les moyennes (pour les petites régions) des composantes d'erreur. Nous avons effectué une simulation suivant la méthode de Monte Carlo pour calculer ces valeurs pour les divers estimateurs dont il a été question dans la section précédente. Pour chaque échantillon r $(r=1, \ldots, R)$ nous avons calculé la valeur de l'estimateur $\hat{T}_{q(r)}$ et enfin

$$B(\hat{T}_q) = \left(\sum_{r=1}^{R} \hat{T}_{q(r)}/R \right) - T_q, \tag{10}$$

et

$$\mathrm{Var}(\hat{T}_q) = \left| \sum_{r=1}^{R} \hat{T}^2_{q(r)} - \left(\sum_{r=1}^{R} \hat{T}_{q(r)} \right)^2 / R \right| /R.$$

Nous avons pu calculer tous les éléments d'erreur au moyen des formules ci-dessus.

4.2 Taille des échantillons

Les estimateurs à modèle prennent un échantillon d'information pour chaque sous-groupe h afin d'estimer le nombre de personnes économiquement actives dans une catégorie d'heures de travail. On s'attend alors à ce que la variabilité d'échantillonnage soit faible même si l'échantillon est petit, disons mille personnes. Il peut être prévu que le biais constituera une proportion importante de l'erreur totale.

L'estimateur DIR, sans biais à dessein est basé seulement sur les observations effectuées dans la petite région et dépend beaucoup plus de la taille de l'échantillon que les estimateurs à modèle.

On se sert de deux tailles d'échantillons, 1,000 et 2,500, pour effectuer la simulation. Les tailles de ces deux échantillons donnent une erreur d'échantillonnage un peu trop forte pour les estimateurs à modèle et trop faible pour les DIR, comparativement à ce qu'on obtiendrait en utilisant les données de l'EPA régulière.

4.3 Résultats

La première mini-population étudiée était composée de l'échantillon de l'EPA pour les mois d'octobre, novembre et décembre 1980 et de groupes hors-échantillon pour tout le reste de la même année. Nous avons tenté, en se servant des registres, d'exclure les personnes qui ne travaillent pas toute l'année, mais ce tri n'a pas donné les résultats espérés. Nous avons tenté également d'exclure les employeurs parce que ce groupe n'appartiendra probablement pas à la population englobée par la SER.

Nous avons fusionné cette population au registre du recensement de 1980, tirant ainsi de l'information au sujet des heures de travail tant de l'EPA que du recensement.

Après ces réductions, la population étudiée se composait d'environ 35,000 personnes économiquement actives. Cette population étant trop restreinte pour permettre une étude des estimations au niveau des municipalités, nous avons décidé de la faire plutôt par comté. Le registre démographique comprenait des données sur le sexe, le revenu, les heures de travail (tirées de l'EPA et du recensement) et un code de comté.

Nous avons effectué des simulations pour diverses catégories d'heures de travail et pour divers regroupements de catégories de revenus, ainsi que pour des échantillons de deux tailles. L'analyse des résultats mène à la même conclusion; c'est pourquoi nous ne présentons qu'une des simulations.

Nous présentons dans le tableau 1 les estimations de l'REQM calculées par voie de simulation, pour un échantillon de 1,000 personnes et lorsque H=16 (catégories de revenus * sexe). Le paramètre T_q représente le pourcentage de personnes économiquement actives qui travaillent plus de 34 heures par semaine. Les divers comtés ont une valeur T_q dans l'intervalle de 68 à 80.

Tableau 1

REQM $(\hat{T}_{q)}$ estimative pour divers estimateurs, lorsque n = 1,000. $(68<T_q<80)$.

COMTÉ	SYNT	SYNTRATIO	LOGIT	SPINK	KORRSYNT	DIR
01	1.59	4.68	1.42	3.36	1.88	3.29
03	2.18	2.98	1.98	1.99	3.69	8.13
04	1.71	1.34	1.42	1.95	3.46	7.85
05	1.47	1.30	1.34	1.45	3.28	7.43
06	1.32	1.37	1.56	1.88	2.72	7.46
07	1.32	1.63	1.44	2.06	3.07	7.59
08	1.42	1.32	1.36	1.72	3.40	8.51
09	3.82	4.19	4.51	5.84	4.85	8.13
10	1.29	2.04	2.00	1.41	3.37	8.03
11	1.36	1.71	1.75	2.66	2.98	8.30
12	1.33	1.56	1.31	1.39	2.44	5.26
13	1.33	1.53	1.27	1.54	3.19	8.51
14	1.58	2.36	1.32	2.31	2.63	5.67
15	2.41	2.44	3.01	2.79	3.39	7.15
16	2.03	2.64	2.76	2.41	3.88	8.29
17	1.71	2.74	1.90	3.00	3.32	7.72
18	1.63	1.84	2.26	1.68	2.79	7.61
19	1.30	1.31	1.49	1.52	2.99	6.82
20	2.15	2.76	2.90	2.73	3.93	7.86
21	1.28	1.39	1.80	1.42	3.28	7.68
22	1.66	2.21	1.93	1.60	3.00	7.90
23	2.29	2.57	3.25	2.70	3.52	7.97
24	1.49	1.40	1.30	1.45	3.67	7.87
25	4.76	3.71	3.60	5.60	5.02	8.52
Moyenne	1.85	2.21	2.04	2.35	3.32	7.48

Le tableau 1 indique que le SYNT est le meilleur estimateur lorsqu'on se sert du REQM comme mesure de l'erreur. Les autres simulations effectuées sur cette mini-population ont aussi mené à cette conclusion. L'estimateur sans biais DIR et l'estimateur asymptotiquement sans biais KORRSYNT donnent une forte EQM qui s'explique principalement par la variabilité d'échantillonnage. Cette constatation ressort indirectement du tableau 2, où nous indiquons le biais estimatif pour les divers estimateurs.

Tableau 2

$B(\hat{T}_q)$ estimatif pour divers estimateurs

COMTÉ	SYNT	SYNTRATIO	LOGIT	SPINK	KORRSYNT	DIR
01	.83	4.45	-.28	3.10	.55	.09
03	1.78	2.65	1.50	1.57	1.04	-.17
04	1.17	.45	.52	1.39	.72	-.77
05	.76	-.03	.27	.17	.51	.92
06	-.41	-.49	-.87	-1.32	-.24	-.39
07	-.45	-.98	-.66	-1.57	.04	.63
08	.65	.02	.26	-1.23	.27	.68
09	-3.58	-3.95	-4.30	-5.66	-1.71	.52
10	-.25	-1.54	-1.49	-.47	-.06	-.62
11	-.52	-1.06	-1.18	-2.34	-.37	-.83
12	.44	.73	-.23	.01	.34	.15
13	.45	.80	-.01	-.62	.76	.45
14	.92	1.93	.17	-1.95	.66	.30
15	-2.05	-2.07	-2.72	-2.32	-.82	-.27
16	-1.58	-2.30	-2.43	-2.05	-.59	.74
17	-1.15	-2.40	-1.37	-2.63	-.26	.52
18	-1.04	-1.32	-1.85	-1.11	-.59	.40
19	.25	.12	-.69	.11	.45	.86
20	-1.73	-2.42	-2.57	-2.31	-.60	.37
21	-.21	-.45	-1.24	-.59	-.04	-.30
22	-1.08	-1.81	-1.41	-.91	-.43	.66
23	-1.87	-2.19	-2.96	-2.20	-1.05	-.60
24	.79	-.57	.08	-.17	.64	-.34
25	4.57	3.46	3.34	5.41	2.27	-.88
Moyenne des valeurs absolues	1.19	1.59	1.35	1.72	.63	.52

Le tableau 2 montre que l'importance et le signe du biais des estimateurs à modèle varient beaucoup entre les comtés. Les comtés où le biais est le plus élevé sont Gotland (09) et Norrbotten (25). Le comté de Gotland est une île avec une forte proportion d'agriculteurs (revenu imposable peu élevé mais longues heures de travail). Norrbotten est le comté le plus septentrional de la Suède, et l'on y trouve un grand nombre de mineurs (revenu imposable relativement élevé). Nous avons constaté qu'en incorporant la variable relative à la branche d'activité (voir section 4.4), les biais diminuaient.

Le biais est l'erreur dominante dans un estimateur à modèle, même lorsque l'échantillon ne comporte que 1,000 observations. On peut le constater en comparant les tableaux 1 et 2. Dans le cas de l'estimateur SYNT, la moyenne des biais relatifs est de 56 %.

Dans un cas, nous avons utilisé un échantillon de 2,500 personnes et nous avons constaté, comme prévu, que les EQM des estimateurs KORRSYNT et DIR diminuaient plus que celles des estimateurs à modèle lorsqu'on comparait les résultats d'échantillons de 1,000 et 2,500 personnes.

Comme on l'a mentionné antérieurement, une difficulté de la fusion de données de diverses sources est que la définition de la variable étudiée n'est pas la même dans tous

les cas, les méthodes de collecte des données diffèrent, etc. Dans le cas de cette minipopulation, nous avons pu comparer les valeurs de la variable étudiée, c'est-à-dire les heures de travail par semaine, de l'EPA et du recensement. Dans le tableau 3, nous présentons les erreurs découlant de l'utilisation des valeurs de l'EPA plutôt que de celles du recensement.

Tableau 3

Classement des erreurs relatives à la variable "heures de travail" dans l'enquête sur la population active

Pourcentage de l'estimation censitaire	Heures par semaine 1-19	20-34	35-s
a) Inclusion erronée	24.6	28.1	6.5
b) Exclusion erronée	49.2	19.8	6.2
c) Erreur brute	73.8	47.9	12.7
d) Classification correcte	50.8	80.2	93.4
e) Erreur nette	-24.6	8.3	0.2

Remarque: Certaines erreurs résultent du fait que la majeure partie des valeurs de l'EPA réfèrent à une période de temps différente.

Les erreurs présentées dans cette section ne comprennent pas celles qui figurent au tableau 3. Cette situation n'entraîne aucune sous-estimation grave de l'erreur totale lorsqu'il s'agit de l'emploi à plein temps mais, dans le cas des autres catégories d'heures de travail, il faut tenir compte des erreurs de classification. Nous avons aussi effectué des simulations sur des mini-populations où les petites régions étaient des municipalités. La conclusion reste la même: dans la plupart des cas, l'estimateur SYNT, de tous les estimateurs considérés, donne les meilleurs résultats.

Même si nous n'avions pas utilisé toutes les variables associées disponibles dans les simulations, nous maintiendrions quand même que l'estimateur SYNT est le meilleur. Son rendement au cours des études était bon, et un autre point favorable, sa simplicité.

Nous présentons dans la section suivante le travail que nous avons fait pour affiner l'estimateur SYNT.

4.4 Affinement de l'estimateur SYNT

Jusqu'à maintenant, nous n'avons utilisé que les variables associées au sexe et au revenu mais, en pratique, nous aurons même des renseignements sur l'âge et la branche d'activité. Nous présentons dans cette section notre tentative d'inclure ces variables aussi, en vue d'élaborer un estimateur SYNT optimal et, à longue échéance, robuste.

Cet affinement est réalisable sans effectuer de simulation parce qu'il est possible de déduire les formules de la variance (attendue) et du biais de l'estimateur SYNT. Les formules relatives à la variance attendue sont les suivantes:

$$E[Var(SYNT)] \doteq \left(\frac{100}{N_{.q}}\right)^2 \sum_{h=1}^{H} N_{hq}^2 \frac{\gamma_{h.}(1-\gamma_{h.})}{nW_h} \cdot K_h , \qquad (12)$$

où

$$K_h = 1 + \frac{1-W_h}{nW_h} \; ; \; W_h = \frac{N_{h.}}{N} \; ,$$

et pour le biais

$$B(SYNT) = \frac{100}{N_{.q}} \sum_{h=1}^{H} N_{hq} \, (\bar{Y}_{h.} - \bar{Y}_{hq}). \qquad (13)$$

Remarque: La formule pour la variance attendue est déduite de la même manière que dans Cochran (1977), section 5A.8 - "Stratification after Selection of the Sample".

Nous voulions utiliser des données sur les 284 municipalités de la Suède; la seule source de données était donc le recensement de la population et du logement. Une difficulté, en l'occurence, est que nous ne pouvions pas distinguer les personnes qui travaillent toute l'année, et nous nous attendions de surestimer l'erreur. Nous nous sommes convaincus en travaillant sur plusieurs mini-populations que, même si le niveau d'erreur était élevé, nous pouvions nous servir des données du recensement pour améliorer les procédures d'estimation.

Une caractéristique importante d'un estimateur est que son rendement soit bon à long terme. Pour éprouver ce rendement à long terme, nous avons élaboré l'estimateur SYNT en nous servant des données du recensement de 1980, puis nous l'avons testé avec celles du recensement de 1975.

Dans cette étude, les paramètres $\bar{Y}_{h\cdot}$, $\bar{Y}_{hq}$ et N_{hq} sont connus, et la variance (attendue) et le biais pour un échantillon d'une taille donnée, n, peuvent être calculés.

Dans le tableau suivant, le travail est résumé par une valeur moyenne (par municipalité) de l'EQM relative dont la forme est la suivante:

$$\text{Rel-REQM} = 100 \; . \; \text{REQM}/T_q \; , \qquad (14)$$

où $\text{REQM} = \{E[\text{Var}(\text{SYNT})] + B^2(\text{SYNT})\}^{\frac{1}{2}}$

Quatre variables associées, soit le sexe, l'âge, le revenu et la branche d'activité, sont classifiées d'une manière particulière et dénotées SEXE, ÂGE2, REV4 et BRA5 respectivement. Nous ne décrivons pas en détail les travaux qui ont mené à cette classification.

En pratique, l'estimateur est appliqué à trois échantillons consécutifs de l'EPA, totalisant 35,000 personnes. Pour que notre étude soit plus réaliste, nous avons aussi utilisé un échantillon de 35,000.

Tableau 4

Valeur moyenne (pour les municipalités) de l'REQM relative
Échantillon de 35,000

Variable associée	Heures de travail			
	Recensement de 1980		Recensement de 1975	
	20-s	35-s	20-s	35-s
SEXE * AGE2 * REV4 * BRA5	0.91	1.54	1.07	1.29
SEXE * REV4 * BRA5	0.92	1.55	1.10	1.34
AGE2 * REV4 * BRA5	0.89	1.94	1.05	1.72
REV4 * BRA5	0.91	1.90	1.09	1.76

Le tableau 4 indique que les résultats sont les meilleurs lorsqu'on utilise les quatre variables associées. L'exclusion de la variable ÂGE2 n'influe pas significativement la moyenne des REQM relatives. On constate par contre que le sexe est une variable associée importante, surtout lorsqu'on estime le pourcentage de personnes économiquement actives qui travaillent plus de 34 heures par semaine.

Pour estimer l'effet des variables associées, nous avons calculé la moyenne des REQM relatives pour un estimateur synthétique, en n'utilisant aucune information associée (c'est-à-dire que l'estimation nationale est utilisée pour chaque municipalité). Pour la catégorie des 35 heures par semaine, en nous servant des données du recensement de 1980, nous avons reçu la valeur 2.22, que l'on peut comparer à 1.54 dans le cas de l'estimateur "optimal". Dans le cas d'un échantillon de 35,000, où l'on utilise un estimateur DIR, la moyenne des REQM relatives est de 7.7.

Le tableau 4 indique que l'estimateur SYNT "optimal" a donné un bon rendement dans le cas des deux périodes étudiées.

Nous avons essayé également plusieurs autres formes d'affinement, que nous décrivons brièvement ci-dessous.

Si le biais est stable dans le temps, nous pourrions nous servir du biais calculé à partir des données du recensement de 1980 et le soustraire à l'avenir de l'estimation SYNT. Dans cette étude, nous effectuons cette rectification de l'estimateur pour le recensement de 1975 et les moyennes des REQM relatives pour les catégories de 20 heures et de 35 heures par semaine sont 1.07 et 1.12 respectivement. Ainsi, nous ne pouvons pas soutenir que le biais reste stable dans le temps.

Nous en sommes arrivés à cette conclusion également lorsque nous avons fait l'essai de ce qu'on appelle un estimateur SPREE de (Purcell (1979)), en utilisant les données du recensement de 1980 pour décrire la structure d'association.

Une manière de réduire le biais (erreur systématique) dans le cas d'un estimateur à modèle est de grouper les petites régions de telle sorte que les hypothèses sur lesquelles reposent les estimateurs aient plus de chance d'être réalisées. On relève dans les publications des essais assez bien réussis par (Purcell (1979) et Lundström (1987)), de réduire le biais (erreur systématique) d'un estimateur à modèle en regroupant des petites régions et en calculant des estimations pour chaque groupe. Il est vrai que la variabilité d'échantillonnage augmentera, mais cette augmentation ne sera peut-être pas aussi importante que la baisse du biais. Nous avons tenté de regrouper les municipalités mais nous avons constaté que les regroupements étaient sans effet après cinq ans.

5. CONCLUSION

L'estimateur synthétique simple SYNT est le meilleur des estimateurs étudiés, mais produit-il des estimations d'une qualité acceptable? Notre étude méthodologique ne répond pas à cette question. Il est difficile d'y répondre, bien sûr, parce que tout dépend de ce qu'on attend de ces estimations et de ce qu'on est disposé à céder en échange. Lorsqu'on se sert d'estimateurs conventionnels, on peut au moins établir approximativement les erreurs à partir de l'échantillon, et ainsi l'utilisateur des données reçoit une information adéquate. Lorsqu'on a recours à un estimateur synthétique, c'est le biais qui constitue l'erreur dominante, et cette erreur ne peut pas être estimée à partir des données présentement disponibles.

Les municipalités sont les plus petites régions utilisées aux fins de l'étude méthodologique, mais on a besoin d'estimations pour des régions beaucoup plus petites. Même dans le cas de ces régions plus petites, l'estimateur SYNT semble le meilleur.

Malgré toutes ces difficultés, Statistique Suède a décidé de se servir de l'estimateur SYNT.

REMERCIEMENTS

Je tiens à remercier Dr. Claes Cassel de ses nombreuses observations utiles durant les travaux, et M. Göran Råbäck pour les calculs statistiques qu'il a effectués aux fins de la rédaction du présent article.

BIBLIOGRAPHIE

Cassel, C. (1984). Optimal Selection of (X) (en Suédois). Note non publiée, Statistique Suède.

Cochran, W.G. (1977). *Sampling Techniques.* 3e édition. Wiley, New York.

Hammshek, E.A., et Jackson, J.E. (1977). *Statistical Methods for Social Scientists.* Academic Press, New York.

Hidiroglou, M.A., Morry, M., Dagum, E.B., Rao, J.N.K., et Särndal, C.E.(1984). Evaluation of Alternative Small Area Estimators Using Administrative Records. 1984 *Proceedings of the Survey Methodology Section, American Statistical Association.*

Lundström, S. (1986). Estimating Population Characteristics and Households in Swedish Municipalities Using Survey and Register Data. *Proceedings of the Second Annual Research Conference, March 23-26, 1986.* Bureau of the Census.

Lundström, S. (1987). An Evaluation of Small Area Estimation Methods: The Case of Estimating the Number of Nonmarried Cohabiting Persons in Swedish Municipalities, *Small Area Statistics, An International Symposium.* Wiley, New York.

Purcell, N.J. (1979). Efficient Estimation for Small Domains: A Categorical Data Analysis Approach. Dissertation de doctorat non publiée, University of Michigan.

Särndal, C.E. (1984). Design-Consistent Versus Model-Dependent Estimation for Small Domains. *Journal of the American Statistical Association,* Vol. 79, 624-631.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

MÉTHODOLOGIE DE LA CONSTRUCTION D'UN REGISTRE D'ADRESSES À PARTIR DE PLUSIEURS SOURCES ADMINISTRATIVES

J. DOUGLAS DREW, JOHN ARMSTRONG, ALEX VAN BAAREN, et YVES DEGUIRE[1]

RÉSUMÉ

Dans le cadre du programme de recherche pour le recensement de la population et du logement de 1991, une étude de la faisabilité de la construction d'un registre d'adresses de logements pour les régions urbaines du Canada est en cours à Statistique Canada. Le test pilote initial montre que l'utilisation d'un registre d'adresses construit à partir des données provenant de plusieurs systèmes de dossiers administratifs pourrait améliorer le champ d'observation du recensement. Le document décrit la méthodologie utilisée pour construire les registres d'adresses pour d'autres essais pilotes prévus pour l'automne 1987. Les sujets examinés comprennent la qualité des renseignements fournis par les différents fichiers administratifs, les procédures utilisées pour la standardisation des adresses et les techniques de jumelage d'enregistrements utilisées pour le non-dédoublement des listes d'adresses. Le document examine également les avantages de l'utilisation des renseignements ne se rattachant pas directement aux adresses dans le processus de couplage.

1. INTRODUCTION

L'idée d'utiliser un registre d'adresse pour les recensements de la population n'est pas nouvelle. Redfern (1987) souligne qu'il existe non seulement des registres d'adresse mais également des registres de la population en Suède, au Danemark et dans quelques autres pays européens et que l'existence et l'utilisation de tels registres ont contribué à modifier le rôle du recensement dans ces pays. De même, le Bureau of the Census des États-Unis se sert d'une liste d'adresses pour son recensement décennal. Cette base est constituée à partir de listes obtenues du secteur privé et elle est ensuite améliorée au moyen de vérifications sur le terrain (Whitford 1987).

Au Canada, le secteur privé n'offre pas de listes de bonne qualité et Statistique Canada a donc dû envisager, à diverses occasions, d'en construire. Je soulignerai ici qu'actuellement, pour le recensement du Canada, on utilise des listes d'adresses établies manuellement avec le concours de quelque 40,000 recenseurs responsables, chacun, d'un secteur regroupant de 200 à 300 logements. Ces listes, qui sont établies au moment de la livraison des questionnaires de recensement, ne sont pas saisies.

[1] Informatique et Méthodologie, Statistique Canada, 4-C2, Immeuble Jean-Talon, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

La première étude de faisabilité ayant trait à la construction d'un registre d'adresses a été effectuée par Fellegi et Krotki (1967). L'approche envisagée consistait à fusionner en une seule liste sans double compte les adresses tirées de sources multiples, c'est-à-dire des recensements antérieurs, des registres d'évaluation des municipalités et des listes utilisées par les compagnies d'électricité pour la facturation. Des registres d'adresses pilotes ont ainsi été construits et évalués pour deux villes de taille moyenne, Waterloo et London. Les auteurs de l'étude ont établi que ces registres répertoriaient 97% des logements, ce qui était un résultat très satisfaisant. Toutefois, en raison des limites de la technologie à cette époque, le processus de construction de ces listes devait être essentiellement manuel, si bien qu'on a renoncé alors à le mettre en oeuvre.

Au cours des années 70, une série d'études a été entreprise dont Booth (1976) donne un résumé. L'approche envisagée consistait à saisir les adresses répertoriées au recensement précédent et à mettre à jour le registre ainsi obtenu à l'aide des données de Postes Canada. Cette façon de faire assurait une couverture comparable à celle que l'on avait avec les méthodes traditionnelles. Toutefois, le coût initial de la saisie des données s'est avéré prohibitif, en dépit des économies que cette méthode devait permettre de réaliser à long terme, et le projet n'a pas été mis en oeuvre.

Royce (1986) a exposé plusieurs utilisations et avantages possibles d'un registre d'adresses dans le cadre des programmes de Statistique Canada, et énuméré plusieurs facteurs favorisant davantage la construction d'un tel registre que ce n'avait été le cas au cours des décennies antérieures. Il s'agit notamment de la plus grande disponibilité de fichiers administratifs qui renfermaient des adresses, de l'utilisation quasi universelle du code postal dans ces fichiers, de la baisse du prix des ordinateurs alliée à l'augmentation de leur puissance et de l'existence de méthodes et de logiciels plus perfectionnés pour le jumelage d'enregistrements. Dans un contexte aussi favorable, des recherches ont été entreprises en vue de construire un registre d'adresses et de l'utiliser au recensement de 1991, en partant de l'approche envisagée par Fellegi et Krotki antérieurement mais en informatisant pour ainsi dire toutes les étapes. Faute de données de qualité pour les adresses dans les régions rurales, l'étude a été restreinte aux régions urbaines.

Le tableau 1 présente les résultats relatifs à un registre pilote de petite envergure construit pour un secteur d'Ottawa qui compte 5,000 logements (Drew, Armstrong et Dibbs 1987). La couverture des logements valides obtenue avec ce registre d'adresses s'est avérée environ 1% moindre que celle que l'on a eue au recensement de 1986 pour le même secteur. Cependant, la combinaison de la liste tirée du recensement de 1986 et de ce registre d'adresses a donné une couverture des logements de 2.3% supérieure à celle que l'on avait eue lors du recensement. Il convient de souligner que les quartiers sélectionnés pour cet essai présentaient des risques de sous-dénombrement important. Le sous-dénombrement enregistré lors du recensement pour les quartiers sélectionnés a été estimé à 3.7% après vérification sur le terrain, de sorte que l'amélioration de 2.3% résultant de la combinaison de la liste tirée du recensement et du registre d'adresses équivaut à environ 60% du sous-dénombrement estimé des logements lors du recensement. Les estimations du surdénombrement pour la liste tirée du recensement et pour le registre d'adresses nécessaires au calcul du nombre net de logements présenté dans le tableau 1 ont été obtenues à la suite de vérifications sur le terrain.

Après des résultats aussi encourageants, un second essai a été planifié. L'essai porte sur deux façons d'utiliser un registre d'adresses pour la collecte des données du recensement, les deux s'appuyant sur la méthode actuelle de livraison des questionnaires de recensement. L'autre possibilité, qui consiste en un recensement postal basé sur l'utilisation d'un registre d'adresses en 1991, a été abandonnée dès le début de cette recherche du fait que l'étude effectuée a révélé que cela ne conduirait à aucune économie par rapport à la méthode traditionnelle dans la mesure où il faut prévoir une vérification sur le terrain préalable à l'utilisation du registre afin d'en améliorer la couverture (Gamache-O'Leary, Nieman et Dibbs 1986).

Tableau 1

Essai effectué à Ottawa: Nombre net de logements en pourcentage du nombre net de logements tiré du recensement*

	Type de logement		
	Individuel	Multiple	Total
Registre d'adresses	97.9	99.8	99.2
Recensement + Registre d'adresses	102.3	102.2	102.3

* Nombre net de logements = Nombre total de logements - Nombre de logements surdénombrés

Selon la première méthode, que nous appellons la méthode pré-liste, des registres d'adresses sont préalablement imprimés pour chaque secteur de dénombrement. La tâche du recenseur consiste à mettre à jour ces listes en faisant les ajouts et les suppressions nécessaires. La mise à jour est effectuée en même temps que la livraison des questionnaires de recensement aux logements valides.

Avec la seconde méthode, que nous appellons la méthode post-liste, les recenseurs doivent construire les listes d'adresses à partir des notes prises à l'occasion de la livraison des questionnaires, comme c'est le cas avec la méthode actuelle. Une fois la livraison terminée, on donne aux recenseurs une copie du registre d'adresses pour leur secteur de dénombrement, avec les instructions pour l'appariement leur liste manuelle et le registre d'adresses. Tout logement supplémentaire répertorié dans le registre d'adresses fait alors l'objet d'une vérification sur le terrain et, si c'est un logement valide, celui-ci est ajouté à la liste du recenseur et un questionnaire de recensement est livré à cette adresse.

L'essai de novembre 1987 se limitait à une comparaison des listes de logements construites selon les méthodes de recensement traditionnelles et, en conséquence, il ne comportait pas de livraison de questionnaires. Pour cet essai, on a embauché des intervieweurs sans expérience que l'on a répartis en équipes chargées d'appliquer strictement la méthode pré-liste ou la méthode post-liste. Le plan de l'essai prévoyait que la construction des deux types de listes serait fait dans les mêmes régions. Les personnes embauchées ignoraient qu'il y avait deux équipes distinctes travaillant dans les mêmes régions. Au moment de l'évaluation, nous allons saisir les listes d'adresses obtenues avec chaque méthode et procéder à un couplage automatise assorti d'une correction des anomalies à partir de vérifications effectuées sur le terrain.

Cet essai a été mené dans les régions métropolitaines de recensement (RMR) de Vancouver, Edmonton, Toronto, Montréal et Halifax. Dans chaque RMR on a tiré un échantillon de 64 secteurs de dénombrement stratifié en fonction du type de logements prédominant d'après les résultats du recensement de 1986. Cet échantillon englobe des régions comptant environ 20,000 logements pour chaque RMR.

Le tableau 2 répertorie les fichiers administratifs utilisés comme sources de données pour construire les registres d'adresses pour chaque RMR. Trois fichiers nationaux, que Statistique Canada possèdait, ont été utilisés pour l'ensemble des villes; il s'agit du fichier de Revenu Canada Impôt (TAX) et des fichiers de Santé et Bien-Être Canada sur les allocations familiales (FAM) et la sécurité de la vieillesse (OAS). En outre, pour chaque ville on a acheté deux listes sélectionnées parmi les registres d'évaluation municipale (MUN), les listes de la compagnie de téléphone (TEL) et celles des compagnies d'électricité (ELE) utilisées pour la facturation. Edmonton constitue une exception du fait qu'à cause du délai requis pour obtenir un des fichiers, le registre d'adresses a été établi à partir de quatre sources seulement.

Tableau 2

Essai de novembre 1987: fichiers sources par RMR

RMR	Fichier source					
	TAX	FAM	OAS	MUN	TEL	ELE
Vancouver	x	x	x	x		x
Edmonton	x	x	x	x		
Toronto	x	x	x	x	x	
Montréal	x	x	x		x	x
Halifax	x	x	x	x	x	

2. ÉTAPES DE LA CONSTRUCTION D'UN REGISTRE D'ADRESSES

Comme nous l'avons souligné précédemment, l'approche utilisée pour construire un registre d'adresses consiste à regrouper les données provenant de multiples sources administratives en une liste sans double compte. Cette approche comprend quatre étapes principales qui sont exposées ci-après.

Standardisation des adresses

Les données relatives à l'adresse contenues dans les fichiers administratifs ont une structure libre, c'est-à-dire que la présentation et l'ordre d'apparition des composantes de l'adresse, comme le nom de rue, le numéro de voirie, le numéro d'appartement, etc. ne suivent pas des règles uniformes. Il est nécessaire d'analyser les données relatives à l'adresse pour en identifier les composantes afin que celle-ci puisse être ré-écrite sous une forme standardisée de façon à simplifier le couplage. Cette tâche s'avère plus complexe qu'on pourrait le penser à prime abord.

Au terme de la recherche relative au registre d'adresses, les études d'évaluation des logiciels actuels de Statistique Canada servant à standardiser les adresses ont révélé des lacunes telles qu'il est apparu nécessaire de revoir complètement ces logiciels afin de pouvoir construire un registre d'adresses. On a adopté une approche très sophistiquée qui comprend plus de cent règles de syntaxe permettant d'identifier les adresses valides (Deguire 1987). Le système divise l'adresse à structure libre en composantes primaires, constitués d'une série de lettres ou de chiffres consécutifs et séparés par des espaces ou des séparateurs tels que la virgule. Certains composantes primaires sont identifiés par le système comme étant des mots clés, par exemple: "Rue", "App.", en français, ou "Street", "Apt", en anglais. Compte tenu des patrons des composantes primaires alphabétiques et numériques et des mots clés connus, nous avons établis qu'il est possible de décoder d'une façon unique et sans ambiguïté les composantes de plus de 95% des adresses. Bien que 52% des adresses puissent être décodées en 8 patrons seulement, le nombre des variantes rencontrées est considérable et il faut plus de 1600 patrons pour décoder 95% des adressses. Actuellement, les 5% restants font l'objet d'un examen et, lorsque cela est possible, d'un décodage manuel. Nous sommes préoccupés de trouver une solution pour décoder cette minorité de cas et une étude est prévue qui vise à déterminer si une amélioration du logiciel ou une modification de la méthodologie de construction du registre d'adresses permettrait de laisser tomber ces cas au lieu de chercher à les décoder manuellement.

Regroupement et élimination des duplicatas

Une fois que l'on a regroupé les adresses standardisées tirées des divers fichiers sources, il faut éliminer les duplicatas, c'est-à-dire les enregistrements multiples

correspondant à une même adresse. Cette étape comprend deux parties: une opérational appariement exact visant à éliminer les duplicata exacts et l'utilisation du jumelage d'enregistrements afin d'éliminer les duplicatas qui diffèrent en totalité ou en partie au niveau d'une ou de plusieurs des composante standardisée. De telles divergences peuvent être dues à divers facteurs tels que les variantes orthographiques, l'utilisation d'abréviations non courantes, etc. Le jumelage l'enregistrements est effectué à l'aide du progiciel de Statistique Canada GIRLS (Hill et Pring-Mill, 1985), fondé sur la méthode mise au point par Fellegi et Sunter (1969).

La section suivante donne une présentation plus détaillée de l'appariement et du jumelage d'enregistrements dans le cadre de la construction des registres d'adresses pour l'essai de novembre 1987.

Codage géographique

Étant donné que nous visons en dernier ressort à construire des listes d'adresses par secteur de dénombrement à partir du registre d'adresses, le couplage des données du registre d'adresses et des codes géographiques utilisés pour le recensement, au niveau du secteur de dénombrement au moins, est essentiel.

Un tel couplage peut se faire de diverses façons qu'il conviendrait d'évaluer. Une méthode pourrait consister à coupler le code postal et le secteur de dénombrement. Cela a déjà été fait, à l'occasion de la saisie des codes postaux, pour un cinquième des logements répertoriés lors du recensement de 1986 et des plans existent en vue d'actualiser ce couplage. On a également prévu des plans pour évaluer l'exactitude du couplage, étant donné que l'utilisation de celui-ci pour lier les adresses d'un registre d'adresses et les secteurs de dénombrement nécessite un degré d'exactitude et de précision de loin supérieur à celui qui est requis pour les utilisations ordinaires.

Contrôle et imputation

La dernière étape de la construction d'un registre d'adresses consiste en une mis au point finale. Il s'agit, par exemple, d'imputer des numéros d'appartement pour combler les lacunes. C'est également l'occasion de repérer manuellement certaines adresses erronées qui ont échappé aux vérifications antérieures et de les éliminer.

3. RÉSULTATS PRÉLIMINAIRES DE LA CONSTRUCTION DE REGISTRES PILOTES

La présente section donne une analyse préliminaire du processus de construction des registres d'adresses, à partir des registres pilotes utilisés pour l'essai de novembre 1987. On pourra procéder à une analyse plus précise et plus détaillée lorsque les résultats de l'essai sur le terrain, actuellement en cours, seront disponibles.

Le tableau 3 indique la couverture brute des registres pilotes à divers stades de leur construction en pourcentage du nombre de logements répertoriés dans les régions sélectionnées lors du recensement de 1986. La colonne 2 donne le nombre initial d'adresses dont le code postal correspond à un des codes postaux des secteurs de dénombrement sélectionnés dans chaque ville selon la plus récente version du fichier de conversion des codes postaux, qui date de février 1987. Il s'agit du nombre d'adresses après le regroupement des adresses standardisées tirées des divers fichiers sources et avant l'élimination des doubles comptes. Les quatre fichiers sources utilisés pour Edmonton contenaient, au total, deux fois plus d'adresses que le nombre répertorié au recensement de 1986 alors que les cinq fichiers sources utilisés pour les autres villes en contenaient en moyenne trois fois plus.

Après l'élimination des duplicatas exactes, la couverture brute (par rapport à celle qui a été obtenue au recensement de 1986) est passée en moyenne de 173% (colonne 1) à 122% (colonne 3). Cela témoigne du succès et de l'importance de l'étape de standardisation des adresses.

Tableau 3

Couverture brute des registres pilotes aux diverses étapes de leur construction, en pourcentages du nombre de logements répertoriés au recensement de 1986

RMR	Après fusion	Après élimination des doubles comptes évidents	Après vérification des codes postaux	Après jumelage d'enregis-trements	Version Final
(1)	(2)	(3)	(4)	(5)	(6)
Vancouver	283	117	109	103	104
Edmonton	194	110	103	99	101
Toronto	283	113	103	102	102
Montréal	312	136	125	111	108
Halifax	297	134	126	109	110
Average	273	122	113	105	105

La colonne 4 est une étape particulière à la construction des registres d'adresses pilotes. Le code postal de chaque adresse a été vérifié au moyen d'un logiciel mis au point par Statistique Canada à cette fin; les adresses dont le code postal était erroné et celles dont le code, après correction, n'était pas situé à l'intérieur d'un secteur de dénombrement faisant partie de l'échantillon ont été rejetées. Soulignons que dans le cadre de la construction d'un registre d'adresses intégral, ces adresses ne seraient pas rejetées mais attribuées au secteur de dénombrement correspondant. À la suite de la vérification des codes postaux, 9% des enregistrements ont été rejetés et, bien entendu, aucun n'a été ajouté; cela constitue une source de sous-dénombrement qui est propre uniquement aux registres pilotes. Nous envisageons de mesurer l'ampleur de ce sous-dénombrement qui risque de varier d'un très faible niveau à un niveau assez significatif, selon que les codes postaux utilisés dans les divers fichiers sont plus ou moins interdépendants.

Le jumelage d'enregistrements a permis de réduire encore de 8% (colonne 5) la couverture brute, ce qui donne une couverture brute moyenne de 105%. La colonne 6 donne la couverture brute après le contrôle et l'imputation. En moyenne, la couverture brute n'a pas été modifiée mais, si l'on considère la couverture de chaque RMR, celle-ci a augmenté ou baissé de 1 à 2%, ce qui est considérable compte tenu du taux de sous-dénombrement net prévu. Si les résultats sont analogues à ceux qu'a donnés un registre pilote antérieur construit pour Ottawa, le sous-dénombrement net par rapport aux chiffres du recensement pourrait être proche de 1% ce qui donne, compte tenu du surdénombrement brut des registres qui est de 5%, un surdénombrement net de 6%. Le surdénombrement est dû aux duplicatas qui n'ont pas été repérées lors du jumelage d'enregistrements ou à la présence de logements qui ne sont plus valides.

Les résultats de l'essai sur le terrain permettront d'établir le surdénombrement et le sous-dénombrement non seulement pour le registre d'adresses proprement dit mais également du point de vue des utilisations qui en seront faites dans le cadre de la collecte des données du recensement. Nous envisageons aussi d'effectuer des études détaillées

pour déterminer les raisons pour lesquelles des addresses ont été oubliées dans le registre et établir s'il est possible d'améliorer les méthodes ou les progiciels utilisés de façon à réduire le sous-dénombrement.

Le tableau 4 présente quelques résultats du jumelage d'enregistrements dans le cadre de la construction du registre d'adresses. Il donne, pour les paires d'enregistrements couplés, le pourcentage de fois où les composantes de l'adresse servant à coupler les données concordaient entièrement, partiellement ou pas du tout. Il convient de souligner que le numéro de voirie était un champ de regroupement dans le cadre de ce jumelage; en effet, les autres liens n'étaient recherchés que pour les enregistrements dont les numéros de voirie coïncidaient. Notons également que lorsqu'il y avait regroupement et appariement exact, il était fait état des fichiers dans lesquels l'adresse figurait et, au moment du jumelage d'enregistrements, on retenait la version qui apparaissait dans le plus grand nombre de fichiers. Deux degrés de concordance étaient admis pour la comparaison des noms de rue et de municipalité. (Ceux-ci sont combinés dans le tableau 4.) Le premier degré correspondait à une divergence mineure dans l'épellation due à l'omission d'une lettre ou à la transposition de deux lettres. Deux noms étaient considérés comme satisfaisant au second degré de concordance partielle lorsque leur codage phonétique à l'aide du NYSIIS (New York State Identification and Intelligence System) était identique. Ce système de codage vise à éliminer les effets des erreurs d'orthographe les plus courantes. Un autre type de concordance partielle était admis, au niveau des trois derniers caractères du code postal: si deux des trois caractères étaient identiques, on considérait qu'il y avait concordance partielle.

Tableau 4

Résultats de la comparaison des paires d'adresses appariés lors du jumelage d'enregistrements (en pourcentage)

Catégorie	Résultats			
	Concordance parfaite	Concordance partielle	Non concordance	Composante manquante
Nom de rue	49	31	20	
Numéro d'app.	93		7	
Suffixe du no de voirie	95		5	
Code postal: caractères 1 à 3	100			
Code postal: caractères 4 à 6	95	4	1	
Municipalité	87	2	11	
Nom de famille	35		18	47

Il est intéressant de souligner le peu de concordance des noms de rue relevé lors du jumelage d'enregistrements, une concordance parfaite n'étant obtenue que dans la moitié des cas. Cela semble dû à la fréquence des erreurs d'orthographe et à l'utilisation d'abréviations. Un autre élément mérite une attention particulière, c'est l'utilisation du nom de famille comme variable de couplage. Celle-ci a été utilisée strictement pour les besoins du jumelage d'enregistrements et a été éliminée du registre final. Du fait que les données des fichiers sources dataient plus ou moins, la non-concordance des noms de famille n'empêchait pas systématiquement le couplage de deux adresses; néanmoins, la concordance des noms de famille était considérée comme un résultat important qui était donné un poids très élevé. Afin d'évaluer l'importance du nom de famille comme variable de couplage, pour une ville en particulier on a refait le couplage des enregistrements sans utiliser le nom de famille et le nombre d'adresses appariées s'est trouvé réduit de 1%.

Les deux tableaux ci-après illustrent la contribution des divers fichiers sources au registre d'adresses final. Le tableau 5 donne la couverture de chaque fichier source par rapport à la couverture brute du registre d'adresses, c'est-à-dire le pourcentage des adresses du registre qui ont été repérées dans chaque fichier source. Il confirme ce que nous soupçonnions, à savoir que la couverture des fichiers de Revenu Canada Impôt, des compagnies de téléphone et d'électricité est très importante. Les fichiers des compagnies d'électricité sont les meilleurs et, du moins pour les deux provinces sur lesquelles nos observations ont porté, il semble que l'existence d'un compteur unique dans les immeubles à logements multiples, qui était auparavant la faiblesse de cette source, ne constitue plus un inconvénient sérieux. La faible couverture des fichiers de Revenu Canada Impôt pour Montréal et Toronto est attribuable au nombre important d'erreurs relevées dans les adresses de ces fichiers en raison duquel ces adresses n'ont pas été retenues. Exception faite de Toronto, la couverture des fichiers d'évaluation municipale est généralement faible du fait qu'il n'y a habituellement qu'un seul enregistrement pour chaque propriétaire d'immeubles à logements multiples.

Tableau 5

Couverture brute des fichiers sources
(en pourcentage de la couverture brute de registre d'adresses)

Ville	Fichier source					
	TAX	FAM	OAS	MUN	TEL	ELE
Vancouver	73	26	26	48		87
Edmonton	82	32	18	49		
Toronto	60	22	18	78	76	
Montréal	57	24	16		72	86
Halifax	78	30	19	47	72	

Le tableau 6 donne le pourcentage d'adresses provenant d'une seule source. Là encore, les fichiers des compagnies d'électricité se classent remarquablement bien, suivis de près par ceux de la compagnie de téléphone. Les fichiers de Revenu Canada Impôt pour Halifax et Edmonton se sont également bien classés. Dans le cas d'Edmonton, le résultat est une anomalie étant donné que, parmi les quatre fichiers utilisés, celui de Revenu Canada Impôt est le seul à avoir une couverture importante.

Il faut également souligner que les résultats du tableau correspondent à la contribution du fichier par rapport à la couverture brute. Il sera intéressant, une fois que les résultats de l'essai de novembre 1987 seront disponibles, de calculer la contribution de chaque fichier par rapport à la couverture nette. L'utilité de fichiers tels que celui des allocations familiales ou de la sécurité de la vieillesse risquerait d'être remise en question si une proportion importante des 0.5% à 1% des adresses dont ils sont la source unique s'avérait erronée.

Tableau 6

Adresses provenant d'un seul fichier source
(en pourcentage de la couverture brute de registre d'adresses)

Ville	Fichier source					
	TAX	FAM	OAS	MUN	TEL	ELE
Vancouver	5	1	1	1		13
Edmonton	28	5	4	4		
Toronto	2	0.5	0.5	7	12	
Montréal	3	0.5	1		9	17
Halifax	10	1	1	2	9	

Perspectives

L'analyse des résultats de l'essai de novembre 1987 sera terminée au printemps de 1988. On disposera alors également d'une estimation des besoins de construction ainsi que du coût et des délais pour différents scénarios concernant l'utilisation d'un registre d'adresses dans le recensement de 1991. Il est prévu qu'il sera décidé au printemps de 1988, à partir de cette double source de renseignements, de la mesure dans laquelle le registre des adresses sera utilisé au recensement de 1991. Si l'on convient d'utiliser le registre des adresses à une grande échelle, il faudra alors accorder une forte priorité aux travaux de construction en vue de 1991.

Les études effectués jusqu'à présent ont mis en évidence les domaines devant faire l'objet de recherches plus approfondies, dont certaines devront être menées parallèlement aux travaux de construction s'il était décidé d'utiliser le registre d'adresses en 1991. Ces recherches devront également être poursuivies si le registre d'adresses devait être utilisé dans le cadre d'un essai et non de la production courante, en 1991. Quelques domaines devant faire l'objet de recherches approfondies sont présentés ci-après.

Mise à jour

Pour le moment, nous avons seulement envisagé l'établissement initial d'un registre des adresses. Les sources et les méthodes qui conviennent le mieux au stade initial ne sont pas nécessairement les mieux appropriées lorsqu'il s'agit de mettre à jour le registre. Il faut tenir compte de la fréquence à laquelle la mise à jour doit être effectuée et de l'incidence sur les systèmes de la fréquence et de l'importance des mises à jour. Une approche possible pourrait être d'effectuer un couplage détaillé des versions successives des fichiers sources afin de mettre en évidence les modifications, ce qui permettrait de les relier au registre d'adresses existant. La façon de procéder pour éliminer des adresses qui ne sont plus valides doit faire l'objet d'une étude distincte. On pourrait également utiliser d'autres sources de données comme les permis de construction ou de démolition ou les mises à jour effectuées par Postes Canada.

Utilisation du registre d'adresses pour délimiter les secteurs de dénombrement

Pour les besoins de la collecte et de la diffusion des données, les secteurs de dénombrement du recensement doivent compter approximativement le même nombre de logements et respecter les limites géostatistiques et géopolitiques de niveau supérieur. Actuellement, afin de délimiter les secteurs de dénombrement, on se fonde essentiellement pour déterminer le nombre de logements sur les chiffres du dernier recensement et il arrive parfois que ceux-ci ne correspondent pas à la réalité. En utilisant les chiffres d'un registre d'adresses mis à jour pendant la période intercensale, on

pourrait améliorer la délimitation des secteurs de dénombrement et réduire le coût et les délais imposés par la nécessité de fractionner des secteurs de dénombrement parce qu'au cours des opérations de recensement sur le terrain on découvre que le secteur a connu une croissance substantielle.

Utilisation d'un registre d'adresses comme base de sondage pour les enquêtes-ménages

Actuellement, la plupart des enquêtes-ménages de Statistique Canada sont basées sur des échantillons aérolaires, ce qui requiert des interviews sur place coûteuses, du moins le premier mois où les ménages sont introduits dans l'échantillon. Les listes de téléphone ne constituent pas en elles-mêmes une solution alternative intéressante pour les grandes enquêtes nationales, en raison du biais associé au sous-dénombrement de l'univers des non-abonnés (Drew et Jaworski 1986). La solution qui consisterait à utiliser une double méthode combinant l'échantillon aérolaire et la liste téléphonique est plutôt inefficace puisqu'il faut avoir recours à un échantillon aérolaire important pour couvrir une petite population de non-abonnés. Un registre d'adresses contenant les numéros de téléphone d'environ 75% des ménages (voir le tableau 5) pourrait constituer une base intéressante offrant l'avantage de pouvoir interviewer par téléphone une proportion importante de la population tout en permettant d'adopter un plan de sondage efficace pour les ménages ruraux et urbains restants.

Des plans prévoient la conversion d'une portion de l'échantillon de l'enquête sur la population active en une base de sondage reposant sur un registre d'adresses dans les secteurs où des registres pilotes sont mis au point en vue de la délimitation des secteurs de dénombrement. Dans le cadre de l'essai, on doit mettre en application des méthodes susceptibles de régler le problème du sous-dénombrement du registre d'adresses.

Perfectionnement de la méthodologie de la construction d'un registre d'adresses

Enfin, des recherches doivent être entreprises concernant le processus de construction du registre d'adresses. Il faut étudier notamment des aspects tels que l'incidence de l'emploi d'un nombre plus grand de fichiers sources ou de fichiers sources différents. Est-il possible de trouver d'autres sources offrant une couverture importante et, le cas échéant, quelle serait l'incidence de leur emploi pour la construction d'un registre d'adresses?

En outre, nous avons vu que le progiciel servant à standardiser les adresses et à valider les codes postaux ne permet pas de traiter tous les cas possibles. Les cas problèmes doivent faire l'objet d'une recherche plus approfondie. S'agit-il d'erreurs dans l'adresse qui apparaissent dans un fichier alors que d'autres fichiers contiennent une adresse valide pour le même logement? Si telle était la réalité, on pourrait décider d'ignorer les cas problèmes qui se présentent dans des fichiers particuliers. Par contre, si les cas non résolus par le progiciel sont dus à une lacune du système qui l'empêche de traiter des adresses pourtant valides ou encore si certains types d'adresses tendent à être erronées dans tous les fichiers, le fait d'ignorer ces cas entraînerait des problèmes de couverture. Les cas problèmes, y compris les adresses manquantes dans les registres pilotes, devront faire l'objet d'une étude détaillée pour que l'on ait une réponse à ces questions.

En résumé, les résultats obtenus jusqu'à présent sont encourageants, tant du point de vue de la faisabilité technique de la construction d'un registre offrant une couverture élevée des adresses à un coût raisonnable que de la possibilité, avec un tel registre, de réduire le sous-dénombrement au recensement. Pour la prochaine année, il est prévu de poursuivre les recherches dans un certain nombre de directions en vue d'appliquer ou d'améliorer les méthodes de construction et de mise à jour du registre d'adresses en les intégrant aux travaux de développement qui seront entrepris s'il est décidé d'utiliser ce type de registre au recensement de 1991.

BIBLIOGRAPHIE

Booth, J.K. (1976). A summary report of all address register studies to date. Rapport interne, Statistique Canada.

Deguire, Y. (1987). Research into the parsing and standardization of free format addresses at Statistics Canada. Rapport interne, Statistique Canada.

Drew, J.D., Armstrong, J., et Dibbs, R. (1987). Research into a register of residential addresses for urban areas of Canada. *American Statistical Association, Proceedings of the Section on Survey Research Methods.*

Drew, J.D., et Jaworski, R. (1986). Telephone survey development on the Canadian Labour Force survey. *American Statistical Association, Proceedings of the Section on Survey Research Methods.*

Fellegi, I.P., et Krotki, K.P. (1987). The testing programme for the 1971 census in Canada. *American Statistical Association, Proceedings of the Social Statistics Section*, 29-38.

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

Gamache-O'Leary, V., Nieman, L., et Dibbs, R. (1987). Cost implications of mail-out of census questionnaires using an address register. Rapport interne, Statistique Canada.

Hill, T., et Pring-Mill, F. (1985). Generalized Iterative Record Linkage System. *Proceedings of the Workshop on Exact Matching Methodologies.* Arlington, Virginie, 327-333.

Redfern, P. (1987). European experience of using administrative data for censuses of population: The policy issues that must be addressed. Article présenté lors du Symposium international sur les utilisations statistiques des données administratives, Ottawa.

Royce, D. (1986). Address register research for the 1991 Census of Canada. *Journal of Official Statistics*, 2, 447-456.

Whitford, D. (1987). Research Program for the 1990 Decennial Census. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 447-452.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

LES UTILISATIONS MULTIPLES DES DOSSIERS ADMINISTRATIFS DANS L'ANALYSE DES DONNÉES SUR L'ÉDUCATION

CHARLES D. COWAN et MARY K. BATCHER[1]

RÉSUMÉ

Les données sur l'éducation posent un problème analytique particulier. Les données disponibles sont souvent hiérarchiques, ce qui signifie qu'il y a plusieurs niveaux auxquels on peut recueillir, compiler et analyser les données. Bien que ce cas ne soit pas propre au domaine de l'enseignement, ce qui est unique à ce dernier, c'est la variété des données disponibles et les possibilités d'intégration horizontale et verticale des données d'enquêtes et administratives. La première partie du document, traite des utilisations actuelles des données administratives par le "Center for Education Statistics". Aux niveaux élémentaire et secondaire, les données administratives sur les districts scolaires, les écoles publiques et les enseignants des écoles publiques sont recueillies et compilées pour chaque état et l'ensemble du pays. L'une des utilisations de ces données est le rassemblement et la présentation de statistiques sommaires. Une deuxième utilisation des données consiste en une analyse dans le temps afin de déterminer les tendances de l'existence et de l'utilisation des ressources.

La deuxième partie du document revoit les utilisations des données administratives pour deux autres utilisations. Une troisième utilisation de ces données sera de fournir des totaux de population pour pondérer l'enquête auprès des écoles et du personnel enseignant, qui est une grande enquête menée auprès des districts scolaires, des établissements d'enseignements et des enseignants, et qui sera exécutée à compter de 1988 pour le "Center for Education Statistics". Une quatrième utilisation des données sera de créer et de valider des modèles de prévision pour les inscriptions, la demande de personnel enseignant au niveau élémentaire et au niveau secondaire, les ratios enseignants/élèves et les taux de promotion de fin d'études secondaires par état.

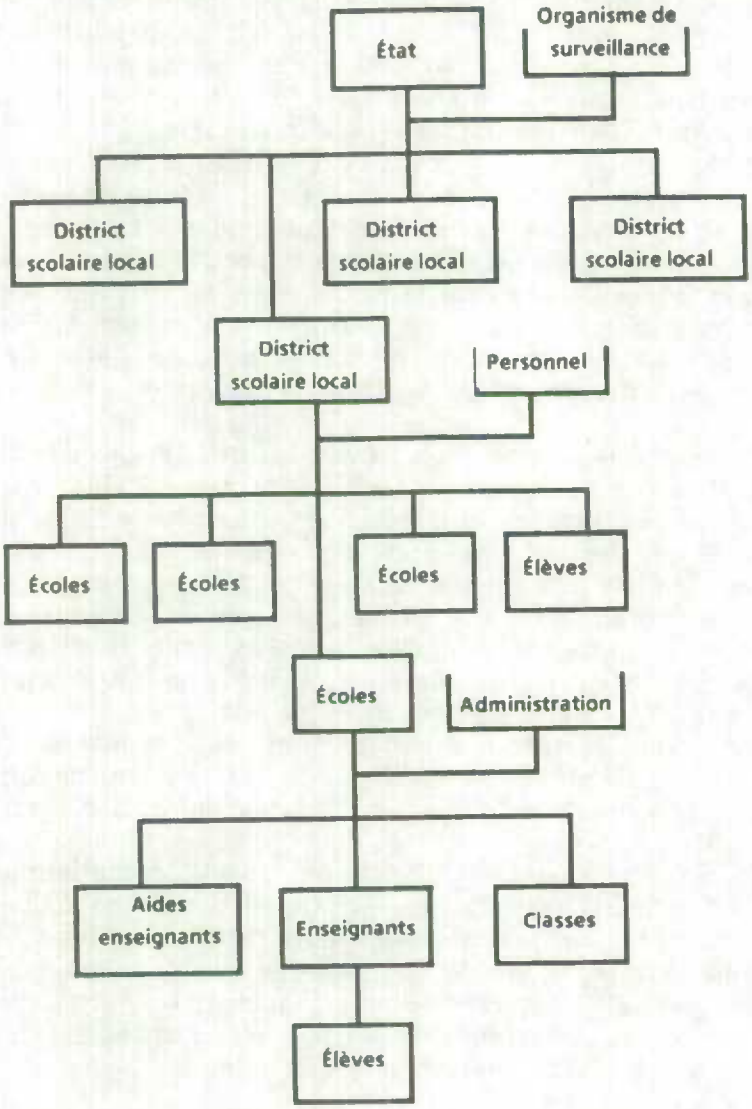
On examinera également les utilisations parallèles de données des dossiers administratifs pour l'enseignement post secondaire telles que les données sur les dossiers scolaires des collèges et des universités. L'intégration de ces données aux données d'enquêtes recueillies par le "Center for Education Statistics" contribuera à réduire le biais et la variance des estimations des tendances et des niveaux du système d'enseignement américain.

[1] Charles D. Cowan et Mary K. Batcher, Center for Education Statistics, 555 New Jersey Avenue North West, Salle 400, Washington, D.C. 20208. U.S.A.

1. INTRODUCTION

Les données sur l'éducation sont différentes de celles qui sont recueillies auprès de l'ensemble de la population du fait qu'elles proviennent de plusieurs niveaux, chacun de ces niveaux représentant des unités administratives qui font l'objet d'une étude ou qui fournissent des renseignements sur les niveaux inférieurs. Le graphique 1 montre les liens qui existent entre les divers niveaux d'établissement de dossiers administratifs du système d'enseignement américain. Ces liens donnent lieu à des problèmes particuliers concernant la collecte, la validation et l'analyse des données administratives. Le présent document décrit la méthode de collecte des données sur l'éducation et l'utilisation de ces données dans les systèmes de dossiers administratifs, ainsi que les questions plus générales liées au couplage des données administratives avec les données d'enquête et de recensement recueillies par le CES (Center for Education Statistics/Centre des statistiques de l'éducation).

GRAPHIQUE: NIVEAUX DES DONNÉES SUR L'ÉDUCATION



2. SYSTÈMES DE DOSSIERS ADMINISTRATIFS

À l'heure actuelle, le CES compte deux grands systèmes de dossiers administratifs: le CCD ou "Common Core of data" (ensemble de données communes) qui porte sur l'enseignement primaire et secondaire, et l'IPEDS ou "Integrated Postsecondary Education Data System" (système intégré de données sur l'enseignement postsecondaire) qui concerne l'enseignement postsecondaire aux États-Unis. Ces deux systèmes sont décrits dans le présent chapitre. Le reste du document portera sur les utilisations des dossiers administratifs à des fins d'estimation et d'analyse.

2.1 **Common Core of Data** (CCD/Ensemble de données communes)

Le CCD est constitué des données d'une série interdépendante d'enquêtes effectuées par les organismes chargés de l'éducation des cinquante États américains, du District de Columbia et des régions éloignées relevant des États-Unis. Ces enquêtes fournissent au CES des renseignements de base sur les quelque 85,000 écoles publiques américaines et leurs 40,000 millions d'élèves, sur plus de 2 millions d'enseignants ainsi que sur les 4 millions de personnes formant l'ensemble du personnel. Elles servent aussi à recueillir des données détaillées sur les dépenses annuelles de $137 milliards de dollars des écoles publiques américaines. La collecte des données dans le cadre du CCD a lieu chaque année et ce, depuis l'automne 1977.

Bien que le ministère de l'Éducation des États-Unis ait recueilli des données de base sur les écoles publiques américaines depuis sa création, le prédécesseur immédiat du CCD était l'enquête ELSEGIS (Elementary and Secondary General Information Survey/enquête générale sur l'enseignement primaire et secondaire) mise en oeuvre en 1967. L'ELSEGIS servait à compléter les renseignements sur les programmes obtenus des systèmes de données sur les programmes fédéraux à l'aide de données générales sur les systèmes scolaires et les élèves. Cette enquête permettait d'obtenir des données sur le nombre d'écoles, leurs niveaux d'enseignement, les inscriptions et le rapport enseignant/élèves dans les systèmes scolaires locaux. Elle servait aussi à recueillir des données sur l'univers des écoles, y compris le nom des écoles, le nombre d'inscriptions et le nombre d'enseignants. Les éléments d'information auxquels l'ELSEGIS s'intéressait sont semblables à ceux que l'on retrouve dans le CCD.

La collecte des données pour le CCD a débuté au cours de l'année scolaire 1977-1978, après une mise à l'essai en 1976. Actuellement, le CCD est soumis à un processus d'évaluation et de révision de trois ans, dans le cadre d'un effort conjoint de l'administration fédérale et des États visant à améliorer la base de données.

Le CCD regroupe les données de quatre enquêtes: la State Nonfiscal Survey (enquête servant à la collecte de données non financières menée auprès des États), la Survey of Revenues and Current Expenditures for Public Elementary and Secondary Education (enquête sur les revenus et les dépenses courantes se rapportant à l'enseignement primaire et secondaire), la Public Elementary/Secondary Education Agency Universe Survey (enquête sur les écoles publiques primaires et secondaires menée auprès des organismes responsables de l'éducation et la Public Elementary/Secondary School Universe Survey (enquête sur les écoles publiques primaires et secondaires). Toutes les enquêtes du CCD ont pour objet de recueillir des données administratives auprès des organismes de l'éducation des États. La State Nonfiscal Survey permet la collecte de données sur le nombre d'élèves par niveau, sur le nombre d'enseignants, d'administrateurs, de conseillers en orientation, de bibliothécaires et d'autres membres du personnel, ainsi que sur le nombre de diplômés d'écoles secondaires. La Survey of Revenues and Current Expenditures for Public Elementary and Secondary Education sert à rassembler des données sur les revenus et les dépenses courantes ainsi que sur les composantes utilisées pour calculer la fréquentation scolaire quotidienne moyenne dans les écoles primaires et

secondaires des États-Unis. Les données recueillies dans le cadre de cette enquête portent sur l'année scolaire écoulée et sont donc en retard d'une année sur les données des autres enquêtes du CCD. Toutes les données de la State Nonfiscal Survey et de la Survey of Revenues and Current Expenditures sont déclarées au CES pour l'ensemble d'un État.

Les données recueillies au moyen des deux autres enquêtes du CCD sont moins regroupées. La Public Elementary/Secondary Education Agency Universe Survey fournit des données comme le nom et l'adresse des écoles, le genre d'organisme, le genre de région desservie, le nombre d'écoles, le nombre de niveaux scolaires offerts et le nombre d'élèves, pour chaque district scolaire et par État. Le Public Elementary/Secondary School Universe Survey fournit des données comme le nom et l'adresse des écoles, le genre d'école et le nombre d'élèves et d'enseignants, et ce pour toutes les écoles publiques de chaque État.

En août 1985, le CES a financé un projet de trois ans et l'a mis en oeuvre avec l'aide du Council of Chief State School Officers (Conseil des principaux administrateurs des écoles des États). Ce projet vise à décrire l'état actuel des données du CCD, à étudier les ajouts possibles de données afin d'enrichir la base de données et de faire en sorte qu'elle puisse fournir une vue d'ensemble appropriée du système scolaire au pays, et à permettre l'établissement de recommandations à l'intention des États et du CES pour rendre la base de données plus complète, comparable et à jour. Le CES a maintenant pris connaissance des recommandations qui ont été faites et des décisions ont été prises concernant toutes les questions non financières à part la dotation en personnel. Les changements visant la collecte de données non financières, la dotation mise à part, seront mis en application pour la collecte de 1987-1988. Le CES recevra et évaluera les recommandations relatives à la dotation et aux questions financières, et tout changement proposé sera appliqué au cours des deux prochaines années (années scolaires 1988-1989 et 1989-1990).

Les nouvelles données non financières qui doivent être recueillies en 1987-1988 comprennent les numéros de téléphone des écoles et des districts scolaires, le genre de région/secteur où se situe l'école, le nombre d'élèves par école d'après la race et l'admissibilité au programme de repas gratuits. Les chiffres fournis au niveau des districts scolaires permettront d'avoir des données plus détaillées sur les effectifs scolaires et d'établir, pour la première fois à ce niveau, le nombre de diplômés d'écoles secondaires.

D'autres mesures ont été prises pour améliorer la base de données. On a notamment procédé à l'uniformisation des définitions à travers le pays et les personnes qui fournissent les données, c'est-à-dire les coordonnateurs du CCD de chaque État, ont reçu une formation plus poussée. L'uniformisation des définitions a interrompu la continuité de la série chronologique. Ces ruptures résultent de deux objectifs qui s'opposent: 1) le besoin d'uniformiser les définitions et les méthodes de collecte pour assurer la comparabilité des données d'un État à un autre, et 2) le besoin de fournir des données comparables dans le temps. Lorsque les révisions seront terminées et que quelques années se seront écoulées, nous espérons pouvoir satisfaire ces deux objectifs en même temps. Toutefois, d'ici à ce que la série chronologique se stabilise, le CES devra établir des correspondances entre les anciennes et les nouvelles données lorsque ce sera possible, et fournir les explications requises lorsqu'il n'est pas possible d'établir une correspondance exacte.

Au cours des années précédentes, les enquêtes du CCD étaient effectuées de façon indépendante, les divers questionnaires étant envoyés par la poste à des périodes différentes et devant être renvoyés à des dates différentes. À compter de 1987-1988, les documents d'enquête du CCD seront groupés en une seule trousse qui sera envoyée par la poste à la fin de décembre et devra être renvoyée pour le 15 mars.

Dans le cas des données non financières, les rapports provisoires seront publiés en juin pour l'année scolaire prenant fin et contiendront les données déclarées au 15 mai. Le CES acceptera les mises à jour faites par les États jusqu'au 15 septembre et publiera en

octobre les données finales se rapportant à l'année scolaire écoulée. Les fichiers de données non financières feront l'objet d'une seule et dernière série de mises à jour en septembre de l'année suivante, et ne pourront plus être modifiés par la suite. Les données financières finales seront aussi publiées en octobre, mais une année après les données non financières. Une fois que les fichiers de données non financières seront fermés en septembre, les données finales révisées de l'année précédente seront mises sur bandes qu'il sera possible de se procurer. Il sera aussi possible d'obtenir les bandes des données financières, mais les périodes de référence seront légèrement différentes. En ce qui a trait aux bandes de données agrégées d'un État, elles regrouperont les données de plusieurs années.

2.2 Integrated Postsecondary Education Data System
(IPEDS/système intégré de données sur l'enseignement postsecondaire)

L'IPEDS consiste en un groupe d'enquêtes visant à recueillir des données auprès de tous les principaux établissements d'enseignement postsecondaire, notamment sur les caractéristiques de l'établissement, les inscriptions, le nombre de diplômés, les aspects financiers, le personnel et leur rémunération. L'IPEDS est censé donner un aperçu de l'évolution de l'enseignement postsecondaire aux États-Unis. À cette fin, le système doit décrire la situation de l'enseignement postsecondaire et tenir compte des changements touchant l'importance de ce niveau d'enseignement, ses caractéristiques générales, ses établissements et ses élèves.

Par le passé, le CES obtenait les renseignements voulus au moyen de trois grandes enquêtes: HEGIS (Higher Eduction General Information Survey/enquête permettant de recueillir des données générales sur l'enseignement supérieur), VEDS (Vocational Education Data System/système de données sur l'enseignement professionnel) et la Survey of Noncollegiate Postsecondary Schools with Occupational Programs (enquête sur les établissements postsecondaires qui ne sont pas des collèges et qui ont des programmes de formation professionnelle). Les données de ces enquêtes étaient complétées par les résultats d'"études spéciales", y compris d'enquêtes périodiques HEGIS (par ex., des enquêtes portant sur le lieu de résidence et les migrations des étudiants de niveau collégial, ou sur les bibliothèques des collèges et des universités), d'une enquête-échantillon sur les diplômés collégiaux des dernières années, d'enquêtes à court délai de réponse, et d'enquêtes menées conjointement avec d'autres organismes fédéraux. Par suite de l'élaboration et de la mise en oeuvre de ces enquêtes, le CES a relevé plusieurs problèmes d'ordre méthodologique. Par exemple, étant donné que les établissements visés par les enquêtes HEGIS et VEDS étaient en grande partie les mêmes, leur fardeau de réponse s'en trouvait accru. De plus, le CES n'était même pas en mesure de faire la synthèse des renseignements additionnels ainsi obtenus à cause de différences dans les définitions, dans les méthodes d'enquête, et ainsi de suite. Enfin, dans le cadre de la collecte de ces données, on n'a ni tenu compte, ni même constaté l'existence de plusieurs segments du vaste univers des établissements d'enseignement postsecondaire. Par conséquent, ce programme d'enquêtes n'a pas réussi à présenter une description complète du système d'enseignement postsecondaire.

Compte tenu de ces problèmes et des recommandations faites par les responsables des établissements d'enseignement postsecondaire, le CES a entrepris pendant trois années de mettre au point le système intégré de données sur l'enseignement postsecondaire (IPEDS). Ce système englobe tous les établissements d'enseignement postsecondaire et fournit une description complète et juste du domaine de l'enseignement postsecondaire. Tout au long de l'élaboration du système IPEDS, on a pris soin de s'assurer qu'il serait possible d'atteindre les objectifs suivants:

- Éliminer les redondances et le chevauchement des opérations de collecte de données sur l'enseignement postsecondaire;

- minimiser le fardeau de réponse;
- assurer la comparabilité des données entre les différents secteurs d'enseignement postsecondaire;
- tenir compte des facteurs uniques propres aux divers secteurs de l'enseignement postsecondaire;
- obtenir des statistiques valables et sûres auprès des établissements d'enseignement postsecondaire.

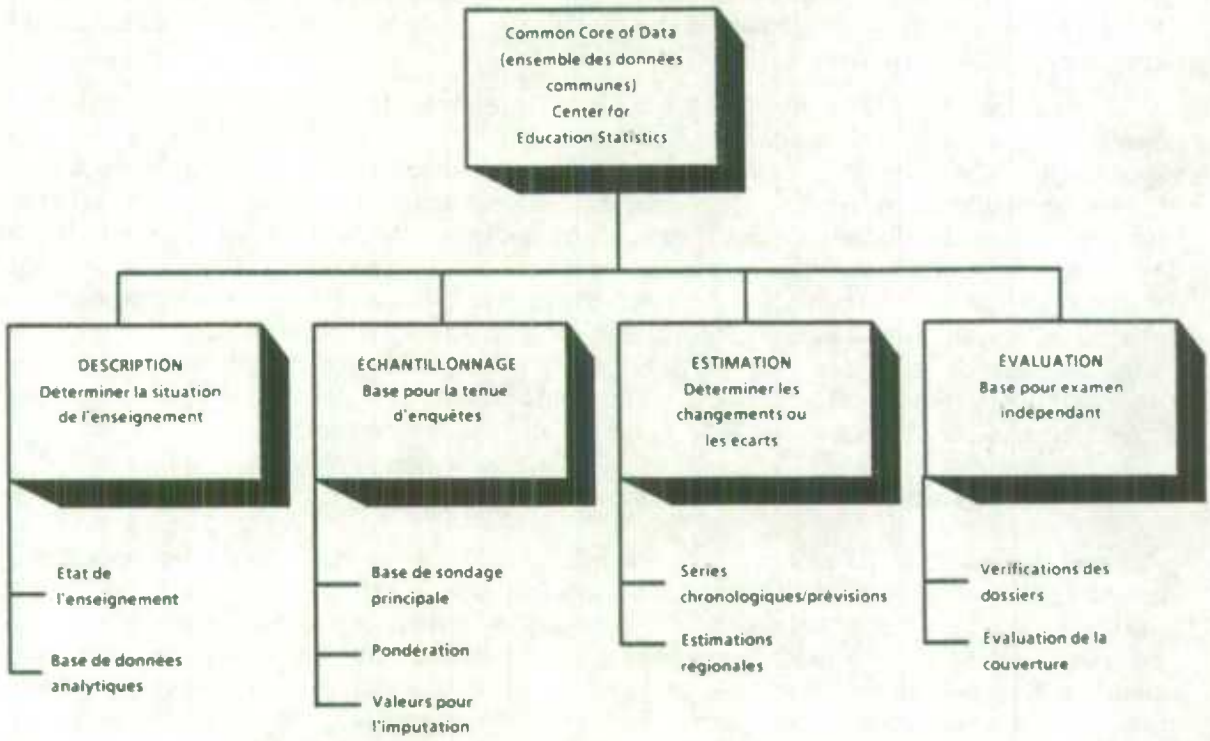
3. UTILISATIONS DES DOSSIERS ADMINISTRATIFS ET PARTICULIÈREMENT DU CCD

Le reste du présent document sera consacré à la description des nombreuses utilisations possibles des dossiers administratifs relativement aux fonctions d'information d'un organisme statistique. Pour illustrer ces fonctions, nous examinerons surtout les utilisations du CCD. Pour un organisme statistique fédéral, les utilisations des dossiers administratifs peuvent être réparties en huit grandes catégories:

1) Les données servent à résumer et à décrire la répartition actuelle des ressources et la situation de l'enseignement aux États-Unis.
2) Les données s'inscrivent dans une série chronologique et servent à décrire les tendances et à déterminer certains indicateurs de base pour prévoir l'évolution de l'enseignement au cours de l'année à venir ou des cinq prochaines années.
3) Les données peuvent servir de base de sondage pour la stratification ou la sélection d'écoles ou de districts avec probabilité proportionnelle à une certaine mesure de la taille qui peut être trouvée dans les dossiers administratifs.
4) Les données peuvent être utilisées pour fournir des totaux et des limites qui formeront la base de la pondération de l'échantillon afin de réduire la variance et les biais attribuables à la sélection dans les enquêtes du CES.
5) Les données peuvent servir à faire des estimations régionales des facteurs échantillonnés.
6) Les données peuvent servir à l'imputation des données manquantes des enquêtes-échantillon.
7) Les autres chercheurs ont accès à la base de données pour obtenir les renseignements dont ils ont besoin pour effectuer leurs analyses.
8) Les données peuvent servir à évaluer les résultats d'enquêtes et de recensements du point de vue de la couverture et de la validité des réponses.

Le tableau 2 montre les diverses utilisations des dossiers administratifs.

GRAPHIQUE 2: UTILISATIONS DES DOSSIERS ADMINISTRATIFS DANS L'ANALYSE DES DONNEES SUR L'ÉDUCATION



3.1 Rapport sur la situation actuelle au chapitre de l'enseignement

Les données du CCD servent avant tout à décrire la situation de l'enseignement public au niveau primaire et secondaire aux États-Unis. Elles constituent une source précieuse d'information pour les publications du CES. Le CCD fournit une grande partie des données requises pour les deux principales publications annuelles du CES, le Digest of Education Statistics et Condition of Education, ainsi que pour plusieurs rapports et autres publications spéciales.

En plus de faire état de la situation dans le domaine de l'enseignement, les données du CCD sont les principaux renseignements utilisés pour établir les formules de répartition de certains programmes fédéraux. Les dépenses par élève prévues pour chaque État sont calculées à l'aide des données du Survey of Revenues and Current Expenditures for Public Elementary and Secondary Education qui sont fournies au CES. Certaines sommes versées en vertu de divers programmes, dont le chapitre 1 de la Education and Consolidation Act of 1981, Impact Aid, et Indian Education, sont déterminées en partie d'après les dépenses qu'un État prévoit faire par élève.

3.2 Modèles de prévision

Le CCD sert de base à l'établissement des projections annuelles du CES concernant les inscriptions, la demande d'enseignants aux niveaux primaire et secondaire, le rapport enseignant/élèves et les proportions de diplômés d'écoles secondaires. Récemment, dans le domaine de l'enseignement primaire et secondaire, des prévisions ont pu être faites au sujet des inscriptions et du nombre d'enseignants par État. Ces modèles sont plus complexes que les estimations nationales traditionnelles en raison des mouvements démographiques et des différences observées entre les États dans certains des taux de

base relatifs à la persévérance scolaire. Les premières projections produites cette année ont été fondées sur les données du CCD des dix dernières années par État, les données sommaires du Bureau of the Census au niveau des États, les prévisions de la population âgée de 4 à 17 ans provenant aussi du Bureau of the Census et établies par État et enfin, sur le calcul des taux de persévérance scolaire, y compris les taux d'admissions et de départs des écoles privées.

Un avantage imprévu résultant de l'établissement de modèles de prévision a été la possibilité d'utiliser les modèles pour chaque État en vue de valider les résultats des rapports les plus récents sur les inscriptions de chaque État. Au cours de l'élaboration de la série chronologique, les rapports provisoires provenant des États ont été utilisés comme sources finales de données. Toutefois, dans certains États, il était évident que certains des effectifs scolaires déclarés dépassaient les limites prévues. On a communiqué avec les responsables de l'information de neuf États afin de vérifier les renseignements reçus et dans deux cas, des données additionnelles ont été fournies, consolidant ainsi le modèle. Dans les autres cas, les écarts par rapport aux prévisions ont été attribués à des changements apportés aux méthodes de déclaration de l'État ou à l'adoption par l'État d'une nouvelle loi ayant eu des répercussions sur l'enseignement.

3.3 Les dossiers administratifs en tant que base de sondage

Les données du CCD peuvent aussi servir de base de sondage pour des enquêtes portant sur les écoles publiques et les districts scolaires aux États-Unis. L'univers des organismes publics d'enseignement primaire et secondaire ainsi que celui des écoles publiques primaires et secondaires comprennent les données sur le nom, l'adresse, quelques renseignements descriptifs additionnels et les effectifs scolaires de toutes les écoles et de tous les districts scolaires du pays. En outre, l'univers des écoles publiques comprend le nombre de titulaires de classe dans chaque école. Ces listes ont souvent été utilisées comme bases de sondage pour les enquêtes sur les écoles publiques et les districts scolaires. Un exemple récent de ce genre d'utilisation est la Public School Survey de 1985 (enquête sur les écoles publiques) du CES, pour laquelle un échantillon représentatif de 2,081 écoles a été tiré à l'échelle nationale à partir des listes de l'univers du CCD.

L'utilisation du CCD en tant que base de sondage a été un facteur déterminant dans l'ajout de nouveaux éléments d'information à ces listes, notamment le nombre d'élèves admissibles au programme de repas gratuits et le nombre d'élèves selon la catégorie raciale/ethnique. L'ajout de ces nouveaux éléments, le traitement plus rapide des listes de l'univers des écoles et des organismes d'enseignement et une collaboration étroite avec les États afin de compléter ces listes devraient contribuer à accroître leur utilité en tant que bases de sondage. Comme il n'existe pas de liste équivalente pour les écoles privées, le CES est en train d'en établir une avec l'aide du Bureau of the Census. L'établissement d'une telle liste se fait en trois étapes. Il faut d'abord se procurer une liste d'une entreprise américaine qui se spécialise en données sur l'enseignement, laquelle est constituée à des fins commerciales à l'intention de sociétés d'édition de livres. Il faut ensuite obtenir les listes contenues dans les dossiers administratifs d'un grand nombre d'associations d'écoles privées. Ces associations sont formées de religieux ou de laïques qui sont responsables d'une proportion importante des grandes écoles privées. La dernière étape consiste à établir une base aérolaire dans environ 75 UPE qui sont soumises à une vérification poussée afin de découvrir toute école privée qui ne figurerait pas sur les deux premières listes. Une fois ces trois étapes terminées, le CES s'attend à ce que la base de sondage des écoles privées soit complète dans une proportion de 95 pour cent, et que les établissements manquants soient surtout ceux de petite taille.

3.4 Pondération des données-échantillon

Des systèmes de dossiers administratifs complets peuvent aussi être utilisés pour pondérer les données-échantillon en fonction de paramètres démographiques connus et précis. Une telle pondération est utile à trois points de vue. Premièrement, elle permet de réduire la variance des estimations de l'échantillon, étant donné que le redressement permet d'incorporer des données qui ne présentent pas de variance d'échantillonnage, le système des dossiers étant complet. Deuxièmement, elle peut contribuer à supprimer le biais attribuable à la sélection qui se produit lors du processus d'échantillonnage. La répartition des éléments de l'échantillon est transformée de façon à imiter la répartition connue de la population. Troisièmement, les estimations publiées des enquêtes sont en accord avec les autres données publiées du système des dossiers administratifs. Cette méthode de redressement par pondération est souvent utilisée lors d'enquêtes générales sur la population, lorsque la dernière étape du processus de pondération consiste à grouper les éléments répartis de l'enquête de façon à ce qu'ils correspondent à une population totale connue (ou estimative). Ce genre de redressement peut aussi contribuer à diminuer le sous-dénombrement affectant la base de sondage si le chercheur qui utilise le système de dossiers administratifs est d'avis que ce système, ou un système modifié pour tenir compte du sous-dénombrement, peut fournir de meilleures estimations des totaux.

3.5 Utilisation des dossiers administratifs pour les estimations régionales

Il arrive qu'un chercheur veuille avoir des renseignements à un niveau géograhique ou démographique très détaillé, mais qu'il ne dispose pas des ressources nécessaires pour mener une étude adéquate à ce niveau. Dans un tel cas, les données des dossiers administratifs peuvent être combinées aux données d'enquête pour établir des estimations à des niveaux d'analyse plus détaillée. À cette fin, le système des dossiers administratifs peut être utilisé de deux façons: premièrement, pour l'établissement d'estimations synthétiques directes fondées sur les rapports entre des éléments de l'enquête, lesquels rapports sont appliqués au système des dossiers administratifs. Les rapports peuvent être estimés au niveau de l'État (par exemple, la disponibilité moyenne des ressources en fonction du nombre d'enseignants et des effectifs scolaires), puis appliqués aux dossiers administratifs pour obtenir une estimation au niveau des districts scolaires. Dans le cas présent, l'emploi du terme "synthétique" est approprié parce que les rapports sont censés être valables à tous les niveaux d'analyse, et aucune utilisation n'est faite des données d'enquête qui peuvent être obtenues à des niveaux géographiques plus détaillés.

La deuxième façon d'utiliser les données des dossiers administratifs est d'établir des estimations au niveau d'analyse voulu à l'aide d'estimateurs efficaces empruntés. Grâce à cette technique, des rapports sont établis et utilisés pour faire des estimations se rapportant à des régions où aucune donnée-échantillon n'a été recueillie. C'est une technique qui convient particulièrement aux études sur l'enseignement étant donné que des renseignements de différents niveaux peuvent être employés (par ex., la région, l'État, le district scolaire et, en dernier lieu, l'école) pour faire des estimations au niveau le plus détaillé, les estimations étant relativement sûres pour le modèle à des niveaux moins détaillés et proportionnellement moins efficaces à des niveaux plus détaillés.

3.6 Imputation des données manquantes

Dans les enquêtes-échantillon, l'imputation des données manquantes est habituellement effectuée à l'aide de la méthode dite du "hot deck". Une autre solution consiste à utiliser des renseignements et des liens établis à partir des données des dossiers administratifs. Cette solution peut servir à imputer une valeur aux données manquantes lorsque la population et plus précisément l'élément qui fait l'objet de l'imputation est relativement stable. Si ce n'est pas le cas, mais qu'il existe une relation qui pourrait permettre de faire

l'imputation, des valeurs antérieures ou actuelles du CCD peuvent être utilisées pour établir une estimation en vue de l'imputation, des perturbations aléatoires ou non étant employées pour déterminer la valeur finale imputée.

3.7 Accès des autres utilisateurs à la base de données

Les bases de données établies pour le CES sont aussi constituées en fichiers à grande diffusion qui sont vendus à un prix nominal à tout chercheur qui en fait la demande. De plus, des listes complètes des codes et des guides fournissant des exemples de l'utilisation appropriée des données à des fins d'analyse sont établis et remis avec les bandes de données. Les chercheurs d'autres organismes fédéraux, des universités et d'associations du secteur privé ont utilisé les données du CCD pour analyser et décrire le système de l'enseignement aux États-Unis. Les données du CCD sont utilisées par d'autres services du ministère de l'Éducation pour l'établissement du rapport annuel du Secrétaire de l'Éducation, par le Bureau of the Census pour calculer les taux de migration et par des organismes tels que le National Governors Association (association nationale des gouverneurs) et le Council of Chief State School Officers (conseil des principaux administrateurs des écoles des États) pour créer des indicateurs de la qualité et de l'évolution de l'enseignement.

4. CONSIDÉRATIONS D'ORDRE QUALITATIF RELATIVEMENT À L'ÉTABLISSEMENT DU SYSTÈME DES DOSSIERS ADMINISTRATIFS

Les considérations qualitatives portent sur trois principaux points:

1) la comparabilité des données d'un questionnaire d'enquête à un autre;

2) la comparabilité des données d'un État à un autre;

3) la comparabilité des données de périodes différentes.

Pour ce qui est de la comparabilité des données d'un questionnaire d'enquête à un autre, quelques données de base se retrouvent sur plusieurs questionnaires. Les données sur les inscriptions sont recueillies au niveau des États, des districts scolaires et des écoles. Récemment, le CES a agrégé les données sur les inscriptions déclarées par les écoles afin d'obtenir les totaux pour les États. Pour la plupart des États, ce total ne correspondait pas aux chiffres relatifs aux inscriptions du State Nonfiscal Report. Les écarts avaient tendance à être plutôt faibles, mais le seul fait qu'on ait observé des différences entre des résultats portant sur le même univers est inquiétant en soi. Certaines des différences sont attribuables à des problèmes de définition et à un manque de précision.

Jusqu'ici, la plupart des États ont déclaré sur le State Nonfiscal Report seulement les écoles qu'ils subventionnent et qui relèvent de leur compétence. Toutefois, ils sont disposés à inclure dans le School Universe Report les autres écoles qui sont financées à l'aide des fonds publics, tout en ne relevant pas du State Education Agency. Par conséquent, le écoles qui sont sous la responsabilité d'autres organismes d'un État, par exemple les institutions d'éducation surveillée pour jeunes, peuvent être incluses dans le School Universe Report, mais pas dans le State Nonfiscal Report. Inversement, dans certains États, des étudiants sont inscrits à un programme, mais pas à une école, et figurent donc sur le State Nonfiscal Report, mais non sur le School Universe Report. Bien que les différences qui résultent de telles situations soient insignifiantes et puissent être expliquées après quelques recherches, elles demeurent un problème.

Le CES travaille en collaboration avec les États en vue d'améliorer sa base de données de façon qu'elle soit plus utile pour les responsables des politiques et les chercheurs. Les

efforts visent surtout à préciser les éléments de données recueillis et à uniformiser les définitions et les méthodes de collecte des données. Bien que le CES publie depuis longtemps des manuels et des définitions, il s'est appliqué récemment, avec le concours des États, à clarifier les définitions et, dans la mesure du possible, à rendre semblables les méthodes de collecte et de déclaration des États. Des négociations ont eu lieu avec chaque État afin de déterminer ce que ces derniers peuvent et ne peuvent pas fournir et de relever toute différence entre les besoins en données du CCD et les données que les États fournissent au CES. Lorsque des différences étaient constatées, elles étaient expliquées clairement et on tentait de mesurer l'importance de ces différences. Dans le cadre de ces négociations, le CES a demandé aux États à quel moment ils pourraient fournir les données demandées tout en tenant compte des définitions du CES. Il y a donc eu un effort concerté de la part du CES pour inciter les États à adopter des définitions et des méthodes de déclaration plus uniformes et pour améliorer la comparabilité des données entre les États.

À mesure que les États travaillent avec le CES à uniformiser les méthodes et les définitions, certains des changements résultant de l'uniformisation causeront une discontinuité dans les données déclarées au CES. Comme il a été mentionné précédemment, une telle situation crée un problème pour quiconque veut effectuer une analyse des tendances ou faire des projections fondées sur les données de séries chronologiques. Bien que la tendance à l'uniformisation à travers le pays ait perturbé les séries chronologiques de certains États et continuera de causer certains problèmes pendant encore quelques années, on ne prévoit pas de difficultés insurmontables. Les projections peuvent être modifiées de façon à tenir compte de ces différences et discontinuités. Grâce aux efforts soutenus du CES, les définitions et les méthodes de collecte et de déclaration se stabiliseront de nouveau, et auront maintenant l'avantage d'être uniformes et comparables d'un État à un autre.

5. RÉSUMÉ

Les utilisations statistiques des données des dossiers administratifs sont nombreuses et variées. Les données administratives constituent une composante importante des séries de produits offerts par le CES. Ces données sont utilisées de façon descriptive pour faire état de la situation de l'enseignement aux États-Unis, pour établir des prévisions, pour tenir lieu de base de sondage, pour pondérer un échantillon en vue de réduire la variance et le biais, pour faire des estimations régionales, pour imputer des données manquantes et pour servir de base de consultation pour les chercheurs de l'extérieur du CES. La validité des résultats de ces utilisations statistiques dépend de la comparabilité des données dans le temps, d'un questionnaire à un autre et d'un État à un autre. Le CES a entrepris d'évaluer et d'améliorer la qualité de cette importante ressource.

6. BIBLIOGRAPHIE

Education in the States, (1987). Volume I: State Education Indicators, Council of Chief State School Officiers, Washington, D.C.

Results in Education: (1987). National Governors Association, Washington, D.C.

Magnani, Robert J., Cowan, Charles D., Biemer, Paul P., et Turner, Anthony G. (1985). Evaluating Censuses of Population and Housing, document de formation statistique ISP-TR-5, Bureau of the Census, Washington, D.C., septembre 1985.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

LA BASE DE DONNÉES DE SIMULATION DE POLITIQUE SOCIALE UN EXEMPLE D'INTÉGRATION DE DONNÉES D'ENQUÊTES ET DE DONNÉES ADMINISTRATIVES

MICHAEL WOLFSON, STEVEN GRIBBLE
MICHAEL BORDT, BRIAN MURPHY et GEOFF ROWE[1]

RÉSUMÉ

Ce document décrit la construction d'une base de données prototype expressément conçue pour permettre l'analyse des politiques touchant l'impôt sur le revenu des particuliers et la taxe de vente ainsi que le transfert de revenus. Les politiques fiscales et de transfert sont de plus en plus considérées comme nécessitant une analyse intégrée qui ne soit pas entravée par les limites traditionnelles des sphères de compétence et des programmes. On a élaboré le modèle de simulation de politique sociale (MSPS) et la base de données de simulation de politique sociale (BDSPS) afin de permettre la modélisation micro-analytique; ce modèle combine des données administratives sur les particuliers provenant des déclarations fiscales des particuliers et des dossiers des demandes d'assurance-chômage avec des données d'enquêtes sur les revenus et les dépenses des familles. On a fait un usage considérable de données administratives agrégées additionnelles lors des phases du projet qui portaient sur la création de la base de données et sur la modélisation. Des données d'entrée-sortie ont aussi été utilisées lors de la modélisation des taxes de vente et des droits dans la mesure où ces taxes et droits se rapportent aux dépenses personnelles en biens et services de consommation. Les techniques utilisées pour créer la base de données et éviter de divulguer des données confidentielles comprennent diverses formes d'appariements et d'imputations stochastiques.

1. INTRODUCTION

De nombreux problèmes et questions courants en matière de politique gouvernementale portent sur la position économique des particuliers et des familles dont ils font partie. Cela comprend des questions portant sur l'impôt des particuliers, les prestations d'assurance-chômage et les prestations d'aide sociale. On ne s'intéresse pas seulement aux chiffres agrégés tels que l'ensemble des impôts payés et les prestations de retraite moyennes, mais aussi à des questions détaillées qui intéressent la distribution, par exemple la répartition par groupe de revenu des prestations versées en vertu des programmes de transfert ainsi que la composition des genres de famille dans les groupes

[1] M. Wolfson, S. Gribble, M. Bordt, B. Murphy et G. Rowe, Division des études sociales et économiques, Statistique Canada, Ottawa, Ontario. K1A 0T6.

de revenu les plus faibles. L'analyse de questions de ce genre exige plus que des tableaux de statistiques. Il faut utiliser une microsimulation: qui est l'usage efficace de techniques de modélisation sur un échantillon représentatif de personnes et de familles qui contient une gamme étendue de données.

Les microdonnées recueillies lors de différentes enquêtes statistiques et par différentes procédures administratives sont conçues pour répondre à des besoins particuliers. Aucun ensemble de données ne donne une image suffisamment intégrée et détaillée des ménages canadiens pour appuyer l'analyse de questions portant sur l'impôt des particuliers et la politique de transfert. Par exemple, les statistiques administratives sur l'assurance-chômage sont tirées d'un programme basé sur les particuliers qui touchent un salaire hebdomadaire; elles ne contiennent donc pas de renseignements sur les familles dont les prestataires font partie pas plus que sur les autres sources de revenu dont ces familles peuvent disposer. Bien que le programme d'assurance-chômage soit basé sur les particuliers, il est quand même souhaitable d'analyser les effets des changements dans les règlements de l'assurance-chômage sur les revenus familiaux ainsi que sur d'autres programmes sociaux tels que les prestations d'aide sociale et les revenus imposables des particuliers.

La base de données de simulation de politique sociale (BDSPS) ainsi que le logiciel connexe (le modèle de simulation de politique sociale - MSPS) ont pour but général de permettre à tous les intéressés de disposer des moyens pour effectuer une analyse complète et intégrée, basée sur la microsimulation, des politiques en matière d'impôt sur le revenu des particuliers et de transfert. Ce document donne un aperçu de la constitution de la base de microdonnées, la SPSD.

Le document décrit d'abord les objectifs généraux de la BDSPS ainsi que le caractère des données de base. Puis, la partie principale du document comprend la description des nombreuses étapes qui ont permis de constituer la BDSPS.

2. OBJECTIFS, SOURCES DES DONNÉES ET TECHNIQUES

Au cours de l'élaboration de la BDSPS, tous les moyens ont été pris afin d'assurer la variété et l'utilité des données de base originales tout en en assurant la confidentialité de sorte que la base de données et le modèle résultants puissent être diffusés sans contrainte. Quatre objectifs principaux ont donc guidé le choix des techniques, des sources de données, des variables ainsi que des méthodes:

- **Accessibilité au public/non-confidentialité**

  Le premier objectif a été de s'assurer qu'aucune personne réelle qui serait représentée dans l'une quelconque des bases de données ne pourrait être identifiée soit à la suite de divulgation explicite ou de déduction par recoupement. Ce n'est que si cette exigence est remplie que la BDSPS et le MSPS peuvent être diffusés dans le public afin d'accroître la richesse de l'analyse et des discussions, au sein du public, dans les domaines de la politique sociale et de la politique fiscale au Canada. Le fait que la base de données ainsi que le modèle puissent être utilisés avec un micro-ordinateur de prix relativement faible équipé d'une mémoire et de périphériques de mémorisation d'une capacité limitée constitue un aspect additionnel de l'accessibilité au public.

- **Précision des données globales et de la distribution des données**

  La BDSPS et le MSPS ont été conçus afin de reproduire le mieux possible des totaux globaux "connus" comme le nombre total de prestataires d'assurance-chômage. Comme objectif secondaire, on a aussi tenté de représenter la

distribution des données globales parmi plusieurs classifications qui ont une importance fondamentale en matière d'analyse de la politique gouvernementale au Canada, telles que la province, l'âge, le revenu, le genre de famille et le sexe. Les autres distributions ont été préservées dans la mesure du possible.

— **Détail et intégralité des données**

Dans le choix et le groupement des variables tirées des principales sources de données, on a tenté de prévoir les options vraisemblables en matière de politique ainsi que de répondre aux besoins des modèles courants de l'impôt et des transferts. Par exemple, les coûts pour la garde d'un enfant sont inclus dans la base de données bien qu'aucun des modèles utilisés actuellement ne les emploie.

— **Cohérence des micro-enregistrements**

Dans la mesure du possible, on a tenté d'éviter la création de personnes et de ménages irréalistes. Par exemple, un couple âgé sans enfant qui pourrait avoir la pleine déduction pour la garde d'un enfant.

Il existe une interdépendance élevée entre ces objectifs centraux et il a fallu faire des compromis. Quand il a fallu effectuer ces choix on a fait appel à des fonctionnaires de plusieurs ministères fédéraux qui étaient intéressés aux résultats et qui possédaient déjà de l'expérience avec leurs propres modèles de simulation. Le résultat ultime représente donc un compromis entre des préoccupations sur les plans méthodologique, de l'information, technologique, ministériel et en matière de politique gouvernementale.

La BDSPS a été constituée à partir de quatre principales sources de données.

— ***L'Enquête sur les finances des consommateurs* (EFC):** C'est la source principale de données dont dispose Statistique Canada sur la distribution du revenu parmi les personnes et les familles; elle a servi d'ensemble de données receveur. La base de données constituée à la suite de cette enquête est riche en détails sur la structure et le revenu des familles, mais elle manque de renseignements détaillés sur les périodes de chômage, les déductions d'impôt et les dépenses des consommateurs.

— Données relatives aux déclarations d'impôt des particuliers: L'échantillon de trois pourcent des déclarations d'impôt (T1) qui sert de base à la publication annuelle de Revenu Canada ***Statistique fiscale*** (le Livre vert).

— Données relatives aux demandes d'assurance-chômage (AC): un échantillon de un pourcent tiré spécialement parmi tous les dossiers des demandes d'assurance-chômage contenus dans le système administratif d'EIC.

— L'***Enquête sur les dépenses des familles (EDF)*:** C'est l'enquête périodique de Statistique Canada qui vise à recueillir des données très détaillées sur le revenu des Canadiens ainsi que sur la structure de leurs dépenses au niveau du ménage, y compris des renseignements sur les variations nettes dans les éléments d'actif et les engagements (épargne annuelle).

Ces sources de données originales sont confidentielles et n'ont jamais été diffusées intégralement. Elles sont plutôt diffusées soit sous forme d'échantillons à grande diffusion dans lesquels de nombreux enregistrements et variables sont supprimés ou sous forme de tableaux récapitulatifs dont certaines cases peuvent être supprimées pour assurer la confidentialité.

Aux fins de la base de données de simulation de politique sociale (BDSPS), ces quatre sources de données ont été transformées en un seul ensemble de microdonnées non confidentiel à grande diffusion. De plus, ces microdonnées ont été augmentées à l'aide de diverses données globales qui sont surtout destinées à servir de repères ou de totaux de contrôle. Ces données globales ont été tirées du recensement de 1981, de rapports

administratifs du Régime d'assistance publique du Canada, des statistiques de l'état civil, de rapports sommaires de Santé et Bien-être social Canada ainsi que des sources de microdonnées elles-mêmes.

Le regroupement des quatre ensembles de microdonnées initiaux, l'addition de nouveaux renseignements ainsi que le remplacement ou la correction de mesures biaisées ont dépendu dans une large mesure de quatre techniques qui ont été beaucoup employées dans la création de la BDSPS: la correction proportionnelle itérative, l'imputation stochastique, le groupement de micro-enregistrements et la fusion stochastique.

La correction proportionnelle itérative (CPI) est une technique utilisée pour réduire le biais en faisant correspondre les données à des totaux de contrôle connus. Par exemple, les poids d'une enquête peuvent être corrigés afin que la répartition de la population de l'enquête selon l'âge et le sexe corresponde à la répartition "connue" de la population par âge et par sexe (c.-à-d. à celle qui est tirée des données d'un recensement).

L'**imputation stochastique** est la création, par tirage aléatoire, de valeurs de données synthétiques pour des personnes dans un ensemble de données receveur à partir de distributions ou de fonctions de densité provenant d'un ensemble de données de base.

Le **groupement de micro-enregistrements** est le processus qui consiste à créer des micro-enregistrements synthétiques par la combinaison statistique de groupes d'enregistrements semblables. Les micro-enregistrements de l'ensemble de données receveur sont groupés d'après des critères relatifs aux politiques. Dans chaque groupe, on fait la moyenne des valeurs des variables pertinentes (par ex., les gains en capital) afin de créer des enregistrements non identifiables qui ressemblent aux microdonnées mais qui sont en fait synthétiques.

La **fusion stochastique** comporte tout d'abord le classement des enregistrements d'un ensemble de données receveur ainsi que d'un ensemble de données donneur d'après des critères relatifs aux politiques communs aux deux ensembles de données (par ex., le mode d'occupation du logement, le statut professionnel, la catégorie de revenu). Les renseignements qui font partie des enregistrements donneurs ainsi classés peuvent alors être attribués aux enregistrements de l'ensemble de données receveur qui ont des caractéristiques semblables sans qu'il soit possible de mieux identifier les personnes qui correspondent à ces enregistrements.

La figure 1 donne un aperçu du processus de création de la BDSPS. Les ellipses représentent des fichiers de données (par ex., l'EFC, le Livre vert) et les rectangles représentent des opérations. Ensuite la partie principale du document contient la description de chacune des étapes de la construction de la BDSPS, telle qu'indiquée dans la figure 1.

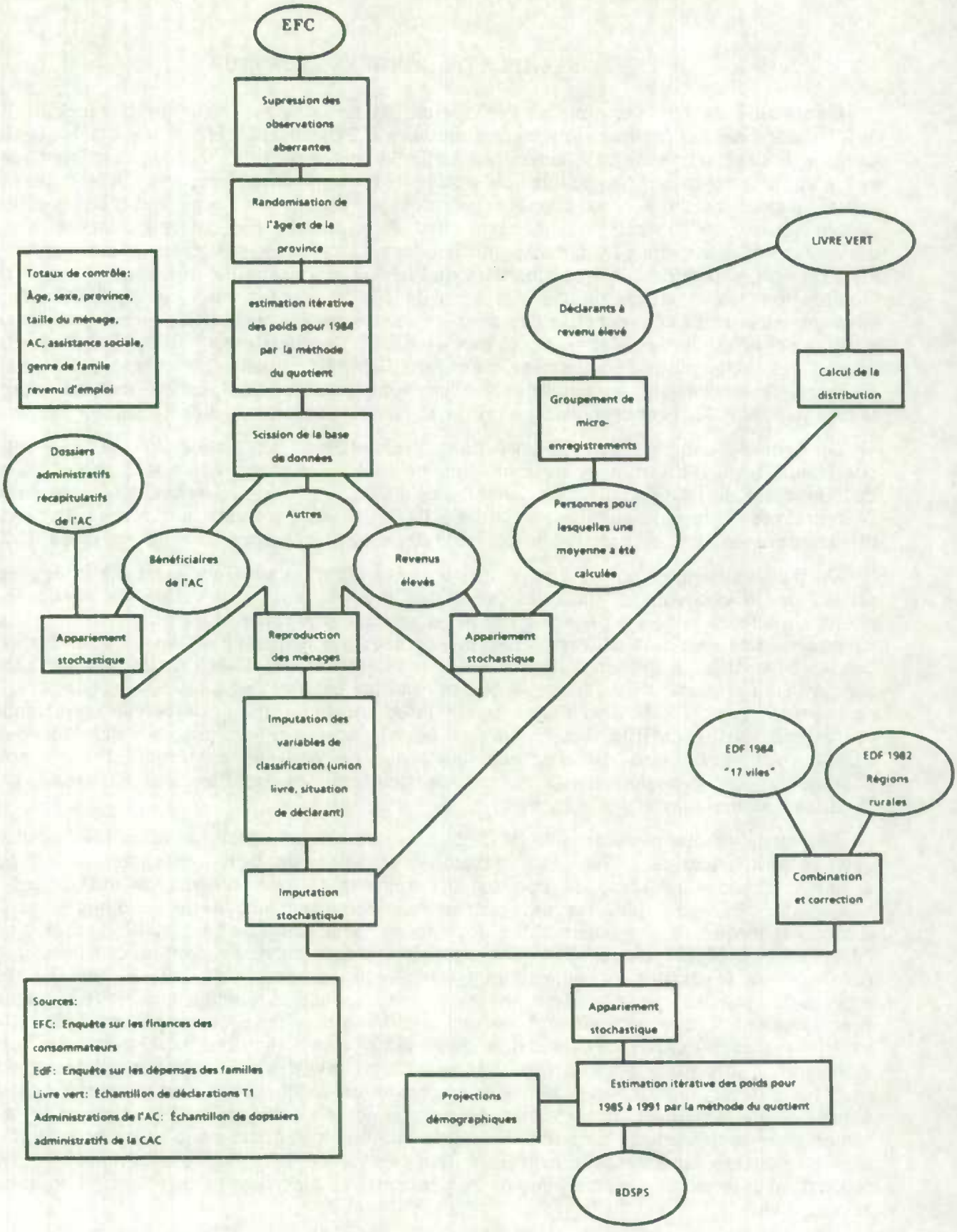


Figure 1: Processus de création de la base de données de simulation de politique sociale (BDSPS)

3. L'ENSEMBLE DE DONNÉES RECEVEUR

L'ensemble de données cible ou "receveur" est tiré des renseignements recueillis lors de l'***Enquête sur les finances des consommateurs*** (EFC) de 1984 effectuée par Statistique Canada; il s'agit d'une enquête annuelle à laquelle participent des ménages choisis tirés du cadre de l'Enquête sur la population active (EPA). Quatre formules différentes sont utilisées pour recueillir les données fournies par chaque ménage qui fait partie de l'échantillon. Le Dossier du ménage sert à recueillir des renseignements d'ordre démographique sur chaque membre du ménage, ainsi que des renseignements sur la structure de la famille. Le questionnaire de l'EPA vise à recueillir des renseignements sur la situation des membres du ménage âgés de 15 ans et plus vis-à-vis de l'activité. Le questionnaire de l'EFC recueille des données sur le revenu, selon la source, pour chacun des membres du ménage âgés de 15 ans et plus. Le questionnaire de l'Enquête sur le revenu des ménages et l'équipement ménager (ERMEM) recueille des renseignements sur les caractéristiques du logement et sur certains genres d'équipement ménager qui s'y trouvent. En 1984, l'enquête visait environ 36,000 ménages composés de 98,000 personnes.

On trouve, pour chaque ménage dans l'échantillon, un Dossier du ménage et un questionnaire de l'ERMEM et pour chaque membre du ménage âgé de 15 ans et plus un questionnaire de l'EPA et un questionnaire de l'EFC. A cause de la quantité considérable de renseignements déjà couplés recueillis à l'aide de ces formules, cette base de données hiérarchique combinée constitue le point de départ du processus de création de la BDSPS.

On peut remarquer que, même si ces données diverses sont entièrement intégrées au niveau des microdonnées, dans les premières phases de la production de l'enquête, le public n'a encore jamais eu accès à ces renseignements considérables portant sur plusieurs variables. Les résultats de l'enquête sont diffusés par Statistique Canada sous forme de bandes-échantillon à grande diffusion ou de publications imprimées distinctes portant sur les revenus des particuliers, les revenus des familles économiques, les revenus des familles de recensement, l'ERMEM et l'Enquête sur la population active. Cette vue traditionnelle et fragmentée de l'utilité des ensembles de microdonnées est mise en question par la BDSPS. L'objectif visé est d'obtenir une base de données entièrement hiérarchique comprenant des renseignements sur les particuliers, les familles de recensement, les familles économiques et les ménages.

Les renseignements obtenus de l'AC, du Livre vert et de l'EDF ont alors été "ajoutés" à ceux recueillis lors de l'EFC. Afin d'exploiter complètement ces renseignements imputés à partir d'autres sources, de nombreux enregistrements originaux de l'EFC ont été reproduits. Par exemple, les enregistrements correspondant à des chômeurs ont été reproduits jusqu'à ce que leur nombre corresponde à la taille de l'échantillon du fichier de l'AC (environ 30,000 personnes). Les enregistrements correspondant à des personnes à revenu élevé (ceux des personnes ayant un revenu de plus de $80,000 en 1984) ont été reproduits jusqu'à ce que leur nombre corresponde à celui des enregistrements correspondant à des personnes à revenu élevé obtenus par regroupements de micro-enregistrements à partir de l'échatillon de Revenu Canada (environ 5,000 personnes). Pour maintenir la structure des familles et la somme globale des poids, les enregistrements de tous les autres membres des ménages comptant un chômeur ou une personne à revenu élevé ont été reproduits de la même façon. Le poids attribué à un enregistrement a été réduit pour tenir compte du nombre de fois que cet enregistrement a été reproduit. La base de données ainsi obtenue renferme plus de 170,000 enregistrements dont une grande proportion représente des ménages où l'on retrouve des personnes en chômage ou avec un revenu élevé.

3.1 Suppression des observations aberrantes

La non-confidentialité de la base de données créée (BDSPS) peut être garantie si chaque ensemble de microdonnées en entrée est lui-même non-confidentiel et si la "fusion" des données ne comporte pas un appariement exact. C'est la stratégie qui a été adoptée, elle commence avec le filtrage du fichier de l'EFC.

Les versions à grande diffusion des données de base (EFC) font l'objet d'un filtrage préliminaire pour trouver les cas potentiellement délicats. Par exemple, les ménages de plus de neuf membres sont supprimés du fichier des ménages qui sera diffusé dans le public et les familles de recensement qui comptent plus de quatre bénéficiaires de l'AC sont supprimées du fichier des familles de recensement qui sera diffusé dans le public. L'étape initiale de la construction de la BDSPS consistait donc à supprimer chaque ménage qui répondait aux critères de filtrage de l'EFC (c.-à-d. les critères appliqués aux niveaux des ménages, des familles économiques ou des familles de recensement).

Outre la suppression de ménages entiers, il y a eu recodage de certaines données de l'EFC. Cela comprenait, par exemple, la fusion de certaines régions géographiques (par ex., la fusion de Brandon avec Winnipeg) ou le fait de donner la valeur "inconnu" au code de profession des conjoints des personnes à revenu élevé.

3.2 Randomisation

La randomisation des combinaisons âge-sexe et des régions constitue une protection additionnelle contre la diffusion de renseignements sur des ménages qui pourraient être identifiés.

On suppose que le fait de révéler la composition d'un ménage selon l'âge et le sexe ainsi que l'endroit où ce ménage se trouve peut augmenter le risque de manquement aux règles de confidentialité. Cependant, ce risque peut être réduit considérablement si l'on randomise l'âge des membres du ménage parmi des groupes d'âge qui couvrent une période de cinq ans et si l'on randomise le sexe des enfants (c.-à-d. des personnes de 15 ans ou moins).

De même, l'endroit où l'on trouve des ménages de genre inhabituel peut être changé si l'on randomise leur province et les codes de catégorie de taille de la région urbaine. On entend par ménages de genre inhabituel ceux qui comprennent plus de huit personnes, plus de deux familles de recensement ou plus d'une famille économique ou ceux qui contiennent des personnes avec des caractéristiques spéciales quant au revenu ou à l'impôt payé (par ex., les femmes avec un revenu de plus de $80,000, ou les hommes et les femmes avec un revenu de moins de $150,000 et un impôt sur le revenu de plus de $150,000).

3.3 Correction proportionnelle itérative (CPI)

Comme la BDSPS comprend des ménages et des structures familiales complets, il est essentiel d'associer à chaque ménage un poids unique qui assurera la cohérence dans les totalisation aux niveaux des ménages, des familles et des particuliers. Cela n'est pas fait actuellement parce que les bases de données à grande diffusion de Statistique Canada comprennent des poids distincts aux niveaux des particuliers, des familles de recensement et des familles économiques.

On a eu recours à la CPI à plusieurs niveaux pour obtenir cette cohérence. Cette procédure est une généralisation de la procédure de la CPI ordinaire (généralement appelée méthode itérative du quotient) utilisée avec les données de l'EFC pour obtenir, au niveau des particuliers, des poids cohérents avec les totaux de contrôle connus pour l'âge et le sexe. On peut considérer qu'il s'agit de corrections (proportionnelles) successives apportées aux poids de l'enquête pour que ces derniers soient conformes avec des totaux

de contrôle déterminés à l'avance. Pour la CPI à plusieurs niveaux, les corrections peuvent être appliquées aux niveaux des ménages, des familles et (ou) des particuliers avec une étape additionnelle dans laquelle on remplace les poids (corrigés) des particuliers dans un ménage par la moyenne du ménage.

Il y a en outre, dans l'EFC, des biais de déclaration qui en limitent l'utilité pour la modélisation des programmes d'impôt et de transfert, par exemple:

- non-déclaration de personnes avec des revenus élevés,
- sous-déclaration des revenus provenant de l'aide sociale et
- sous-déclaration des revenus de placements.

Les poids des enregistrements de l'EFC ont été recalculés à l'aide de la correction proportionnelle itérative afin qu'ils correspondent à des totaux de contrôle externes comme le nombre de personnes à revenu élevé (avec un revenu total de plus de $80,000 en 1984), la taille des familles par province, et les revenus provenant de régimes de pension privés ainsi que les prestations d'aide sociale par province.

Les totaux de contrôle utilisés pour construire les poids pour la BDSPS représentaient: a) les particuliers selon l'âge et le sexe, b) les particuliers selon la catégorie de revenu, c) les prestataires de l'AC, d) les ménages selon la composition de la famille et l'activité, e) les ménages selon les prestations d'aide sociale et f) les bénéficiaires d'une pension. Chacun de ces totaux de contrôle a été désagrégé par province.

Il a été démontré (recherche, dont les résultats n'ont pas été publiés, effectuée par Georges Lemaître de la Division des méthodes d'enquêtes sociales de Statistique Canada) qu'une procédure de pondération par la méthode des moindres carrés, au niveau des ménages, qui est à peu près équivalente à la procédure de la CPI décrite ici, permet d'obtenir des estimations améliorées des caractéristiques de la population. En particulier, dans le cas où l'on n'utilise que des totaux de contrôle relatifs à l'âge et au sexe (cas pour lequel les procédures de la CPI et des moindres carrés sont des équivalents exacts), la variance d'échantillonnage des caractéristiques au niveau de la famille est d'environ 50% inférieure à celle qui est obtenue si l'on emploie la méthode de la personne principale.

3.4 Scission de la base de données

On désigne par scission une étape de préparation mécanique des données au cours de laquelle les données de l'EFC (après suppression des observations aberrantes, randomisation et CPI) sont séparées en trois sous-ensembles mutuellement exclusifs: les personnes à revenu élevé, les prestataires de l'AC et toutes les autres personnes. Pour simplifier les étapes suivantes de la création de la base de données, cette scission est effectuée de façon à ce qu'aucun ménage dans lequel on trouve des personnes à revenu élevé ne renferme aussi des prestataires de l'AC. Il existe, en fait, une poignée de ménages de ce genre, mais dans ces ménages on a traité les prestataires de l'AC comme s'ils n'avaient pas reçu de prestations d'assurance-chômage. Les personnes à revenu élevé sont celles dont le revenu dépasse $80,000 alors que les prestataires de l'AC sont les personnes qui ont déclaré, lors de l'enquête sur les finances des consommateurs, qu'elles avaient reçu des prestations d'assurance-chômage.

4. APPARIEMENT STOCHASTIQUE

L'appariement stochastique comporte la création d'enregistrements composés 'fusionnés' provenant de deux bases de microdonnées. Considérons deux bases de données, une base receveuse A ainsi qu'une base donneuse B. Diverses méthodes peuvent être utilisées pour attribuer certains ou tous les renseignements d'un enregistrement de la base de données B à un enregistrement donné quelconque de la base de données A. Toutes ces méthodes reposent sur la recherche d'un enregistrement de la base de données B qui, en un certain sens, ressemble à l'enregistrement donné de la base de données A. La détermination de la similarité est fondée sur des variables communes aux deux bases de données et elle dépend de l'usage que l'on compte faire des enregistrements 'fusionnés'. Divers algorithmes du 'plus proche voisin', qui utilisent des méthodes ressemblant à celles utilisées pour l'analyse de grappes peuvent être utilisées pour déterminer un appariement mathématiquement 'optimal', compte tenu d'une méthode particulière de détermination de la distance dans un espace à N dimensions. Des complications surgissent en pratique parce que la taille de l'ensemble d'enregistrements 'donneurs' (base de données B dans notre exemple) est limitée et parce qu'on veut utiliser des variables non-continues (par ex., des variables discrètes ou des variables nominales).

Pour la BDSPS, on a utilisé une technique différente plus heuristique. Cette méthode consiste à séparer les deux bases de données en 'cases', définies de la même façon, d'enregistrements, ces derniers sont alors triés d'après l'une des variables continues que l'on retrouve dans les deux bases de données (habituellement le revenu total pour la BDSPS). Les enregistrements dans une case donnée sont alors appariés avec l'enregistrement correspondant de la case de l'autre base de données (c.-à-d. que l'enregistrement n de la case m de la base de données A est apparié avec l'enregistrement n de la case m de la base de données B). Des complications surgissent parce qu'en général le nombre d'enregistrements dans une case donnée n'est pas identique dans les deux bases de données et aussi, à la suite de la présence de poids pour les enregistrements dans une des bases de données ou dans les deux. Ces problèmes sont réglés par reproduction sélective d'enregistrements d'une ou des deux bases de données.

Pour la BDSPS on utilise l'appariement stochastique afin d'ajouter des données recueillies lors de l'EDF, des données sur l'assurance-chômage ainsi que des données sur le revenu figurant dans le Livre vert dans le cas des personnes à revenu élevé. La technique permet de préserver les corrélations entre les éléments de l'enregistrement donneur. Chacune des procédures d'appariement est décrite plus en détail ci-après, on indique aussi à cet endroit que ces appariements stochastiques excluent virtuellement la possibilité d'un appariement exact.

5. RAJUSTEMENT POUR LES REVENUS ÉLEVÉS

Nous savons que l'EFC comporte un biais de déclaration et un biais d'échantillonnage qui font que le nombre de personnes à revenu élevé ainsi que le revenu en dollars de ces personnes sont inférieurs à ce que les dossiers de l'impôt sur le revenu des particuliers laissent supposer. Dans la création de la BDSPS, des mesures ont été prises pour traiter la sous-déclaration ainsi que la non-déclaration de nombreux éléments relatifs au revenu et aux déductions. La figure 2 donne un aperçu de ce processus de rajustement pour les revenus élevés.

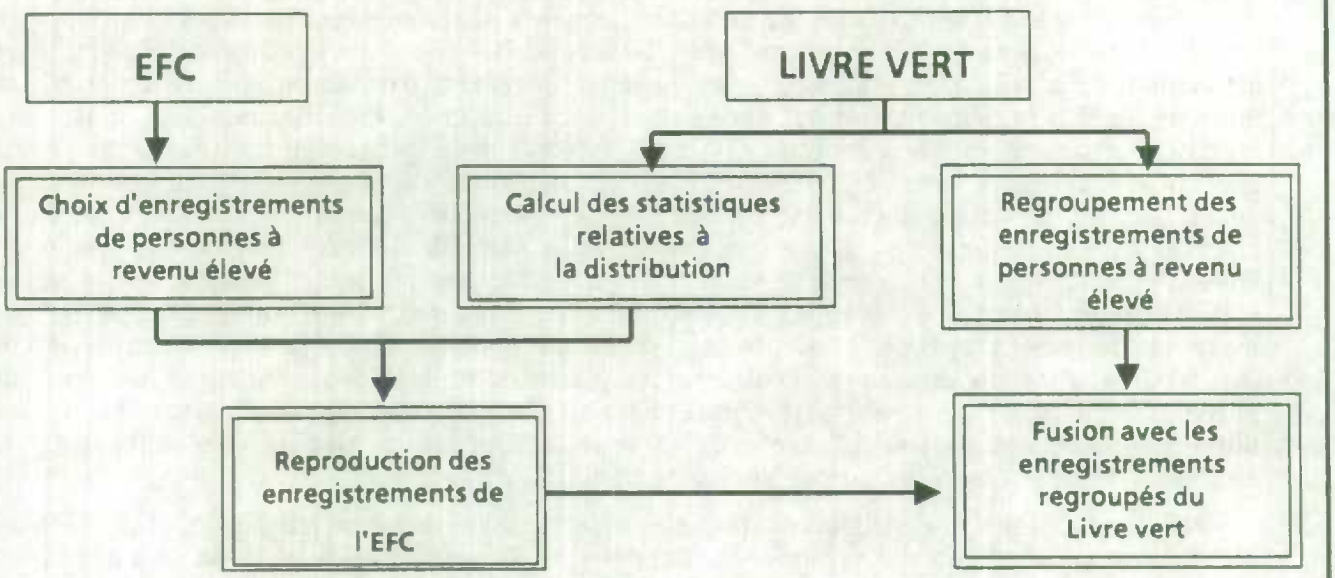


Figure 2: Processus de rajustement pour les revenus élevés

5.1 Regroupement de micro-enregistrements

On corrige l'effet de la non-déclaration de personnes à revenu élevé pour l'EFC en utilisant les totaux du Livre vert pour les personnes avec un revenu de plus de $80,000 comme tolérance pour la CPI. La CPI augmente alors les poids de chaque enregistrement de personne à revenu élevé de l'EFC de sorte que la somme des poids corresponde aux données du Livre vert.

Il existe approximativement 300 enregistrements de ce type. Le processus de CPI les laisse avec des poids très élevés (de l'ordre de 200 à 500). Ces enregistrements sont utilisés comme "receveurs" pour accepter les renseignements plus précis provenant du Livre vert. Ce qui à son tour constitue la base d'un rajustement des éléments de revenu pour le groupe des personnes à revenu élevé.

Même avec un accroissement proportionnel des poids pour les enregistrements des personnes à revenu élevé dans les résultats de l'EFC, il reste encore une sous-déclaration importante des revenus pour ce groupe. Comme deuxième étape, on corrige le biais de sous-déclaration en remplaçant les éléments de revenu de ces enregistrements par des ensembles plausibles mais non identifiables d'éléments de revenu tirés du Livre vert (voir le tableau 1).

Les enregistrements provenant du Livre vert sont groupés en ensembles d'au moins cinq enregistrements. On considère que ces enregistrements groupés constituent un tableau non confidentiel bien qu'ils conservent un bon nombre des caractéristiques des micro-enregistrements. Les groupes représentent des personnes avec un âge et des revenus d'emploi, de placements et de dividendes ainsi que des gains en capital semblables. On calcule, pour ces groupes, ou cinq-uplets, une moyenne pondérée pour les éléments qui figurent au tableau 1. Une fois groupés, les enregistrements sont considérés comme non confidentiels puisqu'ils représentent au moins cinq personnes. Cela équivaut à publier un tableau dans lequel chaque case contient des données sur au moins cinq personnes.

Tableau 1

Éléments du revenu déclarés lors de l'EFC remplacés dans le cas des personnes à revenu élevé

Éléments relatifs à l'emploi
Revenu d'emploi
Revenu net d'agriculture
Autres frais déductibles relatifs à un emploi
Revenu provenant d'un travail autonome - Non agricole
Éléments relatifs aux investissements
Perte en capital déductible pour les autres années
Perte autre qu'en capital déductible pour des années antérieures
Frais financiers
Perte en capital lors de la disposition d'actions de CPCC
Revenu en intérêts
Revenu de location net
Autre revenu de placements
Gain/perte en capital imposable pour l'année
Montant imposable des dividendes de corporations canadiennes imposables
Autres
Autre revenu imposable
Revenu total imputé - Somme des composantes

Le groupement résultant contient 4,676 enregistrements représentant 24,556 enregistrements du Livre vert qui, eux-mêmes, représentent 133,650 déclarants à revenu élevé. Ces enregistrements regroupés, provenant de microdonnées qui seraient confidentielles si elles étaient publiées autrement, peuvent maintenant être intégrés à un ensemble de données à grande diffusion avec peu de perte de renseignements.

5.2 Appariement stochastique

Les 300 enregistrements originaux de la BDSPS sont reproduits jusqu'à ce que leur nombre corresponde à celui des enregistrements regroupés de personnes à revenu élevé (4,676) figurant dans le Livre vert. Ces 300 enregistrements ne constituent pas une base suffisante pour les caractéristiques démographiques de la population des déclarants à revenu élevé. Il ne serait donc pas possible d'effectuer un appariement détaillé selon l'âge, le sexe, la province et le revenu total. On a plutôt procédé par imputation afin de donner aux enregistrements reproduits de la BDSPS une nouvelle valeur pour le revenu total à l'aide d'une division très simple selon l'âge (deux groupes), selon le sexe et selon la région à l'aide de la procédure qui sera décrite dans une section suivante (Imputation stochastique de renseignements sur l'impôt sur le revenu). Cette nouvelle valeur imputée

du revenu total a été utilisée comme clé pour trier les enregistrements de la BDSPS avant la fusion avec les enregistrements regroupés du Livre vert triés de la même façon.

Il faudrait un échantillon original de l'EFC beaucoup plus important pour améliorer l'appariement eu égard à l'âge, au sexe, à la province, au revenu total et à la situation relativement à l'impôt.

5.3 Évaluation

Bien que cette méthode de regroupement de micro-enregistrements assure que les corrélations entre les éléments de revenu et de déductions (qui figurent au tableau 1) sont généralement maintenues, les distributions à une variable des enregistrements synthétiques tendent à avoir une variance inférieure à celle des enregistrements originaux du Livre vert. Cela découle du regroupement de plusieurs enregistrements en un seul. Il arrive très souvent, dans le cas d'éléments rares tels que les "pertes en capital déductibles pour les autres années", que les cinq enregistrements à regrouper contiennent plusieurs valeurs nulles qui sont incluses dans la moyenne. La moyenne est conservée, mais la distribution est biaisée vers la moyenne.

La figure 3 constitue un exemple de la distorsion que cette méthode introduit dans la distribution des gains en capital. En effet, on a attribué à tous les cinq-uplets de personnes de petites valeurs pour les gains en capital plutôt que d'avoir quatre personnes pour lesquelles les gains en capital seraient nuls et une personne qui aurait des gains en capital.

DISTRIBUTION DES GAINS EN CAPITAL
LIVRE VERT COMPARATIVEMENT À LA FUSION STOCHASTIQUE

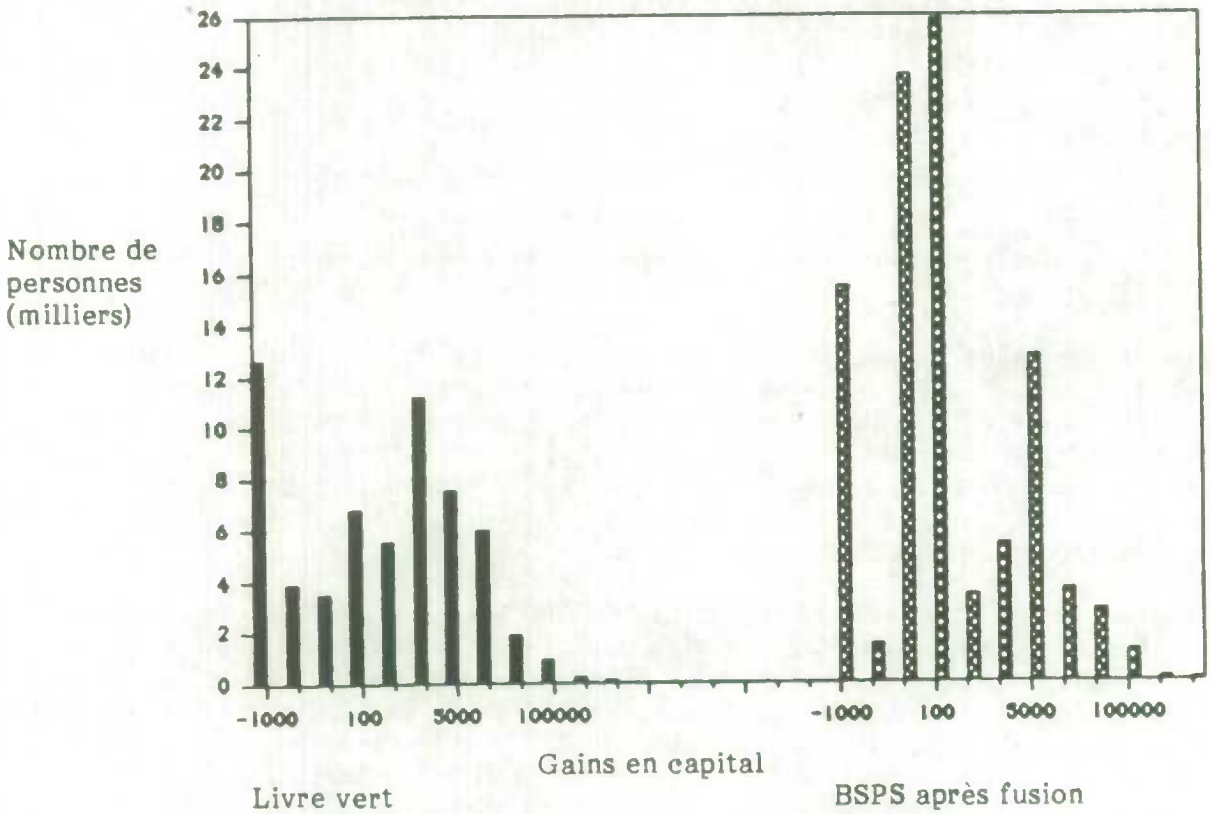


Figure 3: Gains en capital avant et après la reproduction

6. IMPUTATION DES DONNÉES RÉCAPITULATIVES SUR L'ASSURANCE-CHÔMAGE

L'Assurance-chômage (AC) est un programme complexe d'assurance et de maintien temporaire du revenu dont l'administration exige la surveillance de l'activité hebdomadaire des prestataires sur le marché du travail. Les données administratives recueillies en vertu du programme servent i) à suivre les activités des prestataires de l'AC en ce qui a trait aux prestations et aux demandes, ii) à établir l'admissibilité ainsi que le droit aux prestations en surveillant les participations antérieures au programme dans le cas de réitérants ou de personnes qui présentent une nouvelle demande et iii) à surveiller les régimes de travail antérieurs au moyen du document intitulé "Relevé d'emploi".

Le modèle de l'AC a été conçu afin de pouvoir être utilisé comme modèle autonome d'un programme important de paiements de transfert ainsi que pour fournir un élément simulé du revenu d'AC utilisé en entrée dans le modèle de l'impôt. Comme modèle autonome, la sortie demandée sert à préciser les coûts du programme, la population cliente et quels seraient les gagnants et les perdants si d'autres structures étaient utilisées pour le programme; elle permet aussi l'examen de questions relatives au financement du programme. Les entrées dans un système de l'impôt sont constituées par des paiements de prestations simulés basés sur une année civile plutôt que d'après une demande. Ainsi, la tâche initiale lors de la construction de cet élément de la base de données a exigé l'élaboration simultanée d'un module de simulation de l'AC ainsi que la détermination d'une ensemble limité de variables de l'AC (Tableau 2) qui pourraient servir d'entrée dans le module.

6.1 Ensemble de données donneur pour l'AC

Les données récapitulatives administratives de l'AC imputés dans la BDSPS étaient basées sur un échantillon de 1% de la population pour laquelle certaines activités relatives à une demande ont été traitées au cours de l'année civile 1984. L'échantillon est composé d'environ 30,000 personnes et il représente environ 40,000 demandes. Le contenu de cet de données a été spécialement conçu. D'une part, il devait être assez riche pour permettre de saisir les données récapitulatives hebdomadaires sur la population active pertinentes à l'application des règlements du programme de l'AC. D'autre part, il devait être assez compact et général pour ne pas être confidentiel. Pour atteindre ce résultat, on a pensé en fonction de l'histoire d'un événement de sorte que l'on s'est concentré sur la durée des diverses activités plutôt que sur les dossiers hebdomadaires des activités. Le personnel d'Emploi et Immigration Canada ainsi que celui de la Commission d'enquête sur l'assurance-chômage (Commission Forget) ont aidé à concevoir cet ensemble de données. Le tableau 2 montre l'ensemble de variables utilisées en entrée dans le modèle de l'AC.

Parce que ces variables sur les données récapitulatives de l'AC sont interreliées, on a procédé à l'appariement stochastique de chacun des enregistrements des 30,000 prestataires (qui peuvent comprendre une ou deux demandes) avec les enregistrements de l'EFC pour lesquels le répondant avait déclaré avoir touché un revenu d'AC au cours de l'année civile. En plus des variables sur les données récapitulatives de l'AC mentionnées au tableau 2, on a utilisé comme clé d'appariement des renseignements administratifs sur l'âge, la province et le sexe du prestataire. Ces variables étaient disponibles dans l'ensemble de données receveur pour les personnes avec un revenu d'AC.

Les genres de demandes sont un élément important pour l'appariement, puisqu'il existe présentement des différences majeures dans les règles d'admissibilité et dans le droit aux prestations entre les genres de demande. Pour construire une classification des genres de demande dans l'ensemble de données receveur, on a identifié i) les prestataires d'assurance-chômage âgés de 65 ans et plus (prestations de retraite), ii) les prestataires

d'assurance-chômage dont le code d'occupation est "chasse, pêche, piégeage" (prestations de pêcheur) et iii) les femmes bénéficiaires d'AC avec un enfant âgé d'un an ou moins (prestations de maternité). Il a été impossible de faire la distinction entre les prestations de maladie et les prestations ordinaires dans l'ensemble de données receveur.

Tableau 2

Variables récapitulatives de l'AC

Numéro d'ordre de la demande (1re ou 2e de l'année en cours)
Drapeau de réitérant
Genre initial de la demande de prestation
drapeau de changement de genre de demande de prestation
Nombre de semaines de prestations (pour la demande courante)
Nombre de semaines de prestations (au cours des 52 dernières semaines)
Nombre de semaines de travail (avant la demande courante)
Rémunération hebdomadaire moyenne (avant la demande)
Pénalité pour départ volontaire (semaines)
Semaine au cours de laquelle la demande a été établie
Prestations versées au cours de l'année civile (1 ou 2 demandes)
Nombre de semaines de prestations versées au cours de l'année civile

6.2 Appariement stochastique

Pour effectuer l'appariement, on a tout d'abord séparé l'ensemble de données administratives donneur (AC) ainsi que l'ensemble de données receveur (EFC) d'après le groupe d'âge, la province, le sexe et le genre de demande. Les enregistrements dans les cases ont été reproduits afin de faire en sorte que les cases correspondantes des ensembles de données de l'AC et receveur contenaient le même nombre d'enregistrements. Si dans une case quelconque, le nombre d'enregistrements receveurs dépassait celui des enregistrements de l'AC, ces derniers étaient alors reproduits de façon uniforme (les données de l'AC constituaient un échantillon aléatoire simple). De même, si le nombre d'enregistrements de l'AC était supérieur au nombre d'enregistrements receveurs, ces derniers étaient reproduits proportionnellement à leur poids (il faut se rappeler que les données receveuses étaient basées sur un échantillon stratifié). Ce dernier cas était le plus fréquent (170 cas sur 218), mais le premier s'est aussi produit (une conséquence d'un plan d'enquête stratifié). Les poids des doubles des enregistrements de l'ensemble de données receveur ont été ajustés en proportion au nombre de fois que ces enregistrements avaient été reproduits.

Les étapes d'appariement et de reproduction des cases se sont traduites par une augmentation du nombre d'enregistrements représentant l'ensemble des prestataires de l'AC. Initialement, l'ensemble de données receveur contenait 10,381 enregistrements de ce type alors qu'après reproduction il y en avait 31,585. Cet accroissement de l'ensemble de données avait pour but d'assurer l'utilisation complète des données récapitulatives sur l'AC disponibles dans l'échantillon de 1%.

Dans les cases, les enregistrements receveurs et de l'AC appariés ont été définis comme ceux qui avaient le même rang dans les deux ensembles de données. Le classement des enregistrements a été fait d'après les prestations d'AC reçues (en dollars).

6.3 Évaluation

Le tableau 3 nous donne une idée du succès de l'appariement. La corrélation entre les prestations déclarées lors de l'EFC et les prestations correspondantes (appariées) provenant de l'ensemble de données donneur de l'AC indique la 'précision' de l'appariement, puisque les rangs des prestations plutôt que les prestations elles-mêmes ont été utilisés lors de l'appariement. Les quartiles des écarts représentent les limites de 25%, 50% et 75% de la distribution des écarts entre les prestations déclarées pour l'EFC et celles qui figurent dans les dossiers de l'AC.

Tableau 3

Comparaisons entre enregistrements appariés de l'AC

Distributions par province et par sexe ainsi que pour le Canada.
i) n - Nombre d'enregistrements avant la reproduction
ii) r - Corrélation entre les prestations d'AC pour l'ensemble de données receveur et l'ensemble de données donneur ($)
iii) quartiles des écarts - [receveur ($) - donneur ($)]

Province/sexe	n		r	Quartiles des écarts		
	Receveur	AC		25%	50%	75%
T.-N. - Hommes	795	929	0.953	-192	140	417
Femmes	445	549	0.925	-270	-14	232
Î.-P.-É. - Hommes	241	246	0.631	-1,159	-290	789
Femmes	210	213	0.871	-363	11	531
N.-É. - Hommes	496	787	0.931	-271	45	528
Femmes	294	528	0.919	-197	-38	147
N.-B. - Hommes	604	798	0.941	-531	-45	589
Femmes	390	573	0.905	-102	158	669
QUÉ. - Hommes	1,116	5,471	0.970	-162	86	341
Femmes	784	3,961	0.958	-112	103	324
ONT. - Hommes	787	4,990	0.960	-149	36	207
Femmes	687	3,837	0.953	-110	74	306
MAN. - Hommes	343	611	0.932	-360	-69	294
Femmes	272	508	0.866	-115	-49	496
SASK. - Hommes	369	548	0.918	-239	231	489
Femmes	283	394	0.954	-83	75	311
ALB. - Hommes	691	1,648	0.946	-88	68	448
Femmes	482	1,072	0.951	-174	16	264
C.-B. - Hommes	625	2,281	0.953	-112	186	470
Femmes	467	1,638	0.954	-185	68	461
CANADA	10,381	31,582	0.953	-155	69	352

Dans la majorité des cas, les écarts entre les prestations déclarées dans l'ensemble de données receveur ainsi que les prestations tirées des dossiers administratifs de l'AC sont relativement faibles. On peut s'attendre à des écarts aussi élevés que 255 dollars, car ces écarts pourraient représenter un paiement de prestations d'AC pour une seule semaine (c.-à-d. l'écart minimal dans le nombre de semaines de prestations). De plus, les différences sont faibles par rapport aux niveaux médians des prestations, qui étaient de $2,972 pour les hommes et de $2,050 pour les femmes, au niveau national.

Il est prévu que certains enregistrements receveurs peuvent représenter des réponses biaisées et que d'autres peuvent contenir des éléments de prestations qui ne sont pas inclus dans les données ou le modèle de l'AC (par ex., les prestations de formation). Si c'était le cas, l'erreur dans l'ensemble de données receveur représenterait alors un élément important des écarts dans les prestations.

Les corrélations sont élevées, sauf dans le cas de l'Î.-P.-É. où la reproduction des enregistrements ne pouvait permettre que de réaliser un faible avantage. S'il n'y a pas reproduction d'un nombre important d'enregistrements, les contraintes d'appariement relatives à l'âge et au genre de demande réduiront les corrélations marginales pour les prestations.

Une corrélation élevée ne découle pas nécessairement de la technique d'appariement. L'appariement des rangs correspondants assure une association monotone, mais pas nécessaire une association linéaire forte. Cette utilisation des rangs peut être interprétée comme l'appariement de quantiles correspondants d'échantillons indépendants. Ainsi, une association linéaire forte indique que les deux échantillons (l'échantillon receveur et l'échantillon donneur) proviennent de fonctions de densité qui ont la même forme.

Il est difficile de faire une évaluation directe additionnelle des résultats de l'appariement, car essentiellement tous les facteurs communs aux deux ensembles de données ont été utilisés pour l'appariement. Les données de l'AC constituent un prolongement et un remplacement des données receveuses dans lequel les variables de l'AC sont sans biais et cohérentes avec la structure du programme de l'AC.

7. REPRODUCTION DES ENREGISTREMENTS SUR LES MÉNAGES

Des copies des enregistrements des ménages de l'EFC sont créées dans trois situations: 1) lors de l'imputation de données sur l'impôt pour des personnes à revenu élevé, 2) lors de l'appariement stochastique des données de l'AC et 3) lors de la création de personnes âgées synthétiques résidant dans un établissement.

Dans le cas des données sur l'impôt ou sur l'AC, les enregistrements sur les ménages sont reproduits afin d'utiliser le plus possible la gamme étendue de données administratives disponibles. Il faut remarquer que dans ces deux cas, ce sont les copies des enregistrements des personnes qui sont produites tout d'abord. Puis, les enregistrements des autres personnes dans le même ménage sont aussi reproduits. Dans le cas où l'enregistrement de plus d'un membre du ménage est reproduit, il faut effectuer la reproduction d'autres enregistrements afin de faire en sorte que chaque membre du ménage soit représenté de la façon appropriée.

Les enregistrements des personnes âgées résidant dans un établissement sont créés en reproduisant directement les enregistrements de personnes âgées (c.-à-d. de 65 ans et plus) seules qui ne résident pas dans un établissement et qui ne sont pas actives. Cette population d'enregistrements donneurs est choisie parce que ces personnes sont celles qui sont le plus susceptibles de devenir des résidentes d'un établissement et que, par conséquent, leurs caractéristiques par rapport au revenu ressembleront le plus vraisemblablement à celles des personnes résidant dans des établissements. Les poids de ces enregistrements sont ajustés pour refléter les estimations de la population qui réside

dans un établissement selon l'âge, le sexe et la province (d'après des statistiques administratives sur le nombre de jours-lits dans un établissement).

8. IMPUTATION STOCHASTIQUE DES RENSEIGNEMENTS SUR L'IMPÔT SUR LE REVENU

Dans cette section, on décrira l'imputation stochastique, la méthode utilisée pour attribuer les renseignements sur l'impôt sur le revenu des particuliers aux enregistrements de la BDSPS. Les éléments de revenu et de déduction dont la liste suit ont été imputés dans la BDSPS à partir du Livre vert. Ce sont des éléments qui ne sont pas bien représentés dans les données de l'EFC (par ex. les gains en capital), qui en sont absents (comme les frais financiers) ou qui ne peuvent pas être obtenus facilement à l'aide d'un modèle (par ex. les déductions pour handicapés).

1. Autres frais relatifs à un emploi déductibles
2. Frais financiers
3. Frais de garde d'enfants déductibles
4. Dons de charité
5. Perte en capital des autres années déductible
6. Déduction pour handicapés
7. Cotisations syndicales et professionnelles
8. Déduction relative aux études pour les étudiants
9. Autres crédits d'impôt fédéral
10. Crédit d'impôt pour contributions politiques fédérales
11. Gains en capital imposables
12. Perte en capital lors de la disposition d'actions de CPCC
13. Crédit d'impôt fédéral à l'investissement
14. Frais médicaux calculés nets
15. Perte autre qu'en capital d'autres années déductible
16. Autres déductions du revenu net
17. Exemptions pour autres personnes à charge
18. Crédit d'impôt provincial
19. Cotisations totales à un REP et à un REER
20. Proportion du REER par rapport au total (REER + REP)
21. Frais de scolarité.

Ces éléments et d'autres qui peuvent être calculés facilement à partir des données disponibles (par ex. les exemptions personnelles) permettent le calcul complet du revenu imposable ainsi que de l'impôt à payer.

8.1 Les données de base

Les données de base utilisées pour l'imputation ont été tirées d'un échantillon de déclarations d'impôt des particuliers de 1984. Cet échantillon contenait 2.4% de toutes les déclarations d'impôt (380,419 déclarations), le même échantillon a été utilisé pour compiler les données publiées dans l'ouvrage **"Statistique fiscale"** (le Livre vert). L'échantillon est stratifié selon la source de revenu, la région géographique urbaine, la région géographique rurale, l'état relativement à l'impôt (imposable et non imposable) ainsi que le palier de revenu.

Les données de cet échantillon contiennent la majorité des renseignements qui figuraient sur les déclarations d'impôt sur le revenu des particuliers T1 de 1984 aux niveaux fédéral et provincial ainsi que sur les annexes. Cet échantillon ne possède pas de structure familiale explicite (c.-à-d. que les déclarations du chef de famille, de son

conjoint et de ses personnes à charge ne peuvent être analysées ensemble sous forme d'unité familiale identifiable).

8.2 Transformations des données

Afin que ces données sur l'impôt sur le revenu figurant dans le Livre vert puissent être jointes à l'échantillon receveur basé sur les données de l'EFC, il a fallu définir un ensemble de caractéristiques de classification communes. Les attributs dont la liste suit ont été choisis pour leur importance du point de vue des politiques et parce qu'ils étaient disponibles et avaient des définitions semblables dans les deux ensembles de données:

1. province d'imposition
2. groupe d'âge
3. sexe
4. état matrimonial aux fins de l'impôt
5. catégorie du revenu total (à l'exclusion des gains en capital)
6. catégorie pour le revenu d'emploi
7. enfants déclarés pour la déduction pour frais de garde d'enfants (dans l'ensemble de données de l'EFC, nombre d'enfants pour lesquels cette déduction peut être réclamée).

On suppose que les sous-échantillons définis par le classement recoupé de ces éléments ont des distributions qui diffèrent suffisamment pour mériter de conserver le caractère unique de ces distributions. La figure 4 donne un exemple de l'écart dans les gains en capital entre deux groupes de revenu. Une comparaison des dons de charité entre les mêmes groupes est donnée à la figure 5.

Avant l'imputation, il a fallu préparer l'ensemble de données receveur, et pour cela déterminer les déclarants potentiels pour l'impôt, établir l'admissibilité pour certains éléments visés (déductions relatives aux études, déductions pour frais de scolarité et déductions pour frais de garde d'enfants) et créer un système de classification parallèle dans l'ensemble de données receveur de la BDSPS et l'ensemble de données donneur du Livre vert.

On a d'abord utilisé un modèle du système d'impôt sur le revenu des particuliers (le même qui a été utilisé par la suite pour effectuer l'analyse de la politique) afin de déterminer les déclarants probables pour l'impôt ainsi que pour imputer l'état matrimonial aux fins de l'impôt. Par exemple, une personne mariée qui pouvait déclarer son conjoint comme personne à charge aurait été désignée comme mariée-imposée-mariée. Cette imputation s'imposait afin que l'imputation ne s'applique qu'à un univers qui ressemblait à celui de l'ensemble de données donneur.

Trois des éléments de déduction ont été traités spécialement en ce sens que l'admissibilité à ces déductions pouvait être déterminée sur l'ensemble de données receveur. On peut, à partir des renseignements de la base de données de l'EFC, déterminer si la personne peut réclamer la déduction relative aux études (si elle ou une personne à charge fréquente un établissement d'enseignement post-secondaire), la déduction pour frais de scolarité (si la personne fréquente un établissement d'enseignement post-secondaire) et la déduction pour frais de garde d'enfants (accordée aux conjoints à faible revenu avec des enfants de moins de 15 ans présents). Le fait de limiter l'imputation aux personnes qui ont droit à ces déductions assure un certain degré de cohérence interne dans les enregistrements synthétiques. Par exemple, c'est seulement pour les personnes avec des enfants que l'on imputera la déduction pour frais de garde d'enfants.

L'admissibilité à toutes les retenues et éléments de revenu imputés n'est malheureusement pas aussi simple à déterminer.

DISTRIBUTIONS DANS LE LIVRE VERT
GAINS EN CAPITAL SELON LE GROUPE DE REVENU

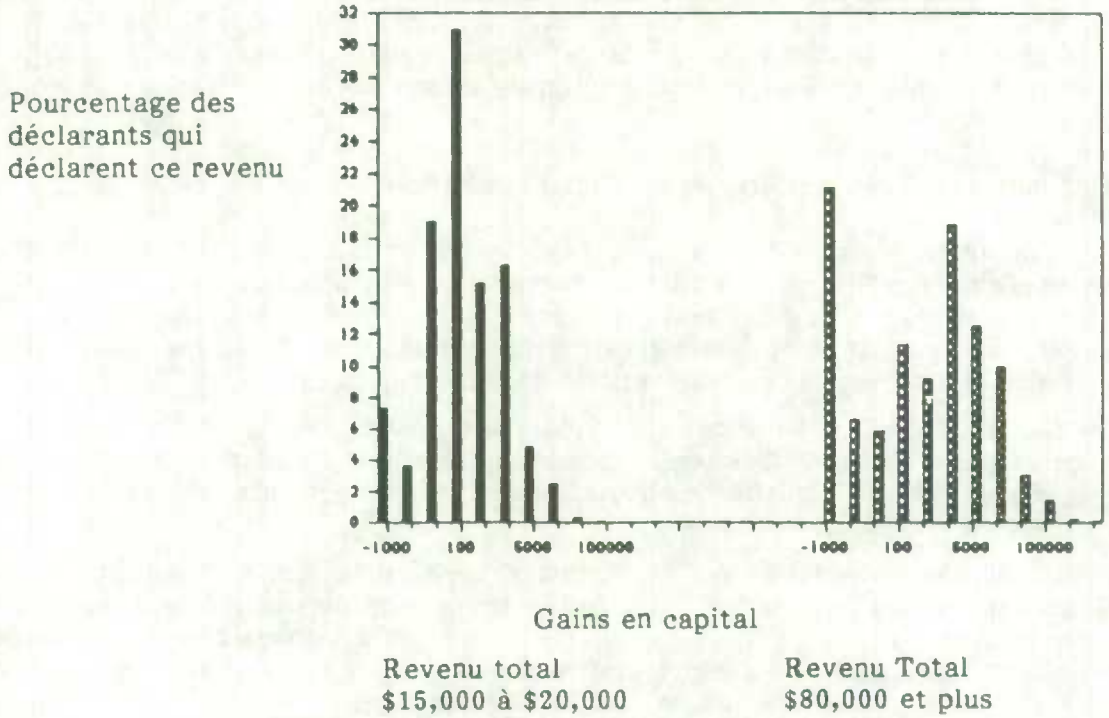


Figure 4. Distribution des gains en capital dans le Livre vert pour deux groupes de revenu.

DISTRIBUTION DES DONS DE CHARITÉ
POUR DEUX GROUPES DE REVENU

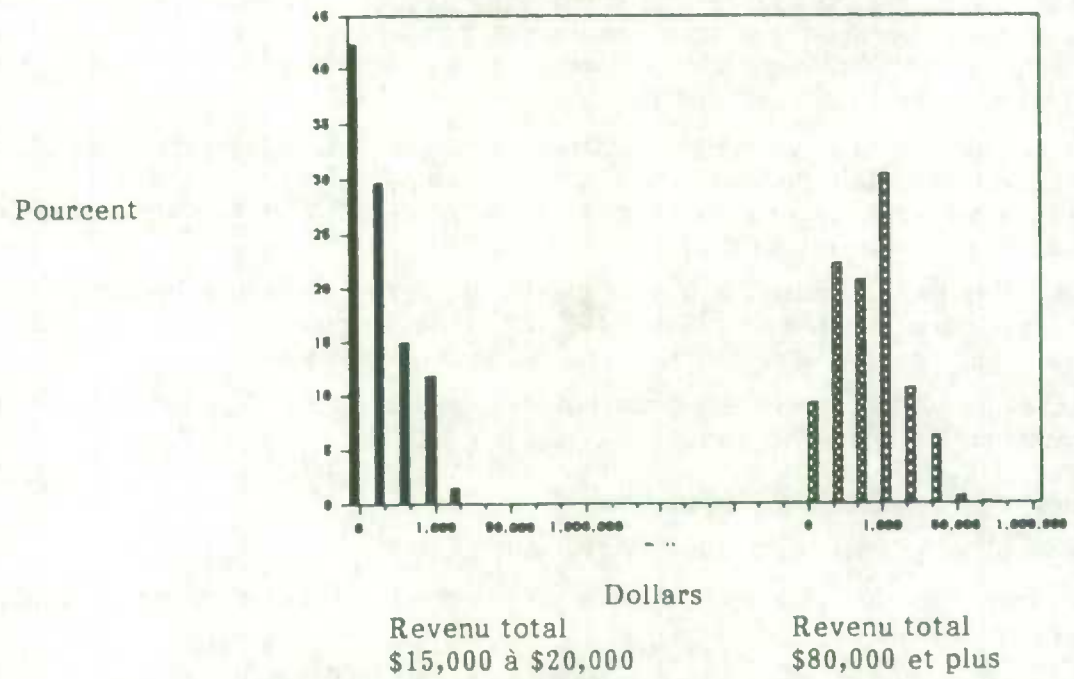


Figure 5. Distribution des dons de charité dans le Livre vert

La distribution conjointe des cotisations à un REP (Régime enregistré de pensions) et à un REER (Régime enregistré d'épargne-retraite) a posé un problème en ce sens que la Loi de l'impôt limite le total des cotisations à ces deux régimes à un certain montant ($3,500 en 1984). L'imputation séparée des cotisations aux deux régimes ne permettrait pas d'assurer que cette limite ne serait pas dépassée. Pour surmonter cette difficulté, on a imputé la somme des cotisations du déclarant au REP ainsi qu'au REER, puis les cotisations au REER seules comme une proportion de la somme des cotisations aux deux régimes.

8.3 Obtention de paramètres statistiques relatifs à la distribution

Un des buts de ce processus d'imputation est de reproduire la distribution des éléments de déduction et de revenu tels qu'ils sont représentés dans le fichier du Livre vert. Pour ce faire, il faut disposer d'une méthode qui permet de représenter une distribution arbitraire. Par exemple, la méthode devrait représenter aussi bien une distribution bimodale, une distribution tronquée et une distribution à longue queue.

Un autre facteur dont il a fallu tenir compte dans le choix de la méthode était la quantité de calculs à effectuer. Comme l'ensemble de données de base contient près de 400,000 enregistrements, les algorithmes utilisés pour produire ces statistiques devaient être raisonnablement efficaces.

La méthode choisie éventuellement consistait à représenter les distributions à une variable d'éléments particuliers tout d'abord par la proportion de ces éléments avec une valeur non nulle dans tout sous-groupe donné puis pour les sous-sous-groupes avec des valeurs non nulles. La fonction de densité était représentée par les limites de ses déciles avec un traitement spécial pour les queues des distributions.

Les mêmes procédures ont été appliquées indépendemment pour deux ensembles de paramètres statistiques: les pourcentages de personnes qui déclarent des revenus et les distributions. Ces paramètres statistiques ont été traités séparément parce que le pourcentage de personnes qui déclarent un revenu exigeait un critère de rejet moins strict et que, par conséquent, on pouvait utiliser des renseignements provenant d'un niveau de regroupement moins élevé. Le paramètre statistique relatif au pourcentage de personnes qui déclarent un revenu était conservé si la somme des poids pour la case était supérieure à 400 ou si le nombre d'enregistrements représentant une valeur non nulle était supérieur à 20. Si ces critères n'étaient pas remplis, c'est le paramètre statistique pour un niveau de regroupement plus élevé qui était utilisé.

Le critère utilisé pour les paramètres statistiques sur la distribution devait être plus rigoureux. La taille minimale pour une case était de 100 enregistrements, c.-à-d. que si une case ne contenait pas au moins 100 enregistrements non nuls, l'on ne calculait pas de paramètres statistiques pour cette case.

Pour chaque élément à imputer, il a fallu classer les enregistrements de déclarations d'impôt sur le revenu, au nombre de près de 400,000, dans des cases pertinentes (par ex., groupe de revenu par âge, par état matrimonial, par sexe, par province).

Pour chaque case, les paramètres statistiques suivants ont été calculés quand l'échantillon était suffisamment important:

- valeurs pour les limites du 1er au 9^{e} déciles,
- la moyenne du premier et du dernier déciles,
- la moyenne des cinq valeurs les plus élevées et celle des cinq valeurs les moins élevées et
- le pourcentage, dans la case, des personnes qui ont déclaré une valeur non nulle pour l'élément.

Ces paramètres statistiques conviennent bien pour représenter une distribution arbitraire et ils sont simples à calculer.

Pour des raisons de confidentialité, on n'a pu utiliser les vraies valeurs maximales et minimales dans une case. C'est la moyenne des cinq valeurs les plus élevées et celle des cinq valeurs les moins élevées dans la case qui ont été utilisées comme substituts.

Les mêmes paramètres statistiques ont alors été produits pour des regroupements de cases, dans ce cas, pour le groupe de revenu par âge, par état matrimonial, par sexe, par région. Si l'on forme cinq régions à partir des dix provinces, cela fait augmenter le niveau de regroupement et par conséquent le nombre moyen de personnes dans une case. Il y aura alors plus de cases qui fourniront des ensembles de données valides pour les statistiques sur la distribution.

Idéalement, toutes les valeurs seraient imputées à partir du niveau de regroupement le moins élevé. Cependant, comme dans beaucoup de cas les données élémentaires sont peu nombreuses, cela est rarement possible. Par exemple, les autres frais relatifs à un emploi déductibles sont concentrés dans les groupes à revenu plus élevé et les cases de cette région contiendraient beaucoup de données. Dans les groupes à revenu moins élevé, il y a moins de données dans ces cases et elles sont souvent vides.

Pour remplir les cases qui sont vides ou contiennent une faible quantité des données recherchées, on a eu recours à la substitution de paramètres statistiques provenant de niveaux de regroupement supérieurs. Si, par exemple, la case représentant la classification suivante:

- groupe de revenu	$35,000 à $39,999
- groupe d'âge	25 à 35 ans
- état matrimonal	célibataire, considéré marié aux fins de l'impôt
- sexe	femme
- province	Québec

était vide ou rejetée à cause du critère portant sur le nombre d'enregistrements qu'elle contient, on lui substituerait des paramètres statistiques provenant du niveau de regroupement supérieur suivant:

- groupe de revenu	$35,000 à $39,999
- groupe d'âge	25 à 35 ans
- état matrimonal	célibataire, considéré marié aux fins de l'impôt
- sexe	femme

représentant ce groupe de revenu, ce groupe d'âge, cet état matrimonial et ce sexe pour tout le Canada. Si cette case aussi était vide ou contenait une faible quantité des données recherchées, on lui substituerait les paramètres statistiques provenant du niveau de regroupement supérieur suivant. Dans le pire cas, les paramètres statistiques pour une case proviendraient de tout l'échantillon, c.-à-d. tous les groupes de revenu, tous les groupes d'âge, tous les états matrimoniaux, les deux sexes et toutes les provinces.

Les paramètres statistiques résultants portant sur la distribution et le pourcentage de personnes qui déclarent des revenus sont non confidentiels car ils ne révèlent jamais les valeurs des données brutes. On obtient les valeurs extrêmes synthétiquement en calculant la moyenne des cinq valeurs les plus élevées et celle des cinq valeurs les moins élevées.

8.4 Imputation

L'utilisation de cet ensemble complexe de paramètres statistiques relatifs à la distribution obtenus à partir du Livre vert permet de reconstituer la même distribution des valeurs dans l'ensemble de données receveur. Pour chaque personne admissible de l'ensemble de données receveur, on tire une valeur synthétique d'une distribution représentant les déclarations d'impôt d'un groupe semblable de personnes.

Pour obtenir les valeurs dans les huit déciles du milieu, on a supposé que la distribution entre les limites des déciles était uniforme.

Le premier et le dernier déciles ont été traités spécialement afin que la forme et la dimension des queues soient représentées avec précision. Il est essentiel de préserver la queue de la distribution si l'on veut maintenir les moyennes et totaux globaux, particulièrement pour les éléments avec des distributions à longue queue comme les gains en capital ou les pertes d'entreprise.

Lors des imputations pour le premier et le dernier déciles, les valeurs sont tirées en supposant une distribution de Pareto afin de produire une queue avec la forme appropriée. La moyenne du dernier décile est conservée, ce qui conserve donc la dimension totale de la queue. Les valeurs extrêmes sont tronquées à la moyenne des cinq valeurs les plus élevées dans le groupe. La même procédure a été appliquée à la queue inférieure des distributions.

8.5 Évaluation

Les figures 6 et 7 constituent des exemples des résultats du processus d'imputation. Elles sont toutes deux groupées au niveau de tout l'échantillon.

GAINS EN CAPITAL
VALEURS DU LIVRE VERT PAR OPPOSITION AUX VALEURS IMPUTÉES

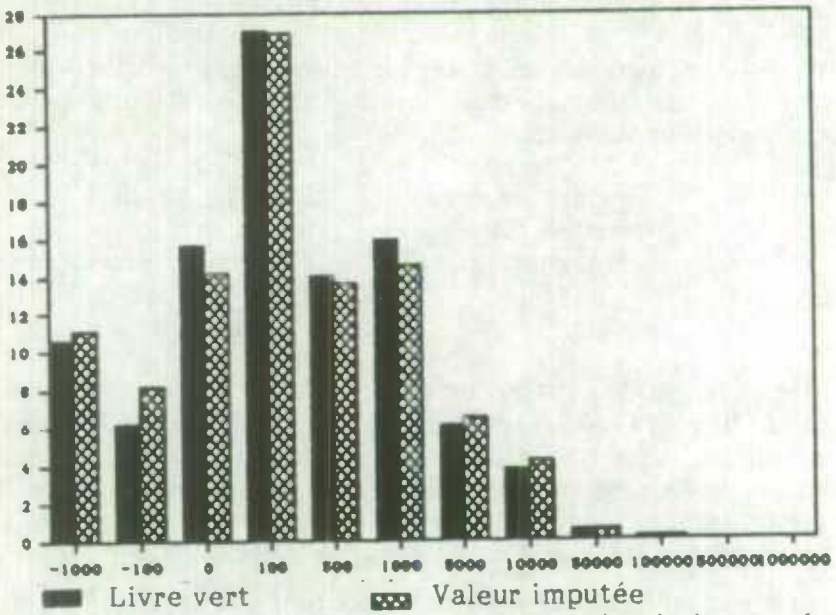


Figure 6: Distributions des gains en capital avant/après imputation

DONS DE CHARITÉ
VALEURS DU LIVRE VERT PAR OPPOSITION AUX VALEURS IMPUTÉES

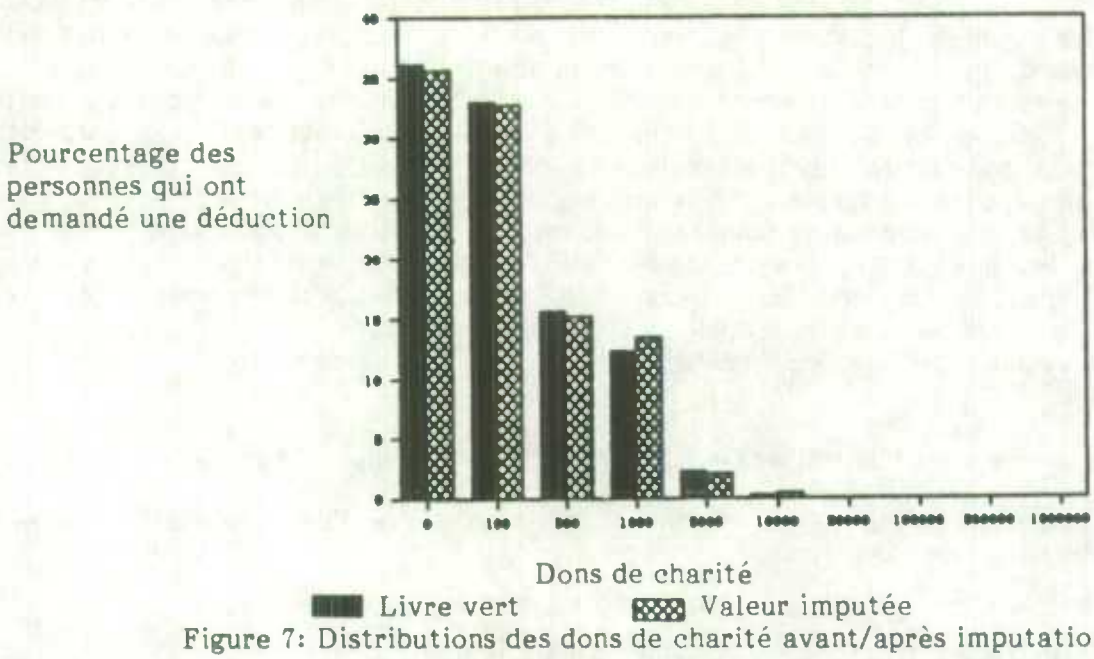


Figure 7: Distributions des dons de charité avant/après imputation

Cette méthode d'imputation tend à utiliser toute la richesse des données de base pour regénérer des distributions plausibles dans l'ensemble de données receveur. Les distributions globales ont du sens mais souvent les cas particuliers n'en ont pas. Par exemple, puisque les gains en capital sont imputés selon le revenu total, l'âge, le sexe et la province, il n'est pas impossible que l'on impute des gains en capital pour un bénéficiaire de l'aide sociale. Dans ce cas, le bénéficiaire de l'aide sociale est traité exactement comme un fermier qui prend sa retraite après avoir vendu sa ferme et qui a reçu plusieurs centaines de milliers de dollars en gains en capital.

Un autre problème que présente cette méthode est que les distributions conjointes sont perdues dans la mesure où les variables de classification ne les expliquent pas. Dans les cas simples, il y a une bonne corrélation entre la majorité des éléments de déduction et le revenu et ce dernier est normalement un critère de classification important. Quand l'intercorrélation entre les éléments (par ex. les cotisations au REP et au REER) est plus importante que leur corrélation avec le revenu, les corrélations sont perdues à moins que l'on ne modifie la méthode.

Un des résultats de la perte de corrélation entre les éléments de déductions est que les personnes, plus particulièrement celles du groupe à revenu élevé, de l'ensemble de données receveur semblent ne pas optimaliser leur situation par rapport à l'impôt. Comme le groupe des personnes à revenu élevé est composé de toutes les personnes dont le revenu dépasse $80,000, une personne dont le revenu total est de $90,000 a la même probabilité de se voir imputer une déduction d'un million de dollars qu'une autre personne dont le revenu est de $2 millions.

Dans la prochaine version de la BDSPS, on pourra surmonter ce problème en ayant recours à la méthode de regroupement des micro-enregistrements pour imputer plus d'éléments de déduction pour les membres du groupe à revenu élevé. Cela aurait pour effet de préserver les corrélations entre le revenu et les déductions ainsi que les corrélations entre les déductions.

9. IMPUTATIONS DES DONNÉES RECUEILLIES LORS DE L'ENQUÊTE SUR LES DÉPENSES DES FAMILLES

L'emploi des données sur les dépenses des familles vise à appuyer les simulations basées sur les coûts du logement (par ex. l'aide sociale), les simulations portant sur les coûts des frais de garde d'enfants ainsi que les simulations des taxes à la consommation. A cause du nombre limité d'enregistrements que contient l'ensemble de données sur les dépenses des familles, on a décidé d'effectuer trois imputations distinctes pour permettre une adaptation spécifique des catégories de classification à la nature du vecteur d'éléments de dépense à apparier. Par exemple, les dépenses d'un ménage pour la garde d'enfants dépendent beaucoup du nombre d'enfants et de la situation des parents vis-à-vis de l'activité et, à ce titre, ces variables devraient être les principales variables de classification pour tout appariement. Par contre, la corrélation pour les frais de logement est plus forte selon le nombre de pièces et le mode d'occupation du logement et une classification selon le nombre d'enfants ne serait pas très utile pour améliorer l'appariement.

Chacune de ces trois imputations comprenaient quatre étapes principales.

- La construction d'une base de données nationale pour l'Enquête sur les dépenses des familles de 1984
- Le choix/le groupement des éléments de dépense pour imputation
- Le choix/la construction de variables d'appariement
- L'appariement stochastique (reproduction pondérée)

9.1 Gonflement des données de l'Enquête sur les dépenses des familles de 1982

C'est en 1982 que la dernière enquête sur les dépenses des familles, menée partout au Canada, a été effectuée. L'enquête de 1984, qui correspond à la période de l'ensemble de données receveur de l'EFC, était limitée à un échantillon de 17 villes. La première étape de l'appariement a consisté à préparer une version pour l'ensemble du Canada en 1984; pour ce faire, il a fallu gonfler les valeurs des enregistrements de 1982 qui ne correspondaient pas aux 17 villes choisies lors de l'enquête de 1984 pour obtenir leurs valeurs pour 1984. L'"augmentation" de tous les éléments monétaires d'un enregistrement donné des dépenses des familles de 1982 a été effectuée simplement à l'aide du même indice plutôt qu'à l'aide d'indices distincts pour l'IPC et le revenu. Le choix de cette procédure simple a été amené par l'exigence du modèle des taxes à la consommation pour lequel il faut qu'une identité comptable complète soit assurée pour la structure du revenu, des dépenses et de l'épargne d'un ménage.

Cette méthode suppose que les structures de la dépense des ménages demeurent constantes, elle évite donc les hypothèses implicites de réaction comportementale aux variations de prix. Cette hypothèse est appuyée par une analyse des changements dans les proportions des dépenses totales consacrées à des éléments de dépense particuliers entre l'échantillon de 1982 et celui de 1984 pour les 17 villes. Les écarts entre tous les éléments n'ont pas dépassé un point de pourcentage. Les plus grands écarts ont été une augmentation d'un point de pourcentage dans le taux d'intérêt hypothécaire exprimé comme pourcentage des dépenses totales (de 4.7 à 5.7), ainsi qu'une diminution de .6 point de pourcentage dans les achats d'automobiles et de camions (de 5.4 à 4.8).

Les indices ont été calculés individuellement pour chacun des enregistrements de l'EDF de 1982 qui ne provenaient pas d'une des 17 villes utilisées pour l'enquête de 1984. Ces indices ont été basés sur la croissance moyenne par genre de famille de chacune des

six sources de revenu déclarées lors de l'Enquête sur les finances des consommateurs. L'indice utilisé pour un ménage est la moyenne pondérée des six taux de croissance individuels pour leur type de famille, où le poids utilisé était la proportion du revenu provenant de chacune des six sources.

9.2 Détermination des variables d'imputation et d'appariement

Le tableau 4 présente un résumé des variables d'imputation ainsi que des variables d'appariement utilisées pour chacun des trois appariements stochastiques. Les chiffres entre parenthèses représentent le nombre de niveaux de classification.

Tableau 4

Variables et classifications utilisées pour les appariements relatifs à l'EDF

	Logement (126)	Garde d'enfants (36)	Vecteur de dépenses (390)
Variables imputées	Loyer Intérêt hypothécaire Impôts fonciers Primes d'assurance Services publics Réparations Autres coûts du logement Valeur du logement Solde de l'hypothèque	Frais de garde d'enfants	"Épargne" Autres rentrées monétaires Revenu du ménage Différences pour solder les comptes Vecteur de dépenses (50) (voir l'Annexe A)
Variables d'appariement	Mode d'occupation du logement Nombre de pièces Catégorie d'habitat Région géographique	Genre de famille Catégorie de travailleurs No. d'enfants (de 0 à 4 ans) No. d'enfants (de 5 à 15 ans) Revenu du ménage	Revenu (variable discrète 6) Genre de famille (5) Mode d'occupation du logement (3) Âge du chef (4) Revenu du ménage Sexe du chef (2) Région géographique (5) Taille de la famille (2) Nombre d'enfants (3) Catégorie d'habitat (2) Revenu (variable continue)

Les variables utilisées pour l'appariement relatif au logement ont été choisies et groupées de manière qu'il soit possible de faire des estimations des principaux coûts et du loyer imputé. La principale utilisation visée était la modélisation des paiements d'aide sociale et, ensuite, la modélisation des crédits d'impôt provincial. Le niveau de groupement élevé reflète le caractère très approximatif du modèle d'aide sociale qu'on peut obtenir quand on manque d'autres données sur l'admissibilité et les niveaux de prestations. Par exemple, les tests d'admissibilité en fonction des actifs sont faits pour une période de cinq ans et les niveaux de prestation sont basés sur l'assurance-incendie alors que l'EDF ne recueille que le montant total de l'assurance pour le logement.

Les frais de garde d'enfants sont composés des frais de garde au foyer ou à l'extérieur du foyer ainsi que des frais du jardin d'enfants. Cette définition vise à suivre les mesures législatives fédérales existantes relatives aux déductions pour les frais de garde d'enfants. Aucune tentative n'a été faite pour exclure les dépenses pour lesquelles une déduction aux fins de l'impôt ne serait pas accordée parce que le contribuable n'a pas de reçus. D'autres éléments comme les vêtements pour bébé ou d'autres variables que l'on pourrait vouloir utiliser quand on établit un modèle qui peut traiter une définition plus étendue des coûts n'ont pas été imputés.

Le choix et le regroupement des variables sur le revenu et sur les dépenses, recueillies lors de l'EDF, pour le vecteur de dépenses ont été basés sur la structure et la composition de la dimension des dépenses personnelles dans les tableaux canadiens d'entrées-sorties et en tenant compte des exigences du modèle des taxes à la consommation. Les dépenses pour lesquelles il existe des impôts et des droits indirects ont été incluses dans la catégorie d'entrées-sorties correspondante pour les dépenses personnelles. Les variables pour lesquelles il n'existe pas d'impôt indirect ou pour lesquelles l'impôt indirect est indéterminé ont été placées dans une catégorie résiduelle (par ex. les commissions des agents immobiliers). Des variables additionnelles ont aussi été incluses dans le vecteur (par ex. le revenu, les impôts, l'épargne) afin de compléter une identité qui permet diverses options en matière de simulation. Au cours de l'appariement on n'a pas fait de corrections pour tenir compte des différences conceptuelles entre les données de l'EDF et celles du système de comptabilité nationale sur lequel les tableaux d'entrées-sorties sont basés.

La détermination des variables d'appariement a été limitée par la disponibilité de variables semblables tant dans l'ensemble de données receveur que dans l'ensemble de données donneur. A partir de cet ensemble limité, on a effectué des analyses particulières pour déterminer le choix et la configuration optimaux des variables d'appariement pour les trois appariements. Les techniques utilisées pour déterminer les variables incluaient la corrélation, l'analyse factorielle et des tests de la différence entre moyennes. Quatre critères interdépendants principaux ont guidé le choix et la création des cases d'appariement:

Niveaux des dépenses: Il devrait y avoir une relation étroite entre les variables utilisées pour classer les ménages, tant au niveau des dépenses qu'au niveau de la distribution pour des produits particuliers.

Catégories de dépenses: Les cases doivent être créées de façon à ce qu'on ne puisse attribuer des coûts qu'aux populations appropriées. Par exemple, les couples sans enfant ne devraient pas avoir de dépenses de garde d'enfants et les femmes seules ne devraient pas avoir de dépenses considérables en vêtements pour hommes.

Catégories pour les déclarations: Les cases devraient refléter le plus possible les catégories qui seront utilisées pour les dernières déclarations. Par exemple, la BDSPS ainsi que le modèle seront vraisemblablement employés pour effectuer des analyses comparatives de provinces et de régions différentes, de niveaux de revenu et de genres de famille différents, ces variables devraient donc être utilisées dans le processus d'appariement.

Taille de l'échantillon dans les cases: Une condition essentielle lors de la création des cases était que chacune d'entre elles contienne au moins cinq observations tant dans la base de données receveuse que dans la base de données donneuse. En pratique, certaines cases contenaient un très grand nombre d'observations de sorte que le tri final effectué d'après le revenu a été un élément clé de l'ajustement pour les trois appariements.

On a tenu compte de l'analyse et des critères lors du choix et de la création des groupes finals ainsi que pour l'établissement des dernières priorités. L'organisation hiérarchique des variables a été faite manuellement et avec une flexibilité qui permettait

différentes séparations pour différents genres de cases. Ainsi, pour les coûts du logement au deuxième niveau de la hiérarchie (nombre de pièces) les propriétaires avec ou sans hypothèque ont été classés en groupes selon le nombre de pièces: moins de six pièces, six et sept pièces et huit pièces ou plus alors que les locataires ont été classés selon les groupes suivants: moins de quatre pièces, quatre pièces et cinq pièces ou plus.

9.3 Appariement stochastique

L'appariement stochastique des enregistrements a été effectué au niveau du ménage et il a exigé la reproduction des enregistrements de l'EDF seulement. Afin d'utiliser le plus possible les données de l'EDF sans avoir à reproduire les enregistrements de l'EFC, on a créé les cases d'appariement de telle sorte que la taille de l'échantillon était toujours plus petite dans les cases de l'EDF que dans les cases correspondantes de l'EFC. Comme l'ensemble de données receveur, même sans reproduction des enregistrements, était près de quatre fois plus grand que l'ensemble de données donneur, il est rarement arrivé que l'on ait eu à redéfinir une case parce que la case de l'EFC avait été épuisée avant la case correspondante de l'EDF. L'appariement a pris la forme d'une reproduction pondérée des enregistrements de l'EDF et d'opérations visant à forcer le nombre d'enregistrements de l'échantillon de l'EDF dans les cases à être égal au nombre d'enregistrements dans la case receveuse correspondante.

Pour effectuer la reproduction, on a calculé la probabilité pondérée que le ménage i se trouve dans la case j. Si l'on multiplie cette probabilité par la taille de l'échantillon dans la case receveuse moins la taille de l'échantillon dans la case donneuse, on obtient une estimation du nombre de fois qu'un ménage donné devrait apparaître dans l'ensemble de données receveur. Si l'on se contente d'arrondir ou de ramener à un nombre entier, par suppression des décimales, le nombre ainsi obtenu, l'erreur d'arrondi peut produire un résultat inexact pour le total du nombre d'enregistrements dans les cases receveuses. Pour corriger cette erreur, on calcule un total cumulatif des fréquences des cases receveuses (D).

$$D_{ij} = \sum_{k=1}^{i} \left(\left(\frac{W_{ij}}{\sum_{i=1}^{n} W_{ij}} \right) \times (N_j^t - N_j^d) \right)$$

Où: i = le i^e ménage

j = la j^e case d'appariement

W = le poids de l'enregistrement donneur de l'EDF

N^t = la taille de l'échantillon dans la case receveuse de la BDSPS

N^d = la taille de l'échantillon dans la case donneuse de l'EDF

Le nombre de fois où chaque enregistrement de l'EDF est ensuite reproduit correspond à la valeur arrondie du total cumulatif moins la valeur arrondie du total cumulatif de l'enregistrement précédent plus un. De cette façon, l'erreur d'arrondi est répartie dans toute la case, chaque enregistrement de l'EDF est assuré d'au moins un appariement et l'on atteint les totaux exacts pour les cellules receveuses.

Cette procédure sert surtout à préserver les distributions pondérées des données de l'EDF, du moins jusqu'à ce que les poids de la BDSPS soient associés avec elles. La

différence entre les poids pour l'EFC et pour l'EDF peut cependant créer des distorsions dans les distributions appariées.

9.4 Évaluation

Plusieurs tests visant à évaluer la qualité des résultats et à aider à l'analyse subséquente ont été effectués. Les distributions des dépenses groupées sont extrêmement semblables avant et après l'appariement. Les seules véritables sources d'écarts, tant au point de vue distribution que regroupement, sont attribuables aux différents poids (pour la BDSPS) maintenant associés aux données de l'EDF et à l'impact mineur sur les poids de l'EDF qui découle du fait qu'on a forcé chaque enregistrement à être reproduit au moins une fois. Il est difficile d'obtenir des totaux de contrôle repères pour la majorité des données relatives aux dépenses. Dans cet esprit, le test principal pour ces totaux regroupés a été de voir comment les totaux après l'appariement se comparaient aux totaux pour l'EDF. Les écarts entre les totaux pour les éléments particuliers imputés lors des appariements pour le logement et la garde des enfants étaient tous inférieurs à cinq pourcent. Le tableau 5 présente les résultats de l'appariement des vecteurs de dépenses.

Le tableau 5 montre les liens entre les totaux agrégés pour l'EDF, la BDSPS et les variables obtenues à l'aide du MSPS. La deuxième colonne montre les écarts en pourcentage entre les valeurs avant et après l'appariement des éléments de l'EDF. Comme on peut le voir, l'écart pour tous les totaux des variables est inférieur à quelques points de pourcentage, la différence étant surtout attribuable aux poids pour la BDSPS associés aux dépenses de l'EDF. Les différences pour solder les comptes sont de 17.2 pourcent inférieures parce qu'il ne s'agit pas en fait d'une dépense réelle mais plutôt de la divergence entre les rentrées et les sorties de fonds pour une famille. La troisième colonne montre l'écart en pourcentage entre les données de l'EDF ainsi que les données produites par le MSPS et/ou les variables imputées. Les différences les plus importantes sont dues aux corrections pour la sous-déclaration qui ont été faites au moyen d'imputations des distributions du Livre vert.

Chaque enregistrement de l'EDF a été reproduit six fois en moyenne pour chacun des trois appariements. Le nombre maximum de reproductions est de 28, 42 et 51 pour les appariements relatifs au logement, à la garde des enfants et au vecteur de dépenses respectivement. Dans 75 pourcent des appariements relatifs au vecteur de dépenses, les enregistrements ont été reproduits moins de douze fois.

La corrélation entre les revenus receveurs et donneurs était élevée, ces valeurs étant de .91 et .96 pour les imputations relatives au logement et à la garde des enfants. La corrélation est inversement proportionnelle au nombre de cases parce que le tri final est effectué d'après le revenu. C'est pourquoi la corrélation pour l'appariement avec les vecteurs de dépenses était plus faible (.86). Les écarts dans les revenus ont eu tendance à être supérieurs dans les queues de la distribution où la majorité des changements avaient été causés, dans la distribution receveuse, par la CPI et l'ajustement pour les revenus élevés. Globalement, dans 90 pourcent des cas, les différences dans le revenu des ménages individuels étaient inférieures à 15 pourcent (voir le tableau 5). Il est particulièrement important de comprendre cet ajustement à cause de ses effets sur les options du modèle de l'impôt pour divers produits ainsi que sur le lien a priori entre le revenu et les dépenses.

Tableau 5

Comparaisons des vecteurs de dépenses, éléments choisis

Catégories de revenus/de dépenses	EDF $ millions	BDSPS EDF	MSPS/ EDF
Aliments et boissons non alcoolisées	30,805	3.10	
Boissons alcoolisées	4,959	0.38	
Tabacs et produits connexes	3,453	3.69	
Vêtements pour hommes et pour garçons	4,462	-0.21	
Loyer imputé brut	19,021	-5.36	
Loyer payé brut	12,773	4.17	
Électricité	4,226	1.26	
Autres combustibles	2,115	8.89	
Appareils ménagers durables	3,292	-0.55	
Biens semi-durables	3,627	-0.98	
Biens non durables	4,301	1.22	
Services domestiques	1,121	-9.25	
Autres services ménagers	2,000	0.40	
Soins médicaux	1,381	1.65	
Soins hospitaliers	86	-3.46	
Médicaments et objets pharmaceutiques divers	1,657	0.87	
Automobiles neuves et usagées	10,014	-1.53	
Pièces et réparations d'automobiles	4,458	3.63	
Frais de transport	3,086	1.43	
Communications	3,583	1.98	
Matériel récréatif, articles de sport et équipement de camping	7,514	-4.28	
Livres, revues et papeterie	2,261	1.50	
Services récréatifs	4,412	0.66	
Bijoux, montres et réparations	1,033	-3.41	
Soins personnels	2,333	-0.49	
Cotisations syndicales et professionnelles	985	2.75	1.80
Impôt personnel	45,148	-5.19	14.77
Cotisations d'assurance-chômage	2,924	0.81	17.25
Paiements de pension de retraite	6,108	0.44	18.88
Éléments non affectés de l'EDF	2,525	8.99	
Changement net dans les actifs/ les dettes	16,021	-5.49	
Primes totales d'un RÉER	3,492	-7.56	36.76
Autres rentrées d'argent	5,612	4.14	
Différence pour solder les comptes	1,245	-17.20	
Revenu total de l'unité de dépense	272,714	-0.87	6.99

Tableau 6

Comparaisons des revenus pour les appariements réalisés à l'aide des vecteurs de dépenses de l'EDF

	Limites des percentiles du revenu BDSPS/EDF						
Quintile	**1**	**5**	**25**	**50**	**75**	**95**	**99**
1	0.010	0.5985	0.918	0.991	1.055	1.319	1.664
2	0.845	0.887	0.954	0.989	1.021	1.078	1.129
3	0.898	0.938	0.980	1.003	1.030	1.074	1.094
4	0.916	0.947	0.978	1.000	1.022	1.072	1.101
5	0.855	0.900	0.961	0.998	1.037	1.130	1.207
>$80,000	1.003	1.014	1.074	1.181	1.418	2.130	3.418
Tous	0.555	0.866	0.965	0.999	1.035	1.154	1.572

Ce tableau montre la distribution des écarts dans les rapports entre le revenu dans la BDSPS et le revenu pour l'EDF après appariement à l'aide des vecteurs de dépenses, par quintiles du revenu. Les enregistrements avec un revenu de plus de $80,000 étaient un sous-ensemble du cinquième quintile. Dans tous les quintiles, l'écart médian entre les rapports pour les revenus avant et après appariement était inférieur à un point de pourcentage sauf pour le groupe des personnes ayant un revenu de plus de $80,000. Cela est dû au fait que le revenu maximum pour l'EDF est de l'ordre de $250,000 alors que pour la BDSPS il est d'environ $11 millions à cause du rajustement pour les revenus élevés. Certaines options du modèle des taxes à la consommation attribuent des impôts indirects d'après la dépense imputée en dollars et, à cause de cela, il devrait exister un lien étroit avec le revenu.

10. CONCLUSIONS

La base de données ainsi que le modèle de simulation de politique sociale, tels que décrits, sont les prototypes d'une fonction complexe d'analyse en matière de politique gouvernementale. Le processus de création de la base de données a subi deux cycles complets pour deux années différentes avec la quantité énorme de données que cela suppose. Les fonctions de modélisation ont aussi été élaborées en entier dans au moins deux environnements informatiques ainsi que pour plusieurs options importantes en matière de politiques qui ne font pas encore partie de la loi. Tous ces travaux se sont produits depuis l'automne 1984.

Pour vérifier la viabilité du principe de la BDSPS et du MSPS, il a été nécessaire d'aller de l'avant et de faire de nombreuses hypothèses en plus de prendre des racourcis qui ont rendu possible la création d'un produit susceptible de subir des essais. Ce processus a eu des avantages inattendus: il a permis de faire des suggestions aux personnes qui ont fourni les ensembles de données et d'isoler les lacunes dans les données dont on dispose pour formuler des politiques gouvernementales. De plus, le modèle a produit des résultats qui se sont déjà révélés utiles à plusieurs reprises pour l'établissement de politiques gouvernementales au Canada.

A court terme, on cherchera de nombreux raffinements méthodologiques du processus de création de la base de données et on en fera l'essai afin de corriger les lacunes et les inexactitudes dans les données. A plus long terme, il ne sera possible d'apporter des perfectionnements que par la collecte de données plus détaillées, plus précises et plus actuelles.

11. ANNEXE A:

Contenu de la BDSPS

a) Structure du ménage

Poids du ménage
Numéro d'ordre du ménage
Numéro d'ordre de la famille économique
Numéro d'ordre de la famille de recensement
Numéro d'ordre du particulier
Lien avec le chef de ménage
Lien avec le chef de la famille économique
Lien avec le chef de la famille de recensement
État du déclarant (imputé)
État relatif à l'union libre (imputé)
Nombre de reproductions de l'enregistrement de l'EFC
Nombre de reproductions de l'enregistrement de l'AC
Nombre de reproductions de l'enregistrement de l'EDF
Nombre de reproductions des données sur la garde des enfants tirées de l'EDF

b) Caractéristiques sociales du particulier

Province
Catégorie d'habitat
Âge (au printemps 1985)
Sexe
État matrimonial
Profession
Activité économique dans laquelle s'exerce la profession
Nombre d'années depuis l'immigration
Situation vis-à-vis de l'activité (la semaine dernière)
Niveau de scolarité
Genre d'établissement d'enseignement
État relatif à l'instruction
Nombre de semaines travaillées l'an dernier
Nombre de semaines de chômage l'an dernier
Principale activité ne relevant pas du marché du travail l'an dernier

c) Composantes du revenu - Revenu gagné

Revenu d'emploi
Revenu provenant d'un travail autonome - non agricole
Revenu provenant d'un travail autonome - agricole
Revenu provenant de chambreurs
Revenu provenant d'une pension de retraite
Autre revenu en argent - imposable
Autre revenu en argent - non imposable
Revenu en intérêts
Revenu en dividendes
Gains/pertes en capital
Autre revenu de placements

d) Composantes du revenu - Transferts

Allocations familiales
Prestations de sécurité de la vieillesse (SV)
Prestations du supplément de revenu garanti (SRG)
Prestations versées à titre d'allocation au conjoint
Prestations complémentaires au SRG versées par une province
Régime de pensions du Canada/Régime de rentes du Québec
Autres transferts - imposables
Autres transferts - non imposables
Revenu provenant de l'aide sociale
Prestations d'assurance-chômage

e) Composantes du revenu - Déductions

Autres frais relatifs à un emploi
Cotisations à un REP
Cotisations à un REÉR
Cotisations professionnelles et syndicales
Frais de scolarité
Déductions pour frais de garde d'enfants
Frais de garde d'enfants - total pour le ménage
Pertes au titre d'un placement d'entreprise
Frais financiers
Autre déduction par rapport au revenu total
Autres exemptions personnelles
Déduction permise pour frais médicaux
Déduction pour dons de charité
Déduction pour handicapés
Déduction relative aux études
Pertes autres qu'en capital
Pertes en capital
Autres déductions par rapport au revenu net

f) Composantes du revenu - Crédits d'impôt

Crédit d'impôt pour enfants
Contributions politiques fédérales
Investissement
Tous les autres crédits d'impôt fédéraux
Tous les crédits d'impôt provinciaux

g) Composantes du revenu - Impôts

Primes d'assurance-chômage
Cotisations au RPC ou au RRQ
Impôt fédéral sur le revenu net
Impôt provincial sur le revenu net

h) Dépenses du ménage

Aliments et boissons non alcoolisées
Boissons alcoolisées
Produits du tabac et articles pour fumeurs
Vêtements pour hommes
Vêtements pour garçons
Vêtements pour femmes
Vêtements pour filles
Vêtements pour bébés
Chaussures et réparation de chaussures
Loyer imputé brut
Loyer payé brut
Autres dépenses relatives au logement
Électricité
Gaz canalisé
Autres combustibles
Meubles, tapis et revêtements de sol
Appareils ménagers durables
Biens semi-durables
Biens non durables
Blanchissage et nettoyage à sec
Services domestiques
Autres services ménagers
Soins médicaux
Soins hospitaliers
Autres soins médicaux
Médicaments et articles divers
Automobiles neuves et usagées
Pièces et réparations d'automobiles
Essence, huile et graisse
Autres services relatifs aux automobiles
Frais de transport local et de banlieue
Transport interurbain
Communications téléphoniques
Toutes les autres communications
Matériel récréatif, articles de sport et équipement de camping
Livres, revues et papeterie
Services récréatifs
Services éducatifs et culturels
Bijoux, montres et réparations
Articles de toilette, produits de beauté, etc.
Soins personnels
Dépenses dans les restaurants et les hôtels
Intérêts sur les prêts personnels
Toutes les autres entreprises personnelles
Dons à des oeuvres de charité
Dons en argent et autres à des personnes résidant au Canada
Cotisations syndicales et professionnelles
Autres frais de fonctionnement (organismes sans but lucratif)
Impôt personnel
Primes d'assurance-chômage
Paiements pour des pensions de retraite (REP, RPC/RRQ)
Éléments non affectés de l'EDF
Changement net dans les actifs/dettes (à l'exclusion d'un REÉR)
Cotisations à un REÉR - total
Autres rentrées d'argent
Différences pour solder un compte

i) Éléments récapitulatifs (Revenu)

Total des revenus d'emploi
Total des revenus d'investissement
Total des autres revenus gagnés
Total des revenus de transfert
Total des revenus en argent
Total des impôts
Revenu disponible total

j) Caractéristiques du logement

Mode d'occupation (s'applique aussi aux établissements)
Nombre de pièces
Nombre de chambres à coucher
Loyer payé
Intérêt sur l'hypothèque
Impôts fonciers payés
Assurance sur le logement
Services publics
Réparations et entretien
Autres coûts d'habitation
Valeur de la maison sur le marché
Solde de l'emprunt hypothécaire qui reste à rembourser

k) Données sur les demandes d'AC (demandeurs d'AC seulement)

Numéro d'ordre de la demande (1re ou 2e de l'année en cours)
Drapeau de réitérant
Genre initial de la demande de prestation
Drapeau de changement de genre de demande de prestation
Nombre de semaines de prestations (pour la demande courante)
Nombre de semaines de prestations (au cours des 52 dernières semaines)
Nombre de semaines de travail (avant la demande courante)
Rémunération hebdomadaire moyenne (avant la demande)
Pénalité pour départ volontaire (semaines)
Semaine au cours de laquelle la demande a été établie
Prestations versées au cours de l'année civile (1 ou 2 demandes)
Nombre de semaines de prestations versées au cours de l'année civile

SESSION IV: COMMUNICATIONS SOLLICITÉES

ÉVALUATION DE LA QUALITÉ

Présidente: N.P. Gendreau, Bureau de la Statistique du Québec

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

DONNÉES SUR LES PERSONNES AGÉES: COMPARAISONS DE DEUX SOURCES ADMINISTRATIVES

N.J. KOPUSTAS[1]

RÉSUMÉ

Compte tenu de l'augmentation de la population âgée au Canada, on s'intéresse davantage à des données actuelles et détaillées sur les personnes âgées. Ce document examine deux sources de données, à savoir le fichier de l'impôt sur le revenu des particuliers et celui du régime de pension de vieillesse et il compare leurs avantages et leurs inconvénients comme source de données sur les personnes âgées. Le document propose également des méthodes de combinaison des données provenant de ces deux sources afin d'obtenir des données plus complètes sur les personnes âgées.

1. INTRODUCTION

La situation des personnes âgées suscitant de plus en plus d'intérêt, la demande de données à leur sujet grandit de jour en jour. Pour répondre à cette demande, on devrait envisager l'utilisation de données administratives. Dans la fonction publique fédérale, c'est à l'application de deux grands régimes que l'on doit la plupart des données sur les personnes âgées. Il s'agit du régime fiscal (déclarations d'impôt des particuliers) et du régime de la Sécurité de la vieillesse. Dans cette étude, nous verrons comment nous pouvons exploiter ces deux sources de données pour dénombrer les personnes âgées, évaluer leur revenu et déterminer leur situation familiale (à savoir, si elles sont membres d'une famille ou si elles sont des personnes hors famille). En outre, nous signalerons quelques-unes des recherches qui pourraient être entreprises à ce chapitre.

2. IMPÔT SUR LE REVENU DES PARTICULIERS

2.1 Représentativité des données

Au Canada, un particulier est tenu de remplir une déclaration, pour une année d'imposition quelconque, s'il a des impôts à payer. Les déductions et exemptions du régime relèvent le seuil de revenu à partir duquel une personne a des impôts à payer. Tous les déclarants ont droit à une exemption personnelle de base ($4,180 en 1986). Certaines personnes peuvent réclamer l'exemption pour conjoint à charge ($3,660 en 1986). Les personnes âgées de 65 ans et plus ont droit à une exemption en raison d'âge de $2,610. Enfin, il y a la déduction ($1,000) relative au revenu de pensions de retraite (revenu autre que les prestations de la

[1] N.J. Kopustas, Statistique Canada, Division des données régionales et administratives, Immeuble Principal, Ottawa, Ontario. K1A 0T6

Sécurité de la vieillesse et du Régime de pensions du Canada ou du Régime des rentes du Québec). Les déductions ci-dessus, auxquelles aurait droit une personne mariée âgée de 65 ans recevant des prestations d'un régime de retraite privé, s'élèvent à $11,450. Mises à part les prestations annuelles de la Sécurité de la vieillesse (environ $3,500) que reçoivent presque toutes les personnes âgées de 65 ans et plus, un déclarant peut recevoir d'autres sources la somme supplémentaire de $7,950 et ne pas avoir d'impôt à payer. D'autres déductions, par exemple, pour intérêts et dividendes, gains en capital, frais médicaux et dons de charité viendraient aussi accroître le montant du revenu sur lequel aucun impôt ne serait perçu. Soulignons par ailleurs que les prestations supplémentaires du régime de la Sécurité de la vieillesse (c.-à-d. le Supplément de revenu garanti) ne sont pas imposables.

Malgré toutes les déductions ci-dessus, environ 61% des personnes âgées de 65 ans et plus produisent une déclaration (ce pourcentage est relativement constant depuis 1980).

Beaucoup de personnes produisent une déclaration pour recevoir un crédit d'impôt. Le plus connu, le crédit d'impôt pour enfants, est versé à la mère. Parmi les autres crédits, mentionnons les remboursements d'impôt, les remboursements de paiements en trop au Régime de pensions du Canada (RPC) et certains crédits d'impôt provinciaux. Notons cependant que peu de personnes âgées ont droit à ces crédits.

En 1986, l'administration fédérale offrait un nouveau crédit, le crédit de taxe sur les ventes, aux personnes à faible revenu (5% du montant du "revenu familial" excédant $15,000 est déduit du crédit de sorte qu'un couple marié dont le revenu est supérieur à $17,000 n'y a pas droit). L'examen des dossiers fiscaux de 1986 révèle, à première vue, qu'un nombre élevé de personnes âgées qui n'ont pas produit de déclaration pour l'année 1985 en ont remplie une pour 1986 afin de recevoir ce crédit fédéral de taxe sur les ventes. Le nombre total de déclarants a augmenté de 450,000, ce qui correspond à une hausse de 3% par rapport à 1985. Si le tiers des déclarants sont âgés de 65 ans et plus, la proportion de déclarants chez les personnes âgées pourrait en ce moment dépasser 65%.

2.3 Revenu

Presque toutes les sources de revenu sont assujetties à l'impôt mais certaines, qui rendent compte d'une part importante du revenu des personnes âgées, ne le sont pas. Parmi celles-ci, la principale est le Supplément de revenu garanti (SRG) versé aux personnes âgées à faible revenu. Toutefois, très peu de bénéficiaires du SRG remplissent une déclaration d'impôt précisément parce que leurs revenus sont bas.

Parmi les autres sources de revenu des personnes âgées qui ne sont pas assujetties à l'impôt, on compte les prestations d'invalidité, les prestations d'invalidité des anciens combattants, les allocations d'ancien combattant ainsi que les crédits d'impôt. Bref, bon nombre des paiements de transfert ne sont pas inclus dans le calcul du revenu imposable et c'est pour cette raison que les dossiers fiscaux ne constituent pas une source complète de données sur le revenu des personnes âgées. Toutefois, étant donné qu'aux fins du calcul du crédit de taxe sur les ventes, ces sources de revenu doivent être déclarées, les dossiers fiscaux devraient dorénavant permettre une plus juste évaluation du revenu des personnes âgées.

2.4 Composition de la famille

Statistique Canada a constitué, à titre expérimental, une base de données sur la famille à partir des dossiers fiscaux (voir Auger, 1987). La définition de la famille utilisée est celle de la famille de recensement ou de la famille nucléaire, c'est-à-dire, un couple marié avec ou sans enfants jamais mariés ou un parent seul avec enfants jamais mariés. Les données sont créées suivant une série d'étapes au cours desquelles on isole et on couple les membres d'une famille ayant produit une déclaration (conjoints et enfants de moins de 30 ans). Pour ce qui des membres n'ayant pas rempli de déclaration (conjoints et

enfants de moins de 30 ans), les données son créées par imputation, c'est-à-dire en utilisant les données fiscales dont on dispose au sujet des autres membres de la famille (exemptions personnelles demandées, allocations familiales reçues, montants du crédit d'impôt pour enfants et autres crédits demandés).

Le tableau 1 compare les données sur les familles des personnes âgées, reconstituées à partir des dossiers fiscaux, aux données correspondantes produites à partir des résultats du recensement de 1981.

Tableau 1

Population âgée de 65 ans et plus, selon le genre de famille, dossiers fiscaux de 1982 et recensement du Canada de 1981

Genre de famille	Dossiers fiscaux	Recensement	Rapport
	(en milliers)	(en milliers)	
Personnes hors famille	563	860	0.65
Membres d'une famille	1069	1501	0.71
familes époux-épouse	1051	1191	0.88
familles monoparentales	18	310	0.06
Population totale	1631	2361	0.69

Comme nous l'avons souligné, Statistique Canada reconstitue les familles en procédant par imputation, c'est-à-dire en utilisant les données contenues dans les déclarations produites par des personnes appartenant à une famille pour déterminer la présence de membres non déclarants dans cette même famille. Le régime fiscal prévoit des exemptions pour différentes catégories de personnes à charge: le conjoint, les enfants de moins de 18 ans, les enfants de plus de 18 ans ainsi que pour d'autres personnes répondant à certaines exigences. À l'heure actuelle, Statistique Canada peut uniquement reconstituer les familles dans lesquelles les personnes à charge sont le conjoint et les enfants. Il ne peut pas déceler la présence dans une famille de membres appartenant à une autre catégorie de personnes à charge. On sait que 61% des personnes âgées ont produit une déclaration. À partir de ces déclarations, Statistique Canada a pu découvrir la présence d'un certain nombre de conjoints âgés de 65 ans et plus (n'ayant pas produit de déclaration), ce qui porte à 69% (déclarants et personnes à charge) la proportion de la population âgée de 65 ans et plus dont permettent de rendre compte les dossiers fiscaux. Le type d'exemption pour personne à charge demandé par la plupart des déclarants âgés est "l'exemption personnelle supplémentaire" (l'annexe 6 de la déclaration, dans laquelle cette exemption doit être demandée, est reproduite à l'appendice 1). Toutefois, en ce moment, Statistique Canada ne peut pas déceler la présence de ces personnes à charge dans les familles; divers facteurs l'en empêchent. Contrairement aux autres exemptions pour lesquelles nous disposons de données permettant de faire des suppositions quant à l'âge et au sexe des personnes à charge, seul le montant réclamé, dans le cas de l'exemption personnelle supplémentaire, est mis en mémoire par Revenu Canada. En fait, nous ne sommes même pas en mesure de déterminer le nombre de personnes à charge visées par une demande, Revenu Canada n'effectue pas la saisie de la date de naissance des personnes à charge pour lesquelles l'exemption est demandée ni leur lien avec le déclarant. Nous pourrions faire des hypothèses très générales, mais les personnes à charge pour lesquelles l'exemption peut être demandée peuvent appartenir à un autre groupe d'âge que celui du déclarant et par conséquent, toute estimation relative à l'âge

serait hasardeuse. Pour l'année 1986, 3.6 millions de déclarants ont réclamé l'exemption personnelle supplémentaire. Il est impossible de déterminer le nombre de personnes à charge visées ni la proportion de ces dernières âgées de 65 ans et plus.

À l'heure actuelle, nous pouvons généralement reconstituer les familles de recensement dans lesquelles un père (ou une mère), qui a des enfants à charge, remplit une déclaration, ainsi que les familles dans lesquelles un père (ou une mère) et ses enfants remplissent une déclaration. Par contre, lorsqu'une famille est composée d'une personne non mariée âgée de plus de 30 ans habitant avec son père (ou sa mère) et que les deux personnes remplissent une déclaration, celles-ci seront comptées comme deux personnes hors famille et non comme une famille monoparentale. Par ailleurs, si dans l'exemple ci-dessus, le père (ou la mère) ne remplit pas de déclaration, les dossiers fiscaux ne nous permettront même pas de savoir que cette personne existe. C'est pour cette raison que lorsqu'on utilise les dossiers fiscaux, il y a, chez les personnes âgées, très forte sous-représentation des familles monoparentales.

En résumé, les dossiers fiscaux permettent de rendre compte de 69% des personnes âgées et d'obtenir des données sur le revenu de 61% de celles qui remplissent une déclaration. Les pourcentages devraient être plus élevés pour l'année d'imposition 1986, lorsqu'entrera en vigueur le crédit fédéral de la taxe sur les ventes.

Les familles époux-épouse dans lesquelles au moins un conjoint est âgé de 65 ans et plus peuvent être reconstituées si l'un des conjoints produit une déclaration. Toutefois, il est impossible d'isoler les personnes à charge âgées qui ne produisent pas de déclaration, d'où la sous-représentation dans ce groupe d'âge des personnes hors famille et des familles monoparentales.

3. PROGRAMME DE LA SÉCURITÉ DE LA VIEILLESSE

3.1 Introduction

Le programme de la Sécurité de la vieillesse (SV) est un programme quasi-universel offert aux personnes âgées de 65 ans et plus. Des prestations mensuelles sont versées à toutes les personnes qui satisfont aux exigences de résidence (en général, 10 années consécutives de résidence au Canada) et qui en font la demande. Des prestations supplémentaires, le Supplément de revenu garanti (SRG), sont versées aux personnes âgées à faible revenu (le montant du SRG est calculé en fonction du revenu de l'année précédente).

3.2 Représentativité des données

Quatre-vingt-seize pour cent de la population âgée de 65 ans et plus reçoit des prestations de la SV (seules les personnes qui ne satisfont pas aux critères de résidence ou qui n'ont pas fait de demande de prestations ne figurent pas dans les dossiers). Par conséquent, ces données administratives permettent d'établir une estimation très fiable de la population âgée.

3.3 Revenu

Lorsqu'une personne demande le SRG, elle doit fournir des renseignements sur ses sources de revenu au cours de l'année précédente. Les renseignements à fournir correspondent à ceux demandés dans la déclaration d'impôt des particuliers, sauf que le demandeur n'a pas à déclarer ses prestations de la SV ni les allocations familiales et qu'il doit indiquer, le cas échéant, le montant de ses indemnités pour accident du travail. Ainsi, pour déterminer le revenu total d'une personne, il faut faire des ajustements, ce qui, dans la pratique, pose

des problèmes étant donné qu'un enregistrement contient le montant des prestations à recevoir pour l'année en cours et le revenu de l'année précédente. Pour y remédier, il faut soit obtenir le montant des prestations et celui du revenu pour les deux années en question, soit faire une supposition en ce qui concerne les prestations reçues au cours de la première année (c'est-à-dire, supposer que le montant des prestations de la SV que la personne a reçu au cours de l'année précédente est le même que le montant déclaré pour l'année en cours).

3.3 Composition de la famille

Les dossiers de la SV contiennent très peu de données sur les caractéristiques démographiques des prestataires, si ce n'est l'âge et le sexe. Toutefois, les dossiers peuvent nous renseigner sur l'état matrimonial des couples dans lesquels l'époux et l'épouse ont 65 ans ou plus, car ces derniers reçoivent tous deux des prestations et ont un compte conjoint au ministère de la Santé. Pour ce qui est des couples mariés dans lesquels seulement un des conjoints est âgé de 65 ans ou plus, les dossiers ne contiennent aucune indication de l'état matrimonial du prestataire et par conséquent, il est impossible de déterminer s'il a un conjoint ou non. Lorsqu'une personne présente une demande pour recevoir le Supplément de revenu garanti, elle doit fournir, si elle est mariée, le revenu de son conjoint. Les dossiers nous permettent donc d'isoler ces couples. Cependant, ils ne nous permettent pas de savoir si les personnes âgées ont des personnes à charge ou si elles sont elles-mêmes à la charge de quelqu'un. Comme nous l'avons dit, les dossiers de la SV permettent d'isoler les familles époux-épouse où les deux conjoints ont 65 ans ou plus mais ils ne permettent pas la reconstitution complète des familles parce qu'ils ne nous renseignent pas sur la présence d'enfants.

4. COUPLAGE DES DONNÉES CONTENUES DANS LES DOSSIERS FISCAUX ET LES DOSSIERS DE LA SÉCURITÉ DE LA VIEILLESSE

4.1 Introduction

Les dossiers fiscaux ne contiennent pas de données sur les personnes âgées à faible revenu (environ 40% de la population âgée). Par contre, ces personnes à faible revenu sont justement celles qui ont droit de demander et qui demandent effectivement les prestations supplémentaires de la SV: le SRG. À cette fin, elles doivent fournir des données au ministère de la Santé et du Bien-être social. Si on pouvait combiner ces deux sources de données, dossiers fiscaux et dossiers de la SV, on pourrait probablement produire des données assez complètes sur le revenu de cette population. Un essai de couplage des données contenues dans ces dossiers a été effectué pour l'année d'imposition 1984 en Colombie-Britannique.

4.2 Représentativité des données

Comme on pouvait s'y attendre, l'essai effectué en Colombie-Britannique en 1984 a révélé que 96% des personnes âgées recevaient des prestations de la SV et que 64% d'entre elles avaient rempli une déclaration d'impôt. La proportion des personnes pour lesquelles le couplage a pu être effectué est de 61%, ce qui signifie que les 3% restants sont des personnes qui ont rempli une déclaration mais n'ont pas reçu de prestations de la SV. Ainsi, les dossiers de la SV et les dossiers fiscaux rendent compte de 99% de la population âgée de 65 ans et plus. Soulignons cependant qu'il s'agit ici de résultats obtenus pour une année et une province seulement. Il serait donc hasardeux d'étendre ces résultats aux autres provinces. En effet, il se peut que les résultats soient différents dans les provinces où le profil d'âge de la population et les tendances de l'immigration diffèrent.

4.3 Revenu

Nous avons vu que les dossiers fiscaux contenaient des données sur le revenu de 64% des personnes âgées et nous avons établi que le pourcentage de celles qui reçoivent le (SRG) et de celles qui ne le reçoivent pas correspond à 15 et 49% respectivement de la population âgée totale. Les dossiers de la SV contiennent des données sur presque toutes les personnes qui reçoivent le SRG et ces dernières constituent 41% de la population âgée totale. Si l'on combine les données des 49% des personnes âgées qui produisent une déclaration d'impôt mais qui ne reçoivent pas le SRG et celles provenant des dossiers de la SV sur les personnes qui reçoivent le SRG, on peut obtenir des données sur le revenu de 90% des personnes âgées.

Tableau 2

Personnes âgées et représentativité des données sur le revenu, Colombie-Britannique, 1984

Prestations	Nombre de personnes âgées	% des personnes âgées	Source des données: Dossiers de la SV	% des personnes âgées	Dossiers fiscaux	% des personnes âgées	Dossiers fiscaux -dossiers de la SV	% des personnes âgées
Pas de SRG	207,141	58	227	0	174,027	49	174,254	49
SV et SRG	148,926	42	144,132	40	52,780	15	146,154	41
Total	356,067	100	144,359	41	226,807	64	320,408	90

(Tous les pourcentages renvoient au total de 356,067).

4.4 Composition de la famille

L'essai de couplage n'a pas permis d'obtenir des renseignements sur la composition ou le type de la famille parce qu'il n'a porté que sur les données sur le revenu contenues dans les dossiers et non sur les données relatives à la famille que nous avons décrites plus haut.

5. RECHERCHES À VENIR

Dans cette étude, nous avons soulevé diverses questions qu'il faudrait examiner de plus près afin de tirer le maximum des données relatives aux personnes âgées contenues dans les dossiers administratifs.

Il faudrait définir les caractéristiques du sous-ensemble de la population âgée ne recevant pas de prestations de la SV afin d'améliorer la technique d'estimation qui sera retenue.

Beaucoup de travail reste à faire sur le rapprochement des données sur le revenu que l'on trouve dans les dossiers fiscaux et dans les dossiers de la SV. Bien que les catégories de revenu y soient presque identiques, certains ajustements doivent être faits. Il faut également trouver une solution au problème posé par le fait que, dans les dossiers de la SV, les prestations reçues au cours d'une année sont additionnées au revenu de l'année précédente. Enfin, il faudrait également se pencher sur la question des sources de revenu qui ne sont pas assujetties à l'impôt.

Pour améliorer la représentativité des données sur la famille, il faudrait définir le groupe de personnes âgées qui sont des personnes à charge et ne produisent pas de déclaration d'impôt. L'introduction, en 1986, du crédit fédéral de la taxe sur les ventes va amener un certain nombre de ces personnes à produire une déclaration. Les données pour l'année d'imposition 1986 seront disponibles en octobre et on est en train d'effectuer une étude pour évaluer l'effet de ce crédit sur les données fiscales. Une autre façon d'améliorer la représentativité des données sur les personnes âgées (il s'agit d'une solution à plus long terme) serait d'effectuer la saisie des données recueillies dans l'annexe 6 des déclarations d'impôt des particuliers.

6. CONCLUSION

Il est clair qu'on peut isoler la population âgée en utilisant les dossiers de la SV, à la condition toutefois de faire un très léger ajustement (de l'ordre de 4%). Les caractéristiques que l'on peut définir se limitent à l'âge et au sexe. Pour ce qui est de l'état matrimonial, on peut uniquement déceler les personnes mariées. Selon le genre d'ajustement effectué pour rendre compte de l'ensemble de la population âgée, il serait également possible d'obtenir des données géographiques (à partir du code postal).

En couplant les données contenues dans les dossiers de la SV et celles contenues dans les dossiers fiscaux, on peut obtenir des données sur le revenu de 90% de la population âgée. Si, dans la base des données fiscales, on pouvait trouver et isoler les personnes qui reçoivent le SRG, les deux ensembles de dossiers seraient complémentaires. En effet, c'est en prenant dans les dossiers de la SV les données sur le revenu des personnes qui reçoivent le SRG et, dans les dossiers fiscaux, les données sur les personnes qui ne le reçoivent pas que l'on obtiendrait les données les plus complètes. Les données sur la famille posent davantage de difficultés. Étant donné que les dossiers sur la SV ne contiennent pas de données sur les personnes à charge et que la base des données fiscales ne contient pas de renseignements sur les personnes âgées qui sont elles-mêmes des personnes à charge, il est difficile de reconstituer les familles si ce n'est les familles époux-épouse.

Dans l'ensemble, les premiers résultats du couplage des données administratives (famille et revenu) sont prometteurs. De plus, avec l'introduction du système de crédits d'impôt, on devrait pouvoir produire des données plus complètes sur les familles des personnes âgées.

BIBLIOGRAPHIE

Auger, E. (1987). Family Data from the Canadian Personal Income Tax File, Statistique Canada, (étude non publiée).

Podoluk, J. (1987). Utilisation des données administratives comme complément des données du recensement, Statistique Canada, (étude non publiée).

Selley, O. (1987). Pilot Study - Microrecord Linkage of the Tax and Old Age Security Records for British-Columbia, Statistique Canada, (étude non publiée).

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

UNE ENQUÊTE À DEUX VOLETS : L'ÉCHANTILLON PERMANENT D'ASSURÉS SOCIAUX EN FRANCE

ANDRÉE MIZRAHI et ARIÉ MIZRAHI[1]

RÉSUMÉ

En France, 99 % de la population est protégée par l'un des régimes d'Assurance Maladie obligatoire. En règle générale, en médecine de ville, le patient paye ses soins et se fait rembourser ensuite les sommes à la charge de l'Assurance Maladie. En cas d'hospitalisation, l'Assurance paye directement l'établissement pour la part qui lui incombe. Un échantillon permanent d'assurés sociaux (au 1/1200) est suivi depuis 1977 et couvrira 40 000 personnes en 1989 ; les informations administratives, sur les natures précises et les prix des consommations médicales de chaque personne relevées en routine, seront complétées par une enquête auprès des assurés eux-mêmes (1/4 de l'échantillon chaque année). Elle porte sur les caractéristiques socio-économiques, la protection contre la maladie, les affections et invalidités et la perception du système de santé ; un carnet de compte de 3 semaines sur les consommations médicales, remboursées ou non, y est joint.

1. L'ASSURANCE MALADIE EN FRANCE

1.1. Les différents régimes

En France les régimes de Sécurité Sociale protègent, à un titre ou à un autre et d'une manière plus ou moins importante, environ 99 % de la population pour les soins de santé.

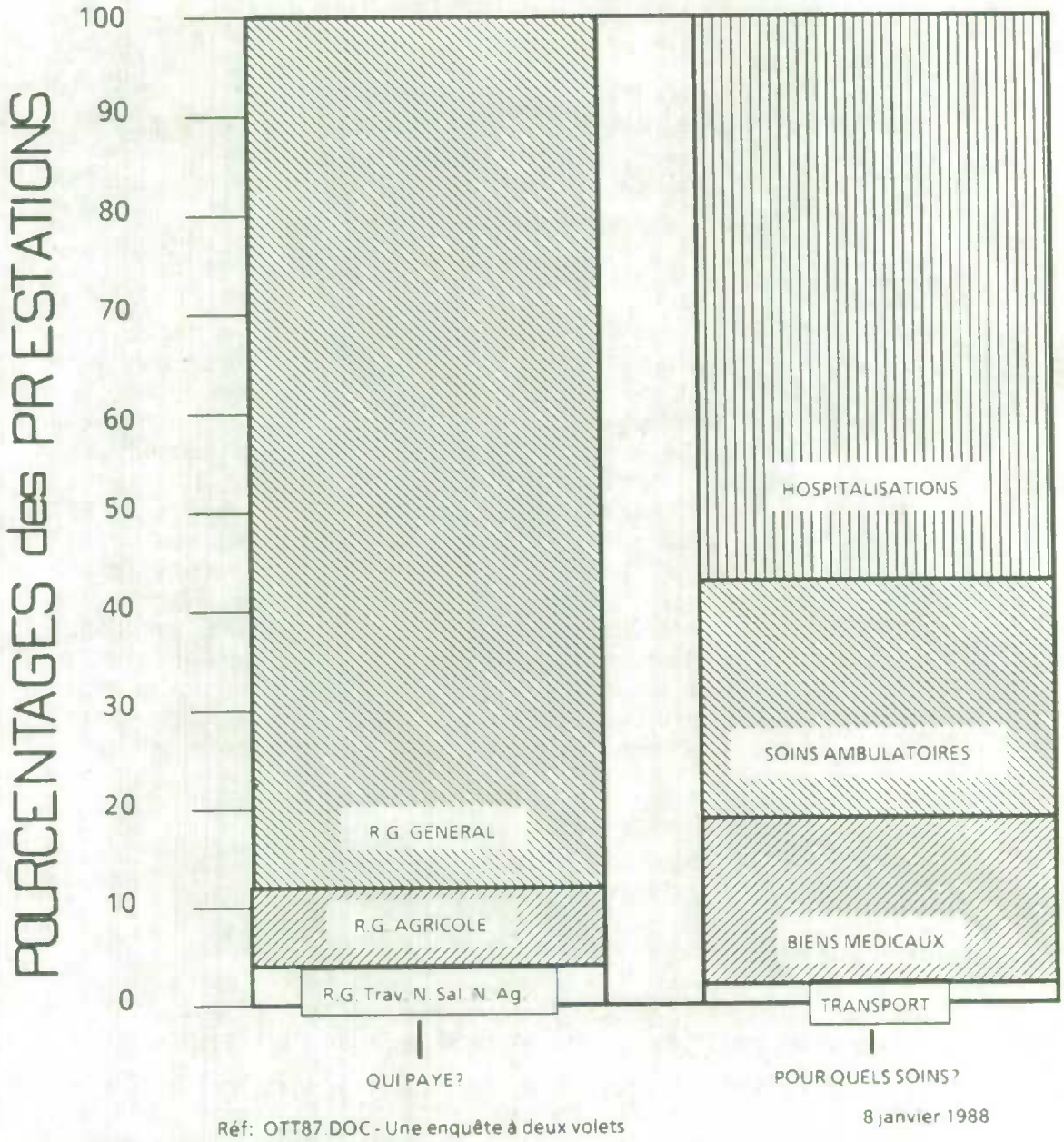
Cette couverture obligatoire, assise sur l'activité professionnelle est effectuée essentiellement par les Assurances Maladie, Maternité et Accidents du travail de trois régimes: le Régime Général des Salariés (87 % des prestations versées), le Régime Agricole (9 % des prestations) et le Régime des Travailleurs Non Salariés des Professions Non Agricoles (4 % des prestations).

L'assurance maladie est financée à 94 % par des cotisations proportionnelles aux salaires (ou aux revenus professionnels pour les non salariés). En contre-partie elle prend en charge 76 % des soins de santé: 86 % des frais d'hospitalisation, 66 % des soins ambulatoires et 61 % des biens médicaux (cf. graphique 1).

[1] Andrée Mizrahi et Arié Mizrahi, Centre de Recherche, d'étude et de documentation en économie de la santé, 1 rue Paul-Cézanne, Paris, France, 75008.

Graphique 1
PRESTATIONS de SECURITE SOCIALE

FRANCE 1986



Réf: OTT87.DOC - Une enquête à deux volets

8 janvier 1988

1.2. Les modalités de prise en charge

L'intervention de l'Assurance Maladie se fait sur la base de tarifs fixés par les pouvoirs publics (établissements hospitaliers, produits pharmaceutiques, forfaits soins des personnes âgées, etc) ou négociés avec les représentants des professions de santé sous forme de convention quadriennale (médecins, dentistes, kinesithérapeutes, ...).

Dans le cas général, pour les soins ambulatoire ou dispensés à domicile, l'assuré paye directement le médecin, le dentiste, l'auxiliaire, le laboratoire, le pharmacien, etc...et se fait rembourser ultérieurement par la Caisse Primaire d'Assurance Maladie dont il dépend ; pour le Régime Général 129 Caisses Primaires sont réparties sur le territoire. Les remboursements sont de 75 % des tarifs pour les dépenses de médecins et de dentistes, 65 % pour les dépenses d'auxiliaires et de laboratoires et 40, 70 ou 90 % pour les produits pharmaceutiques.

Le travailleur ou le retraité est assuré : c'est lui qui perçoit les éventuels remboursements pour ses dépenses de soins et pour celles des personnes à sa charge, dites ayants-droit de l'assuré.

Pour les hospitalisations d'une manière générale, la Sécurité sociale paye directement les établissements hospitaliers, les assurés règlant uniquement les montants à leur charge, 20 % de la dépense ou un forfait de 25 francs par jour d'hospitalisation.

Dans certains cas cependant, l'Assurance Maladie prend à sa charge 100 % des tarifs soit pour certains soins : ceux liés à 30 affections particulièrement longues et coûteuses (tumeur maligne, maladie de Parkinson, hémophilie etc..), à la maternité, aux interventions chirurgicales à partir d'une certaine importance, aux journées d'hospitalisation au-delà du 30ème jour, aux accidents du travail etc..., soit pour certaines personnes : bénéficiaires du Fonds National de Solidarité, anciens combattants, accidentés du travail, certains handicapés, etc... 10% des personnes protégées (assurés + ayants-droit) bénéficieraient d'une au moins de ces mesures.

La réalité est plus complexe car les tarifs ne sont pas toujours les prix effectivement pratiqués, ainsi les dépenses des familles sont supérieures aux sommes règlementairement laissées à leur charge, dans la mesure ou près de 30 % des médecins dépassent officiellement les tarifs, la plupart des dentistes, lorsqu'ils effectuent des prothèses, certains auxiliaires, les opticiens etc...

In fine, en 1986 les dépenses médicales en France ont été financées à 76,7 % par la Sécurité Sociale et 14,6 % par la population ; 7,2 % par les couvertures complémentaires (4,3 par les mutuelles et 2,9 par les assurances privées) qui remboursent une partie des dépenses non prises en charge par la Sécurité Sociale, et 1,5 % par l'Aide Médicale (programme d'assistance aux personnes ayant des ressources insuffisantes).

1.3. Les fichiers et les statistiques de sécurité sociale

L'Assurance Maladie crée en permanence, en sous-produit de sa gestion, de très nombreuses statistiques recensant l'ensemble de ses opérations et largement utilisées par les administrateurs du secteur médical, les partenaires sociaux, les économistes etc... ; ces chiffres constituent en particulier la base de l'élaboration des Comptes Nationaux de la Santé et des Comptes Nationaux de la Protection Sociale largement utilisées pour la définition des politiques de santé aussi bien au niveau national qu'international.

Mais les fichiers détenus par l'Assurance Maladie ont pour fonction première de permettre le versement à bon escient des prestations, aux assurés sociaux lors d'un paiement direct des soins, et aux producteurs (ou aux distributeurs) dans le cas de tiers payant.

Les informations utilisables se rapportent donc aux assurés (fichier immatriculation : nom, prénom, adresse, âge, sexe, nature des droits ouverts...), aux producteurs de soins (fichiers de producteurs : identification, spécialité, droit aux dépassements...), aux sommes à payer (fichiers prestations : identité de l'assuré et du bénéficiaire, lettre-clé et cotation des soins, selon la nomenclature des actes professionnels, dates...).

Ces fichiers sont détenus par les Centres Informatiques de l'Assurance Maladie et dispersés sur tout le territoire français.

Les informations nécessaires aux paiements des assurés ou des producteurs, tels le nom et l'adresse de l'assuré et les sommes versées, sont relevées avec beaucoup de précision ; les informations non directement nécessaires au paiement le sont beaucoup moins. Ainsi l'hospitalisation qui représente plus de la moitié des dépenses médicales n'est pas toujours intégralement rapportée dans les dossiers des personnes concernées, les établissements hospitaliers étant financés par un budget global, ces informations ne sont pas indispensables au règlement des prestations ; il en est de même pour les forfaits soins

versés aux institutions qui accueillent des personnes âgées dépendantes qui sont pris en charge par l'Assurance Maladie.

La mise à jour des fichiers est parfois assez longue, en particulier pour les informations n'ayant pas d'incidence sur les versements des prestations, assuré décédé ou se faisant rembourser à une adresse antérieure ou figurant dans plusieurs fichiers (le système rend de toute manière impossible de se faire rembourser plusieurs fois le même acte), d'où une certaine incertitude sur les effectifs de personnes protégées par les différents régimes.

1.4. L'échantillon permanent d'assurés sociaux

De plus, compte tenu de la complexité et de l'éparpillement des informations, il est difficile d'établir des statistiques au niveau de la personne ; pour pallier ces lacunes, le Département Statistique de la Caisse Nationale d'Assurance Maladie des Travailleurs Salariés a mis en place, en collaboration avec le Centre de Recherche, d'Études et de Documentation en Économie de la Santé (CREDES, anciennement CREDOC) le suivi permanent d'un échantillon représentatif d'assurés sociaux, désigné au 1/1 200ième de manière aléatoire.

Ce panel d'assurés sociaux constitue l'un des instruments majeurs dont dispose la Sécurité sociale pour analyser les dépenses de l'Assurance Maladie sous l'angle des consommateurs de soins (analyse micro-économique) et mesurer l'incidence des modifications législatives et règlementaires en matière de protection maladie. Il a été mis en place en 1977 (7 Caisses Primaires d'Assurance Maladie volontaires, sur un total de 129 Caisses) et a progressivement été étendu, lorsque son efficacité a été démontrée ; il portera au début de 1988 sur environ 32 000 personnes, soit 82 % du champ retenu ; l'extension complète est prévue pour 1989 et les informations concerneront alors environ 39 000 personnes protègées par le Régime Général de Sécurité sociale, soit un sondage au 1/1 200 environ.

La méthode de désignation retenue (toutes les personnes dont le Numéro National d'Identité vérifie une certaine propriété) assure le renouvellement automatique de l'échantillon.

Toutes les informations concernant les assurés ainsi désignés et leurs ayants-droit (identification et prestations) sont relevées en routine (de manière informatique au moment où elles sont créées) et centralisées pour donner lieu aux exploitations statistiques et aux analyses économiques.

Les caractéristiques des fichiers d'Assurance Maladie se retrouvent dans ceux du panel d'assurés sociaux, et des informations d'un grand intérêt pour l'analyse socio-économique du comportement des consommateurs (profession, couverture complémentaire, niveau d'instruction, etc...) ne sont pas dans ces fichiers. Il a donc été décidé d'adjoindre à l'information déjà disponible des renseignements à obtenir auprès des assurés : ce volet auprès des ménages est dit **Enquête sur la santé et la protection sociale.** L'ensemble des informations issues des deux volets figure dans le tableau 1.

Tableau 1

Échantillon permanent d'assurés sociaux
Source et nature des informations

Source:	
SONDAGE DANS LES FICHIERS DE SÉCURITE SOCIALE	ENQUÊTE AUPRÈS DES MÉNAGES
Champ:	*Champ:*
assuré + ayants-droit	assuré + ses ayants-droit cohabitant + autres personnes vivant dans le logement + ayants-droit non cohabitant
Nature des informations:	*Nature des informations:*
Démographie	
âge et sexe	âge et sexe (y compris ayants-droit non cohabitant)
Protection sociale	
régime détaillé de Sécurité sociale de l'assuré	régime d'Assurance Maladie et protection complémentaire de chaque membre du ménage
Socio-économie	
	activité et profession, niveau d'instruction pour tous les membres du ménage et les ayants-droit non cohabitant
Épidémiologie	
	morbidité ressentie, invalidité
Soins médicaux	
nature du producteur, lieu, nature des soins (nomenclature des actes) et importance (lettres-clé), date, tarifs, sommes déboursées, prestations versées *(information relevée en routine).*	nature du producteur, lieu, date, nature des soins, dépenses pour les consommations médicales *(information relevée pendant 3 semaines seulement).*
	Opinions
	avenir du système de Sécurite Sociale, mesures proposées

2. LE VOLET MÉNAGE DU PANEL D'ASSURÉS SOCIAUX

Le volet ménage portera chaque année sur un quart de l'échantillon permanent d'assurés sociaux, désigné de manière aléatoire, de manière à couvrir l'ensemble de l'échantillon en quatre ans.

2.1 L'enquête pilote

La méthode d'enquête a été mise au point à partir des résultats d'une enquête pilote qui s'est déroulée en Mai-Juin 1987 au cours de laquelle cinq modalités ont été expérimentées ; elles diffèrent par le mode de contact (téléphone ou visite d'enquêteur), la durée (forme normale et forme allégée) et le nombre de contacts (2 ou 4 appels téléphoniques), la nature du réseau d'enquêteurs (institut de sondage professionnel, cadres de gestion de la Sécurite sociale, chercheurs CREDES).

Trois modalités diffèrent uniquement par la nature des enquêteurs:

a) - enquêteurs professionnels,
b) - chercheurs du CREDES,
c) - agents d'encadrement de la Sécurite sociale ;

} dans ces 3 cas l'enquête est dite normale, elle comporte au minimum 4 appels téléphoniques,

d - l'enquête, par téléphone, par enquêteur professionnel, dite allégée, comporte au minimum 2 appels téléphoniques.

e - l'enquête, réalisée au cours de 2 déplacements d'enquêteurs professionnels, dite face à face.

Pour une même méthode d'enquête (normale), les taux de refus observés par les non-enquêteurs (chercheurs du CREDES et personnel d'encadrement de la Sécurite sociale) sont nettement supérieurs à ceux obtenus par le personnel spécialisé.

Les taux de retour des carnets de soins et des "questionnaires santé" sont légèrement inférieurs dans les modalités téléphoniques à ceux obtenus par déplacement d'enquêteur ; en revanche, les taux d'acceptation du premier appel téléphonique sont nettement plus élevés que ceux de la première visite.

2.2. La modalité retenue

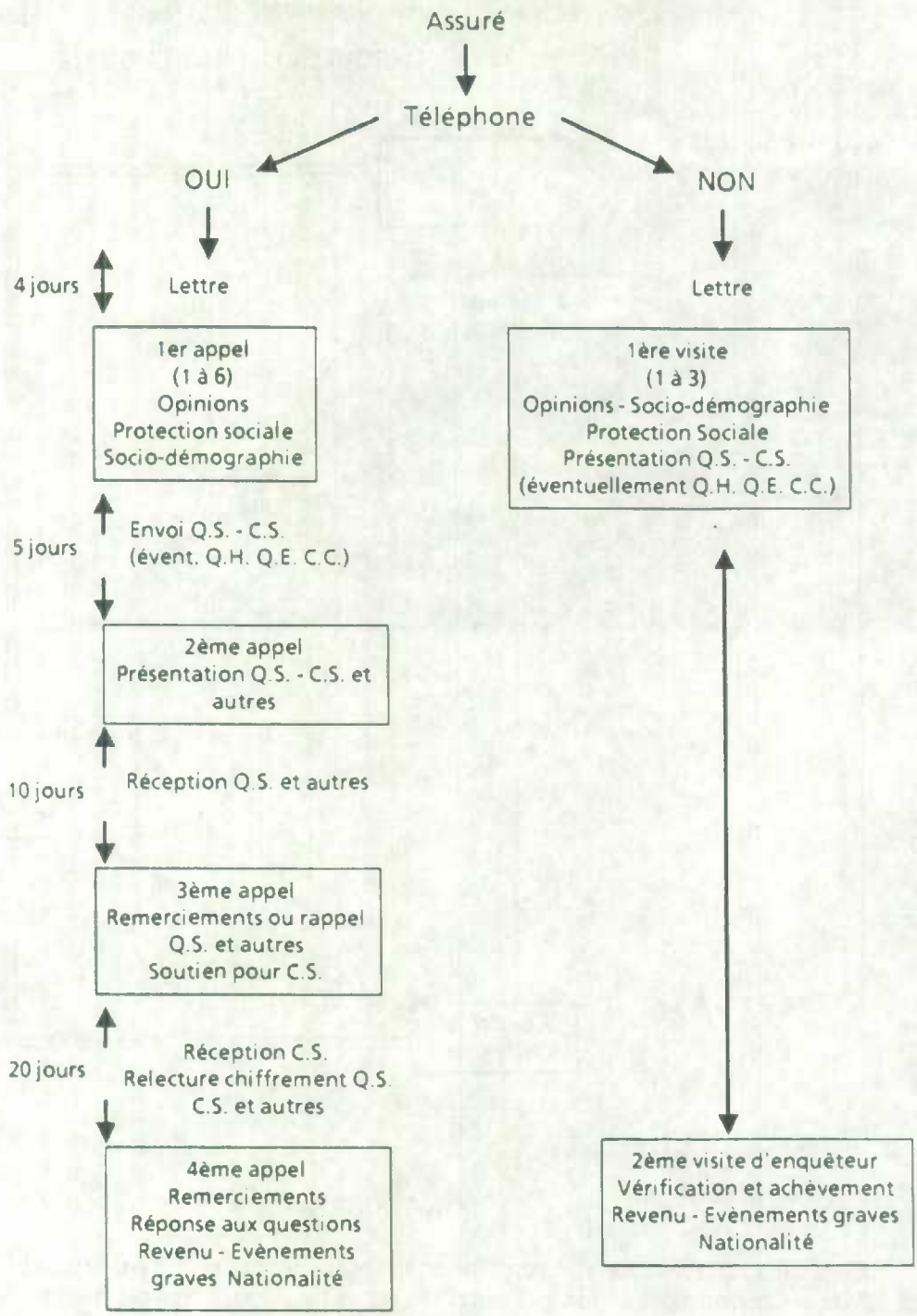
En conséquence, les opérations de terrain seront confiées à des instituts spécialisés dans les enquêtes auprès des ménages.

Les contacts seront téléphoniques et par correspondance pour tous les ménages ayant le téléphone, par déplacement d'enquêteur pour les autres. Le questionnaire téléphonique sera scindé en deux parties (et non trois) pour être administré au cours du premier et du dernier appel, cependant deux appels de soutien sont prévus en cours de la période de remplissage du carnet de soins.

Pour les ménages n'ayant pas le téléphone, deux visites sont prévues, la première pour administrer le questionnaire principal et les questionnaires annexes et déposer le carnet de soins, la deuxième pour relever les carnets de soins et vérifier la cohérence de l'ensemble. L'enchaînement de ces opérations figure sur le graphique 2.

La relecture et la codification de l'ensemble des documents sont effectuées au fur et à mesure de leur réalisation ou réception pour détecter le maximum d'anomalies ou d'imprécisions et des questions sur ces points sont posées aux enquêtés lors du contact téléphonique suivant ; cette procédure améliore sensiblement la qualité de l'information obtenue.

Graphique 2
ECHANTILLON PERMANENT D'ASSURES SOCIAUX
DEROULEMENT DU VOLET MENAGE



Réf: OTT87.DOC - Une enquête à deux volets

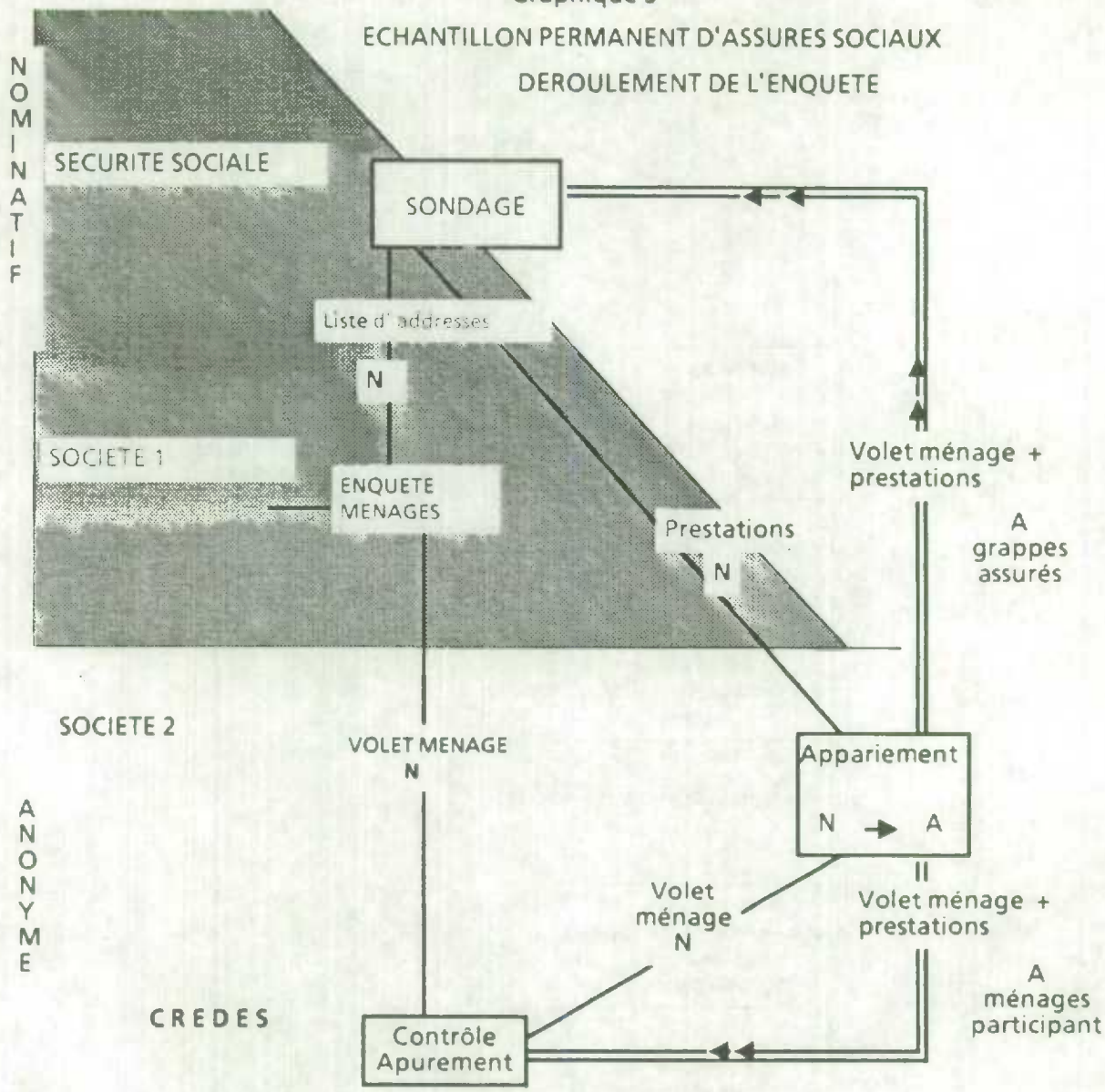
8 Janvier 1988

Pour satisfaire les restrictions protègeant la vie privée et la liberté publique toutes ces données sont traitées de manière completement anonyme : cette condition ne pose pas de problème en ce qui concerne chacune des 2 parties, sondage dans les fichiers de Sécurité sociale et enquête auprès des particuliers ; l'appariement de ces 2 sources est effectué en double aveugle par un troisième organisme auquel les fichiers sont fournis après avoir été rendus entièrement anonymes (cf. graphique 3).

Graphique 3

ECHANTILLON PERMANENT D'ASSURES SOCIAUX

DEROULEMENT DE L'ENQUETE



Réf. OTT87.DOC - Une enquête à deux volets

8 janvier 1988

3. CONCLUSION

De nombreuses difficultés devront encore être résolues : traitement statistique de données micro-économiques longitudinales (prestations), consolidation d'observations relatives à des périodes successives (ventilation de l'échantillon en 4 sous-échantillons observés chacun une fois au cours de la période), harmonisation de données d'origine différentes et suppression des incohérences éventuelles.

Cependant, par la richesse des informations obtenues, ce travail ouvre la voie à des recherches de grande importance, il permettra, nous l'espérons, de suivre le déroulement dans le temps d'événements liés à la santé, et tout particulièrement l'évolution de l'influence des critères socio-économiques et de la morbidité ressentie sur le niveau et la stucture des consommations médicales et de disposer ainsi de séries micro-économiques dans le secteur médical.

BIBLIOGRAPHIE

C.N.A.M.T.S. (1986). Statistiques des Régimes d'Assurance Maladie en 1985. Département Statistique

C.N.A.M.T.S. (1982). Qui consomme quoi ? Département statistique, Paris. Mise à jour des résultats issus de l'échantillon permanent d'assurés sociaux, Décembre 1984.

S.E.S.I. (1987). Comptes Nationaux de la Santé 1984-1985-1986. Ministère des Affaires Sociales.

Mizrahi Andrée et Mizrahi Arié (1978). Méthode de sondage: Enquête permanente dans les dossiers de sécurité sociale. CREDOC - Paris.

Volatier, J.L.. Enquête sur la santé et la protection sociale. Methodologie de l'enquête pilote. CREDES (à paraître).

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

SYMPOSIUM SUR LES UTILISATIONS STATISTIQUES DES DONNÉES ADMINISTRATIVES
UTILISATION DES DOSSIERS DE L'IMPÔT SUR LE REVENU DES SOCIÉTÉS À DES FINS D'ANALYSE DE LA POLITIQUE FISCALE

F. HOSTETTER, C.D. McCANN et B. ZIRGER[1]

RÉSUMÉ

Depuis plusieurs années, les analystes du ministère des Finances effectuent un échantillonnage des données fiscales et financières contenues dans les déclarations de l'impôt sur le revenu des sociétés en vue d'étudier l'incidence de variations hypothétiques de la politique fiscale sur les recettes fiscales et sur le statut fiscal. Bien que les dossiers de l'impôt sur le revenu soient une source appropriée pour de telles données, des problèmes de qualité peuvent se poser du fait qu'il s'agit d'utiliser à des fins statistiques des dossiers établis à des fins administratives. Les difficultés peuvent être dues aux processus de déclaration et d'évaluation employés, à l'absence de normalisation des déclarations fiscales et financières produites par les sociétés, aux définitions adoptées, à l'interprétation des données et au processus d'extraction de celles-ci. Des procédures de contrôle de la qualité des données sont appliquées afin de permettre, dans une certaine mesure, d'identifier et de corriger les erreurs dues à ces aspects particuliers.

1. INTRODUCTION

Le présent exposé décrit les opérations de collecte des données statistiques effectuées par Revenu Canada Impôt (RCI) à partir des dossiers de l'impôt sur le revenu des sociétés. Il présente également les procédures de contrôle de la qualité des données actuellement utilisées et celles qui pourraient l'être. La première partie donne un aperçu des priorités du client et de la façon dont il en est tenu compte au moment de la sélection de l'échantillon. Elle est suivie d'une présentation des données fournies par les sociétés visées par la collecte, de la pertinence de ces données du point de vue de l'utilisation statistique des dossiers de l'impôt sur le revenu des sociétés, des opérations de collecte des données et des procédures connexes de contrôle de la qualité. Cet exposé donne, en conclusion, quelques orientations possibles des opérations de contrôle de la qualité des données tirées d'un échantillon de dossiers de l'impôt sur le revenu des sociétés. Voici tout d'abord un aperçu global des utilisations finales des données recueillies.

[1] F. Hostetter, C. McCann et B. Zirger, Revenu Canada, Impôt, Salle 303 Édifice MacDonald, 123 rue Slater, Ottawa, Ontario, Canada K1A 0T6.

2. RENSEIGNEMENTS GÉNÉRAUX

Le contrôle de la qualité des données statistiques sur les sociétés s'inscrit dans un processus permanent de plus vaste portée dont le but final est la révision par le Parlement de la Loi de l'impôt sur le revenu. Les modifications qui doivent être apportées à la Loi portent surtout sur l'imposition des sociétés et résultent de l'évaluation permanente des mesures fiscales en vigueur et de celles qui ont été proposées. Cette évaluation est faite par le ministère des Finances et se fonde, entre autres, sur l'estimation de l'incidence sur les recettes de telles mesures fiscales. La principale méthode utilisée pour estimer les recettes consiste en une micro-simulation au moyen du modèle de RCI relatif à l'impôt sur le revenu des sociétés.

Les micro-unités utilisées pour la simulation sont des dossiers de l'impôt sur le revenu des sociétés sélectionnés au moyen d'une base de sondage définie en fonction du fichier de données administratives sur les sociétés de RCI. Aux données tirées de ce fichier sont ajoutées d'autres données extraites des états financiers et des déclarations d'impôt des sociétés faisant partie de l'échantillon. Sous l'année d'imposition visée, les données sont ensuite regroupées de façon à avoir une représentation transversale des sociétés imposables qui est fournie sous la forme d'un fichier ordinolinque, le fichier-échantillon des sociétés. Ce fichier-échantillon constitue la principale source de données pour le modèle relatif à l'impôt sur le revenu des sociétés.

Revenu Canada Impôt (RCI) a mis sur pied ces opérations de collecte de données statistiques en 1979. Ces dernières ont été conçues à partir de l'énoncé des besoins prioritaires du ministère des Finances, notre client, concernant un modèle fiscal applicable aux sociétés et en fonction des données requises pour satisfaire ces besoins.

Le ministère des Finances a besoin de données pour l'analyse fiscale, c'est-à-dire plus particulièrement de données relatives à la structure juridique des sociétés, au statut fiscal et à l'industrie. La structure juridique d'une société détermine son admissibilité aux diverses dispositions fiscales visant les petites entreprises. Le statut fiscal permet d'estimer l'incidence sur les recettes de toute modification des mesures fiscales relatives aux sociétés imposables. Il arrive que les sociétés se classant dans une industrie donnée bénéficient de dispositions fiscales particulières. En outre, les analystes du ministère des Finances ont besoin d'une classification basée sur la région géographique afin d'assurer une qualité adéquate des estimations des révisions fiscales par région. Comme il y a souvent un lien direct entre la taille des sociétés et le nombre d'éléments d'analyse inclus dans les simulations des politiques fiscales, il est apparu que la taille des sociétés devait également être une priorité. En résumé, les priorités du client sont les suivantes:

comparaisons régionales
secteurs industriels particuliers
grandes et petites entreprises
statut fiscal et structure juridique
estimation de l'incidence des simulations sur les recettes fiscales

Compte tenu de ces spécifications, on a tiré un échantillon aléatoire simple dans une population stratifiée de sociétés. Pour définir les strates, on a utilisé 13 régions fiscales ou géographiques, 25 groupes d'industries, 6 catégories de taille de l'actif variable, 2 catégories de statut fiscal et 2 catégories de structure juridique. L'échantillon actuel comprend quelque 16 000 sociétés pour chacune desquelles on dispose d'un millier de données. La majorité de ces données sont obtenues à l'occasion de la collecte de données statistiques et le reste est tiré du fichier de données administratives sur les sociétés de RCI. À titre d'exemple, ces dernières années, le nombre de données que l'on peut recueillir concernant chaque société se présente comme suit:

Bilan		56
État financier		44
Données financières comparatives et diverses		21
Sous-codes		
Provenance et utilisation des fonds		22
T2S(1)	Conciliation du revenu net	47
T2S(3)	Relevé des dividendes reçus et des dividendes versés	4
T2S(4)	Relevé des pertes	20
T2S(6)	Résumé des dispositions de biens immobilisations	8
T2S(7)	Analyse du revenu	8
T2S(8)	Déduction pour amortissement	164
T2038	Crédit d'impôt à l'investissement	59
T2S(12)	Déduction pour épuisement gagnée	36
T2S(13)	Continuité des réserves	2
T2S(15)	Régimes de revenu différé	4
T2S(27)	Calcul des bénéfices de fabrication et de transformation	11

Le processus manuel d'extraction de ces données est continu et est appliqué tout au long de l'année. Il fait normalement partie des tâches du personnel permanent du centre de données fiscales d'Ottawa. La Division des services statistiques du bureau central assure la direction fonctionnelle, ce qui inclut les éléments suivants:

- allocation des ressources,
- détermination des cibles pour la production et des taux de productivité,
- élaboration des instructions,
- transcription des données,
- acheminement de la charge de travail en temps voulu,
- contrôle de la production et de l'utilisation des ressources,
- contrôle de la qualité des données.

Avant d'exposer les modalités du contrôle de la qualité des données, nous allons voir quelques aspects essentiels de l'utilisation des données administratives.

3. UTILISATION À DES FINS STATISTIQUES DES DOSSIERS DE L'IMPÔT SUR LE REVENU DES SOCIÉTÉS

Le système fiscal canadien est fondé sur la collaboration volontaire des contribuables qui déterminent eux-mêmes l'impôt à payer. Les tâches essentielles de RCI par rapport à ce système consistent à vérifier l'exactitude des calculs faits par les contribuables et à traiter chaque cas de façon équitable et professionnelle.

Pour un organisme dont la responsabilité principale est la gestion des impôts, la collecte de données à des fins statistiques est une préoccupation secondaire. La grande majorité du personnel de RCI est employée à administrer le système fiscal, c'est-à-dire à évaluer l'impôt à payer, à faire les vérifications nécessaires et à percevoir les montants dus. Beaucoup de ces employés ont une formation comptable. Par ailleurs, RCI compte une poignée de statisticiens chargés d'analyser des variables telles que le nombre de contribuables présentant des caractéristiques précises comme particulier ou comme entreprise et l'impôt payé par ceux-ci compte tenu de leurs données financières.

En d'autres mots, le milieu auprès duquel est faite la collecte de données statistiques sur les sociétés est dominé par des exigences et des priorités liées à l'évaluation et à la perception de l'impôt. Le milieu, et le système fiscal à proprement parler, sont des facteurs importants qui influent à la fois sur la qualité des données statistiques et sur les mesures qu'on peut avoir à prendre pour améliorer cette qualité. Dans le cas de RCI, la

couverture de l'enquête, l'accessibilité des données, le taux de réponse et le contrôle témoignent de l'incidence généralement positive de ces facteurs.

En ce qui a trait à la couverture de l'enquête, le fichier de données administratives de RCI peut être considéré comme un recensement intégral de la population des sociétés. Il contient, pour chaque année d'imposition, un relevé sur bande magnétique des données sur chaque société qui remplit une formule T2 au Canada. Toutes les entreprises constituées en société, qu'elle soient imposables ou actives, doivent remplir une formule T2. En conséquence, pour chaque année d'imposition, le fichier contient des données sur environ 700 000 sociétés.

Les données sur chaque société imposable saisies et stockées dans ce fichier sur bande sont le fruit de l'application des procédures administratives du ministère et comportent les limites inhérentes à cette caractéristique. La majorité des données détaillées contenues dans les déclarations d'impôt et dans les états financiers normalement joints à la formule T2 ne sont pas introduites dans le fichier de données administratives. Étant donné que le fichier de données administratives sur les sociétés est très vaste mais très peu détaillé, il ne peut être utilisé avec profit pour effectuer des micro-simulations. Il faut donc tirer un échantillon pour les micro-simulations.

Les statisticiens de RCI et d'autres personnes ont directement accès au fichier de données administratives et aux formules T2 pour des fins statistiques. Bien entendu, la priorité d'accès va d'abord aux évaluateurs et aux vérificateurs de RCI. Pour ce qui est des autres personnes, y compris les statisticiens des autres ministères, l'accès aux données du fichier est rigoureusement régi par les dispositions de la Loi de l'impôt sur le revenu qui vise à préserver le caractère privé et confidentiel des données fournies par les contribuables.

Le taux de réponse n'est pas non plus un problème véritable. Du fait que l'échantillon est tiré à partir d'une population cible qui est un sous-ensemble du fichier de données administratives, on peut donc avoir des données sur toutes les unités déclarantes sélectionnées. Un léger problème peut se présenter, de façon très occasionnelle, lorsque la formule T2 sélectionnée est entre les mains d'un vérificateur. Toutefois, le problème est contourné en photocopiant la formule T2 originale, sélectionnée pour faire partie de l'échantillon, une fois qu'elle a été vérifiée par l'ordinateur. Si une unité déclarante a défié la loi et n'a pas rempli de formule T2 pour une année d'imposition particulière, elle ne figure pas dans le fichier de données administratives. En conséquence, elle ne peut être sélectionnée pour faire partie de l'échantillon. Or, les risques qu'un tel problème se pose sont minimes car la grande majorité des sociétés imposables remplissent une déclaration.

De la même façon, le manque de contrôle sur les unités déclarantes, sur les données, sur le taux de réponse et sur l'accès aux données ne constitue nullement un problème. Il convient de rappeler, toutefois, que le système administratif d'évaluation et de traitement des déclarations d'impôt a toujours préséance sur l'utilisation statistique des données.

En résumé, la couverture, l'accessibilité des données, le taux de réponse et le contrôle ne posent pas véritablement de problèmes, principalement du fait que les utilisateurs administratifs aussi bien que statistiques des dossiers fiscaux des sociétés appartiennent au même ministère et s'intéressent au même univers d'entreprises.

Les besoins opérationnels de RCI et l'univers des entreprises qui l'intéresse influent sur les activités permettant de prévoir, d'identifier et de corriger les erreurs contenues dans les dossiers fiscaux des sociétés qui servent à des fins statistiques. Ces activités sont décrites ci-après.

4. CONTRÔLE DE LA QUALITÉ

Au départ, nous avions l'intention de décrire un projet possible de contrôle de la qualité appliqué aux données fiscales et financières des sociétés. Cela nous a amené à définir ce que l'on entend par qualité et à voir de quelle manière on peut mesurer la qualité des statistiques sur les sociétés. Cependant, comme le nombre de données en cause se monte à environ 700, il est apparu irréaliste d'envisager de mesurer la qualité de chacune. Nous avons donc défini la qualité d'un produit comme étant la somme de toutes les caractéristiques qui en déterminent l'acceptation compte tenu des fins pour lesquelles le produit a été conçu, de ses spécifications et des priorités de l'utilisateur. Cette définition est relativement claire et précise mais elle ne dit pas de quelle manière la qualité peut être mesurée. Au lieu d'exposer la méthode idéale de mesure de la qualité, nous avons choisi de décrire plusieurs méthodes de contrôle de la qualité actuellement appliquées au moment de la collecte des données sur les sociétés.

Le contrôle de la qualité vise à assurer que le produit répond aux spécifications établies à partir des besoins du client. Les spécifications des données qui doivent être recueillies sont exposées en détail dans le Manuel d'analyse des données sur les sociétés de RCI. Ce manuel décrit les procédures utilisées par les analystes du centre de données fiscales d'Ottawa pour extraire les données des rapports d'impôt fournis par les sociétés et des états financiers qui y sont joints.

Il existe nombre de sources d'erreurs possibles qui risquent d'entamer la précision des données. La première source d'erreur est le contribuable lui-même. L'auteur de la déclaration peut s'être trompé en remplissant la formule d'impôt ou les états financiers. Quand la formule T2 arrive à RCI, des évaluateurs vérifient l'exactitude des données relatives au calcul de l'impôt en se fondant sur les niveaux de tolérances établis. Cette façon de procéder peut nuire à la précision requise lorsque les données sont utilisées à des fins statistiques. Bien que les évaluateurs recueillent quelques données statistiques pour les besoins d'un certain nombre d'utilisateurs de RCI et de Statistique Canada, ces dernières ne sont pas soumises à un contrôle automatique et peuvent aussi être erronées. Le fichier principal de données administratives du Ministère est constitué à partir des transactions effectuées par les évaluateurs.

Un manque d'uniformité dans l'interprétation des spécifications ou les pressions dues à la charge de travail peuvent être à l'origine d'erreurs dans les données extraites. Une insistance trop grande sur la production quantitative peut également nuire, à l'occasion, à la qualité des données extraites.

Une fois que les données sur les sociétés ont été analysées par les spécialistes, vérifiées et transcrites, la feuille de données transcrites est envoyée à la vérification par clavier afin d'être introduites dans l'ordinateur. Des erreurs peuvent être commises à cette étape également.

Dans la suite du processus, on peut aussi découvrir des erreurs résultant de l'application du plan de l'échantillon. Il arrive que les déclarations des sociétés sélectionnées pour faire partie de l'échantillon soient placées dans la mauvaise strate. Les observations que l'on tire du fichier sont alors inappropriées ou incohérentes par rapport au but du plan de l'échantillon.

Nous disposons de sept activités différentes pour compenser ces erreurs possibles et assurer la qualité des données sur les sociétés. Au centre de données fiscales d'Ottawa, ces activités consistent en l'examen par des analystes des fichiers de données administratives et des données transcrites, la vérification des données saisies, le contrôle automatique et la correction manuelle. À la Division des services statistiques du bureau central, nous évaluons la qualité des données après la collecte, nous identifions les secteurs problèmes et nous décidons des mesures correctives à prendre. Les deux dernières activités de contrôle de la qualité consistent à vérifier toutes les observations

propres à l'échantillon, ajoutées au système tout au long de l'année, et à comparer les statistiques sommaires obtenues aux données provenant d'autres sources. Chacune de ces sept activités, décrites ci-après, permet d'améliorer la qualité des données fournies au ministère des Finances pour l'analyse des politiques fiscales.

L'examen des données administratives du fichier est effectué par les analystes des données sur les sociétés, au centre de données fiscales d'Ottawa. Pour chaque déclaration d'impôt sélectionnée pour faire partie de l'échantillon, on prend quelque quatre-vingt données contenues dans le fichier principal du Ministère et on les imprime sur une feuille de transcription particulière. Ces données sont ensuite contre-vérifiées au moyen du document source, c'est-à-dire la formule T2 et l'état financier correspondants. S'il y a une différence, l'analyste fait la correction requise. Lorsque l'erreur porte sur les données du fichier, la correction est apportée au fichier et également au fichier-échantillon avant son extraction.

Une fois les données transcrites, un analyste supérieur expérimenté effectue une révision manuelle des données transcrites. Les erreurs identifiées sont transmises à l'analyste concerné et corrigées avant que les données transcrites soient introduites par clavier dans l'ordinateur. Le pourcentage de données à réviser inclus dans la charge d'un analyste dépend du rendement de l'analyste, de son expérience et du nombre d'années-personnes disponibles.

L'activité de contrôle de la qualité suivante a lieu au moment de l'introduction par clavier des données transcrites. Un certain nombre de vérifications alphanumériques sont directement effectuées et les erreurs sont corrigées. Les données ainsi corrigées sont ensuite introduites en machine dans le système informatique qui constitue le fichier de données-échantillon sur les sociétés avec les observations-échantillon. À cette étape, une série de vérifications sont effectuées, vérification des totaux arithmétiques, vérification de la logique des données, comparaison des données des diverses zones, etc. Lorsque des erreurs sont repérées, le document source et les données transcrites sont révisés par l'analyste.

Au bureau central, des spécialistes passent en revue un lot de déclarations d'impôt et d'états financiers d'où sont tirées les données qui ont été soumises aux quatre étapes précitées et acceptées comme données épurées. Ces déclarations d'impôts sont sélectionnées par échantillonnage statistique, les objectifs visés variant selon les années; par exemple, l'accent peut être mis sur un groupe d'industries particuliers ou sur une formule d'impôt particulière. Une fois les erreurs repérées et corrigées, dans lequel on recommande des mesures correctives, comme l'amélioration des instructions que doivent suivre les analystes et la correction du programme informatique.

L'activité de contrôle de la qualité suivante est appelée "validation finale". Tout au long de l'année, à l'occasion de la collecte des données à partir des déclarations d'impôt remplies par les sociétés, on relève certaines incohérences dans les spécifications ou dans les vérifications. De nouvelles vérifications doivent donc être ajoutées en cours d'année. Celles-ci révèlent parfois des erreurs additionnelles qui doivent être corrigées et la déclaration d'impôt doit être re-sélectionnée et re-travaillée.

Enfin, les données du fichier sont mises en tableaux et celles de plusieurs zones sont totalisées. Ces derniers totaux sont alors comparés à ceux des fichiers-échantillon des années antérieures et à d'autres sources de données, telles que les données publiées par Statistique Canada dans "Statistique fiscale des sociétés" et les données intégrales du fichier de données administratives de RCI. Les écarts importants inexplicables font l'objet d'une recherche et les données du fichier sont ajustées en conséquence.

5. PERSPECTIVES

Dans l'avenir, nous prévoyons apporter un certain nombre de changements qui auront une incidence sur la qualité des données. Nous avons entrepris une étude de faisabilité de la saisie automatique des données en vue de remplacer le système actuel de saisie et de correction par clavier. Avec un système automatique, la personne qui entre les données par clavier soumet celles-ci à un programme informatique qui effectue les vérifications et repère les erreurs qui sont immédiatement corrigées par la personne qui a introduit les données par clavier.

Nos activités actuelles de contrôle de la qualité sont fondées principalement sur une planification de la production et des méthodes de contrôle de type classique et elles impliquent uniquement des techniques limitées telles que la sélection des déclarations d'impôt par échantillonnage aléatoire. Nous envisageons de faire une étude de faisabilité portant sur l'introduction de méthodes statistiques pour mesurer la qualité des données et pour déterminer les niveaux de tolérance à appliquer pour établir l'acceptabilité des données.

S'il est décidé d'adopter la saisie automatique des données et des techniques statistiques plus complexes, nous espérons avoir une plus grande assurance que la qualité des données fournies est appropriée à toutes les fins pour lesquelles celles-ci sont recueillies.

REMERCIEMENTS

Nos remerciements s'adressent en particulier aux spécialistes de la Section de la statistique des sociétés, Jean Wyman, Wendy Blais et Madeleine Gadoua, qui nous ont fourni des renseignements sur les activités de contrôle de la qualité existantes.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

UTILISATION DES DONNÉES DE DOSSIERS ADMINISTRATIFS POUR L'ÉVALUATION DE LA QUALITÉ DES ESTIMATIONS D'ENQUÊTES

JEFFREY C. MOORE et KENT H. MARQUIS[1]

RÉSUMÉ:

La SIPP (Survey of Income and Program Participation/enquête sur le revenu et la participation aux programmes) est une nouvelle enquête par panel du Census Bureau destinée à fournir des données sur la situation économique des particuliers et des familles aux États-Unis. Chaque ménage de la SIPP est interviewé huit fois au cours de l'existence prévue de deux ans et demi du panel, soit tous les quatre mois. La donnée de base de la SIPP est le revenu mensuel qui est déclaré pour chacun des quatre mois de la période de référence qui précède le mois de l'interview. L'étude de vérification des enregistrements SIPP utilise les données des dossiers administratifs pour évaluer la qualité des estimations de la SIPP pour un grand nombre de sources de revenus et de programmes de transfert. Le projet utilise des techniques d'appariement statistiques pour identifier les personnes échantillonnées dans quatre États qui, selon les dossiers, ont reçu des paiements de n'importe lequel de neuf programmes administrés par l'État ou par le gouvernement fédéral. On compare ensuite les dates et les montants des paiements déclarés lors de l'enquête aux données correspondantes des dossiers officiels. Le document décrit les principaux points dont on a tenu compte lors de l'élaboration du projet et présente quelques-unes des premières conclusions.

1. INTRODUCTION

La présente communication traite des questions relatives à l'utilisation des données de dossiers administratifs pour évaluer la qualité des estimations fondées sur des données d'enquête et décrit une application précise aux données de la SIPP aux États-Unis.

L'appariement des données des dossiers administratifs et des observations de l'enquête effectué pour chaque cas et que nous appelons "vérification des enregistrements", fournit des renseignements utiles aux personnes chargées de l'élaboration de l'enquête et à celles qui utilisent ses données. Une vérification des enregistrements permet à l'analyste de faire une gamme complète d'estimations des paramètres des erreurs de mesure à des fins d'évaluation. Ces estimations facilitent à leur tour la réalisation de deux genres d'activités fondamentales:

[1] Jeffrey C. Moore et Kent H. Marquis, Center for Survey Methods Research, U.S. Bureau of the Census, pièce 2737 FB3. Washington, DC 20233. États-Unis.

1. l'ajustement des estimations concernant des sujets spécialisés tels que les moyennes, les proportions, les coefficients de corrélation et les coefficients de régression à plusieurs variables afin de corriger les erreurs de mesure; et
2. l'établissement de plans de sondage plus efficaces qui tiennent compte, par exemple, de l'équilibre à maintenir entre la qualité des mesures et les coûts.

1.1 Principaux termes utilisés

Nous nous concentrerons ici sur les erreurs de mesure ou de réponse, bien que la méthode de vérification des enregistrements puisse aussi servir à évaluer d'autres types d'erreurs non dues à l'échantillonnage et d'erreurs d'échantillonnage. Nous ne nous présenterons pas un exposé technique, mais il est quand même nécessaire de définir en premier lieu certains des principaux termes que nous employons. Nous supposons que l'observation d'enquête à partir du point d'échantillon i peut être exprimée comme étant la somme de la valeur vraie et d'une erreur, e: observation d'enquête$_i$ = valeur vraie i + e_i.

Le biais moyen dans un ensemble d'observations d'enquête N, que nous appelons l'erreur systématique de réponse ou déviation systématique de l'enquête, est $\bar{e} = \Sigma_i / N$, et la variance des erreurs de réponse est simplement Var e.

De même, le modèle servant à mesurer une observation fondée sur un dossier administratif est: dossier$_i$ = valeur vraie$_i$ + u_i, de sorte que le biais relatif au dossier est $\bar{u}$ et la variance des erreurs dans le dossier est Var u.

1.2 Comparaison des méthodes d'évaluation

La vérification des enregistrements présente des aspects qui peuvent être comparés à ceux d'autres méthodes d'évaluation telles que les réinterviews et les expériences. Les réinterviews et autres plans de mesures répétitives servent à estimer un ensemble très limité de paramètres des erreurs de mesure; c'est ce qu'on appelle habituellement la variance de réponse simple ou la variance des erreurs de réponse. Ces méthodes sont fondées implicitement sur des hypothèses solidement étayées concernant les changements réels qui surviennent avec le temps et la notation valable ou le paramètre de biais (Marquis, 1986).

Une solution que l'on utilise souvent consiste à introduire des mesures servant de critères dans le programme de réinterview, par exemple en comparant des réponses divergentes obtenues avec celles que fournit un répondant bien informé ou en posant des questions beaucoup plus détaillées et précises au cours de la réinterview. Mais la validité de telles mesures peut être mise en doute. Bailer (1986) et Koons (1973) ont démontré, par exemple, que les réponses de réinterview ayant fait l'objet d'un rapprochement sont affectées d'une erreur systématique. Et bien que l'on préfère souvent un questionnaire comportant des questions précises et détaillées à une approche plus globale, il n'existe aucune preuve en soi indiquant qu'une telle méthode permettrait de supprimer, ou même de réduire, le biais. Les vérifications d'enregistrements peuvent fournir de meilleures données-critères, nécessitant des hypothèses beaucoup moins solides (et peut-être plus réalistes) pour l'estimation de la qualité des données d'enquête.

L'expérience est une autre méthode d'évaluation des aspects d'une enquête. On peut avoir recours, par exemple, à un plan factoriel entièrement croisé ou à un plan de sondage à réseaux superposés pour l'affectation des intervieweurs. Les analystes comparent les groupes visés par l'expérience et plus précisément, les statistiques concernant, par exemple, les moyennes ou les proportions relatives aux sujets spécialisés, et déterminent les taux de déclaration relatifs aux domaines spécialisés d'intérêt qui sont obtenus par les différentes méthodes. Toutefois, il y a controverse lorsqu'il faut établir laquelle des méthodes est la meilleure pour effectuer des mesures. Cette difficulté se trouve

considérablement aplanie lorsque nous disposons de données-critères, par exemple les dossiers administratifs.

Sans données servant de critères, l'analyste doit souvent s'en remettre à des hypothèses reconnues au sujet des erreurs de mesure, par exemple:

- plus il y a de répondants, meilleurs sont les renseignements;
- la possibilité d'oublier des faits importants augmente avec le temps;
- les interviews sans période de référence donnent lieu à des déclarations excessives alors que ce problème ne se pose pas dans le cas des interviews qui en comportent;
- la qualité des déclarations diminue lorsque l'interview se prolonge ou que le répondant fait partie de l'échantillon depuis un certain temps;
- les gens ont tendance à être paresseux et à avoir l'esprit tortueux, ils mentiront pour éviter d'avoir à répondre à une série détaillée de questions; et
- les déclarations faites par le répondant visé sont préférables à celles qui sont faites par un enquêté-substitut.

De fait, ces hypothèses sont devenues les fondements classiques des plans de sondage dans le monde occidental. Et pourtant, il est difficile d'étayer ces hypothèses quand on procède à des vérifications d'enregistrements appropriées. Les expériences et autres techniques du genre constituent d'excellents moyens pour déterminer exactement les sources de la variation et pour démêler les problèmes d'estimation de la colinéarité, mais s'avèrent souvent inutiles et rarement suffisants pour évaluer un processus de mesure établi.

En somme, ces autres méthodes d'évaluation nous obligent à nous reposer en grande partie sur la supposition: 1) que la mesure initiale et la mesure d'évaluation sont indépendantes alors qu'elles sont clairement dépendantes; 2) qu'il existe un rapport entre la mesure initiale et un critère donné alors qu'il n'y a aucun lien externe objectif; et (ou) 3) que les processus cognitifs sont valables alors qu'ils ne sont pas étayés par des recherches.

Pour les vérifications d'enregistrements, on utilise aussi des hypothèses pour évaluer les mesures. Ainsi, la façon habituelle d'estimer l'erreur systématique de réponse consiste à supposer qu'il n'y a pas de biais relatif aux dossiers ($\bar{u} = 0$) et de prendre la moyenne des écarts entre les valeurs observées de l'enquête et du dossier qui sont comparées: déviation systématique estimative de l'enquête = $\Sigma(S_i - R_i) / N$. Bien qu'il ne soit pas possible de corroborer directement l'hypothèse selon laquelle il n'y a pas de biais relatif aux dossiers, on peut effectuer des tests de sensibilité pour déterminer les effets sur les conclusions de l'évaluation de la non-vérification de cette hypothèse. (Éventuellement, l'étude de vérification des enregistrements SIPP utilisera ces tests et d'autres analyses pour évaluer les erreurs dans les dossiers.)

1.3 Points à retenir pour l'élaboration d'un plan de vérification d'enregistrements

Plusieurs points doivent être pris en considération lors de l'élaboration d'un plan de vérification d'enregistrements pour évaluer la qualité des données d'enquête, notamment les plans d'observation incomplets, les erreurs d'appariement, les erreurs dans les dossiers, les différences dans les valeurs vraies et l'absence de mesures répétitives ou de caractéristiques du plan d'expérience.

1.3.1. Plans d'observation incomplets

Les vérifications d'enregistrements antérieures ont souvent été faites à l'aide d'un plan unidirectionnel ou partiel établi pour la collecte des données, par exemple lorsque nous faisons enquête auprès des gens pour savoir s'ils possèdent une carte de bibliothèque et que nous vérifions les dossiers pour trouver ceux qui ont dit en avoir une, ou encore lorsque nous prélevons un échantillon à partir d'une liste de personnes souffrant d'une maladie chronique diagnostiquée et que nous les incluons dans une enquête afin de voir si ces personnes déclareront leur maladie dans le questionnaire d'enquête. Étant donné que ces plans partiels ne tiennent pas compte, dans les bonnes proportions, de la gamme complète des erreurs de réponse, ils produisent des estimations biaisées des paramètres classiques des erreurs de mesure, tels que l'erreur systématique de réponse et la variance des erreurs de réponse. Les plans unidirectionnels peuvent ne pas déceler une partie ou la totalité de la déviation systématique réelle de l'enquête, porter l'analyste à considérer jusque la moitié de la variance des erreurs de réponse comme une erreur systématique de réponse et déterminer à l'avance la valeur positive ou négative de l'erreur systématique estimative de réponse si la variable mesurée est binaire (Marquis, 1978). Il est nécessaire (mais non suffisant) que les plans soient complets pour obtenir des estimations non faussées des erreurs de réponse.

1.3.2. Erreurs d'appariement

L'objet de la vérification d'enregistrements consiste à apparier une à une les données d'enquête et des données de source administrative. C'est une opération qui est difficile à faire correctement, et les erreurs d'appariement (faux appariements, faux non-appariements) peuvent entraîner un biais dans les estimations des erreurs de mesure. Neter et ses collaborateurs (1965) ont montré que lorsqu'il n'y a pas de cas de non-appariement, les mauvais appariements donnent lieu à une erreur systématique par excès affectant la variance des erreurs de réponse. Pour ce qui est de la sûreté d'une mesure dichotomique (qui est une fonction de la variance des erreurs de réponse), l'estimation est réduite dans une proportion équivalente au taux des erreurs d'appariement (Marquis et coll., 1986). Par conséquent, il serait bon qu'il n'y ait qu'un minimum d'erreurs d'appariement et que l'on soit renseigné sur les erreurs qui restent.

1.3.3. Erreurs dans les dossiers administratifs

Comme nous l'avons mentionné précédemment, on peut habituellement compter sur le fait que les dossiers qui font l'objet d'une étude de vérification sont d'excellentes mesures du sujet d'intérêt. Si les hypothèses implicites concernant le biais relatif à la mesure fondée sur des dossiers et la variance des erreurs de la mesure se montrent inexactes, cela peut introduire un biais dans les estimations des erreurs de réponse. Par exemple, un biais touchant les données observées dans un dossier peut se manifester par un biais affectant les données observées de l'enquête, mais le biais serait de valeur contraire. Feather (1972) a décrit cet effet dans le cadre d'une vérification des dossiers des visites des médecins de la Saskatchewan. Le taux apparemment élevé de déclaration par excès enregistré au cours de l'enquête a été attribué au fait que le dossier contenait de l'information portant sur l'ensemble du traitement plutôt que sur les visites individuelles faites pour obtenir un diagnostic. De même, la présence d'une variance des erreurs de mesure dans les dossiers peut gonfler les estimations de la variance des erreurs de réponse dans l'enquête (Marquis, 1978).

1.3.4 Différences dans les valeurs vraies

Des problèmes surviennent lorsque les définitions de l'enquête et du système de dossiers administratifs diffèrent. C'est souvent le cas lorsque des "comparaisons agrégatives" des estimations des paramètres de la population sont faites séparément par chaque source. Une des différences entre les deux systèmes est l'étendue des populations visées. Ainsi, la base de l'enquête peut être limitée à la population civile hors institution alors que les données des dossiers administratifs peuvent porter sur l'ensemble de la population. L'appariement de chaque cas peut contribuer à réduire les problèmes causés par la différence de champ d'observation, mais même les estimations faites à partir de ces études peuvent encore être affectées par le fait que les concepts ou les caractéristiques d'un concept diffèrent. Par exemple, nos dossiers administratifs indiquent souvent la date d'émission d'un chèque établi en vue d'un paiement de transfert et les répondants de l'enquête SIPP nous donnent la date à laquelle ils ont reçu le paiement. De telles différences peuvent nuire considérablement à nos estimations chronologiques concernant, par exemple, les erreurs de réponse de téléscopiage.

1.3.5 Absence d'expériences et de réinterviews

Les études de vérification d'enregistrements peuvent déceler les erreurs, mais n'indiquent pas comment les corriger. Pour déterminer l'efficacité d'un plan de sondage différent, il faut habituellement tester les autres plans qui pourraient être utilisés ou estimer les paramètres d'un modèle sous-jacent à partir duquel des plans de sondage peuvent être établis (par ex., un modèle des effets du manque de mémoire). Ainsi, un plan de vérification d'enregistrements peut servir à estimer et à comparer les erreurs de réponse chez les répondants eux-mêmes et chez les enquêtés-substituts. Toutefois, sans hypothèses solides, un tel plan ne permet pas de déterminer dans quelle mesure les paramètres des erreurs de mesure changeraient si les règles de déclaration de l'enquête étaient modifiées (par ex., seuls les répondants visés pourraient fournir les réponses).

De même, une vérification des enregistrements faite sans réinterview ou autre série de mesures indépendantes ne peut servir qu'à estimer un nombre limité de paramètres des erreurs de base. Ainsi, nos définitions initiales comportaient trois paramètres: la valeur vraie, les erreurs dans les données d'enquête et les erreurs dans les dossiers. Sans réinterview (ou toute autre mesure indépendante), il n'y a que deux mesures qui permettent d'estimer ces trois inconnues. Une mesure additionnelle comme la réinterview peut aider à reconnaître les estimations des paramètres du modèle.

2. CARACTÉRISTIQUES DE L'ENQUÊTE SIPP

Nous ferons ici une brève description des caractéristiques de l'enquête avant de discuter du plan de vérification des enregistrements.

2.1 Aperçu du contenu de la SIPP

La SIPP a pour objet de fournir des renseignements plus précis sur la situation économique des particuliers et des ménages aux États-Unis. Elle permet de recueillir des données longitudinales complètes sur le revenu en espèces et toute forme d'aide autre que financière, sur l'admissibilité et la participation aux programmes de transfert du gouvernement, sur les éléments d'actif et de passif, sur l'activité et sur une foule de sujets connexes. Les données SIPP aident à l'évaluation du coût et de l'efficacité des programmes actuels de l'administration fédérale, de l'incidence éventuelle des changements que l'on propose d'apporter aux programmes et des effets réels de ces changements une fois qu'ils auront été appliqués. En général, le Census Bureau et d'autres

organismes gouvernementaux qui ont encouragé et appuyé l'élaboration de la SIPP s'attendent à ce que celle-ci soit d'une aide précieuse pour la planification de la politique intérieure (Nelson et coll., 1985).

Les questions de base de la SIPP, qui sont répétées à chaque vague d'interview, portent sur l'activité sur le marché du travail ainsi que sur le revenu selon la source et le montant, y compris les paiements de transfert et l'aide autre que financière accordée dans le cadre de divers programmes pour chaque mois de la période de référence. Les questions de base portent sur presque cinquante sources de revenu, y compris les paiements de transfert gouvernementaux des fonds de retraite, les allocations d'invalidité et les prestations d'assurance-chômage et les programmes sociaux qui assurent, par exemple, une aide financière aux familles à faible revenu avec enfants à charge. Des renseignements sont aussi recueillis sur des programmes d'assistance non financière tels que le programme de distribution de bons alimentaires, Medicare et Medicaid; sur les transferts privés comme les prestations de retraite, les pensions alimentaires et les paiements pour garde d'enfant(s), la propriété de biens qui génèrent des revenus tels que des intérêts, des dividendes, des loyers et des redevances, et diverses autres sources de revenu, par exemple les successions.

2.2 Plan de collecte des données SIPP

La SIPP a été menée pour la première fois en octobre 1983 et visait un échantillon d'environ 25,000 unités de logement (le "panel de 1984") choisies pour représenter la population américaine hors institution. En février 1985, l'enquête a été menée auprès d'un nouveau panel légèrement moins important. Des panels supplémentaires sont censés être introduits dans le champ de l'enquête en février de chaque année pour toute la durée de l'enquête. En raison de restrictions budgétaires, la taille de l'échantillon des nouveaux panels s'élève actuellement à environ 15,000 ménages.

Chaque ménage de l'échantillon est interviewé au cours d'une visite à domicile tous les quatre mois pendant deux ans et demi, ce qui équivaut à un total de huit interviews par ménage. La période de référence pour chaque interview correspond aux quatre mois qui précèdent le mois d'interview. Lors de chaque visite, chaque membre du ménage de 15 ans et plus doit fournir des renseignements le concernant. Les déclarations faites au nom d'un membre absent au moment de la visite sont permises. Les renseignements concernant les déclarations faites par un enquêté-substitut sont consignés et disponibles à des fins d'analyse.

Pour faciliter les opérations sur le terrain, chaque panel-échantillon est divisé en quatre sous-échantillons (groupes de renouvellement) qui ont à peu près la même taille et dont un est interviewé chaque mois. Ainsi, une "vague" (cycle) d'interview est complétée après une période de quatre mois pour chaque panel. Grâce à ce plan d'enquête, les activités sur le terrain et les tâches de dépouillement se déroulent plus régulièrement; par contre, chaque groupe de renouvellement se trouve à avoir une période de référence de quatre mois différente.

A partir de la deuxième vague d'interview du panel de 1984, des réinterviews sont menées auprès d'un petit échantillon de ménages qui doivent répondre à une sous-série de questions (y compris sur leur participation aux programmes). Ces données servent principalement à relever les erreurs de l'intervieweur, mais peuvent aussi aider à estimer les réponses incohérentes.

3. PLAN DE VÉRIFICATION DES ENREGISTREMENTS

L'objet de la vérification des enregistrements est de fournir une évaluation d'une partie des données recueillies dans le cadre de la SIPP. Nous allons maintenant mettre l'accent sur les principales caractéristiques du plan de vérification des enregistrements et traiter des aspects suivants: les échantillons, les dossiers administratifs, la méthode d'appariement et l'analyse.

3.1 Échantillons de la vérification des enregistrements

La vérification des enregistrements de la SIPP est fondée sur un plan "complet" plutôt qu'unidirectionnel. En d'autres mots, les enregistrements dont nous disposons nous permettent de valider toutes les données observées de l'enquête. D'autres plans dont nous n'avons pas tenu compte consistaient: 1) à vérifier uniquement les enregistrements des personnes qui déclarent participer à un programme, ou 2) à tirer un échantillon de bénéficiaires connus et à interviewer ces derniers afin de déterminer si les renseignements qu'ils ont fournis sont vrais. Ces deux types de plan sont incomplets et auraient faussé les estimations des paramètres des erreurs de réponse.

L'étude de vérification des enregistrements porte sur un sous-ensemble de données SIPP fournies par le panel de 1984. Premièrement, l'échantillon est limité aux ménages résidant dans quatre États: la Floride, New York, la Pennsylvanie et le Wisconsin. Pour le panel de 1984, cela équivaut à approximativement 5,000 ménages. Deuxièmement, la période de référence de l'étude correspond seulement aux mois d'enquête des deux premières vagues du de 1984. La figure 1 montre la vague, le groupe de renouvellement, le mois d'interview et la période de référence qui correspondent aux données cibles de l'enquête.

Vague	Groupe de renouvellement	Mois de l'interview	Mois de la période de référence											
			Juin	Juil.	Août	Sept.	Oct.	Nov.	Déc.	Janv.	Fév.	Mars	Avr.	
1	1	Oct. 83	X	X	X	X								
	2	Nov. 83		X	X	X	X							
	3	Déc. 83			X	X	X	X						
	4	Janv. 84				X	X	X	X					
2	1	Fév. 84					X	X	X	X				
	2	Mars 84						X	X	X	X			
	3	Avr. 84							X	X	X	X		
3	4	Mai 84								X	X	X	X	

Figure 1: Structure de l'enquête pour les données visées par l'étude de vérification des enregistrements SIPP

Troisièmement, l'étude de vérification des enregistrements met l'accent sur la qualité des données sur la participation aux programmes et sur les sommes versées dans le cadre de certains programmes de transfert gouvernementaux. Nous comparons les documents d'enquête et les dossiers administratifs concernant cinq programmes fédéraux (Federal Civil Service Retirement/fonds de retraite des fonctionnaires fédéraux; Pell Grants/subventions Pell; Social Security - OASDI/sécurité de la vieillesse, pensions de survivant et assurance-invalidité; Supplemental Security Income/revenu supplémentaire de sécurité sociale; et Veterans' Compensation and Pensions/indemnités et pensions des

anciens combattants) et quatre programmes administrés par les États (Aid to Families with Dependant Children/aide aux familles à faible revenu avec enfants à charge; les bons alimentaires; les prestations d'assurance-chômage et l'indemnisation des accidents du travail).

Nous avons limité l'étude à quatre États, soit la Floride, New York, la Pennsylvanie et le Wisconsin, afin d'avoir des proportions plus faciles à traiter. La sélection de ces États a été faite en fonction des critères suivants: 1) l'existence d'un système de dossiers automatisé, accessible et complet pour tous les programmes visés; 2) un vaste échantillon SIPP; 3) une diversité géographique appropriée, et 4) le consentement au partage des données individuelles pour les besoins de notre étude. Ainsi, les États ont été choisis intentionnellement, aucune tentative n'a été faite pour choisir des États représentatifs du pays.

Nous avons demandé à chaque organisme participant de ces États de fournir des données d'identification sur toutes les personnes qui ont reçu un revenu au titre du programme cible entre mai 1983 et juin 1984 ainsi que les données sur les sommes reçues. La même demande a été adressée aux organismes fédéraux participants, mais seules les données sur les bénéficiaires résidant dans un des quatre États choisis devaient être fournies.

Nous avons obtenu ces dossiers administratifs, étant entendu que la confidentialité des données obtenues serait assurée au même titre que les données recueillies en vertu de l'intitulé 13 du U.S. Code. Ainsi, seuls les employés assermentés du Census Bureau qui sont affectés à l'étude de vérification peuvent consulter ces dossiers. Sauf dans le cas des données sommaires qui ne permettent pas d'identifier les particuliers, le contenu des dossiers ne doit pas être diffusé ou divulgué à d'autres pour quelque raison que ce soit.

Certains organismes ont décidé d'envoyer les données requises en deux étapes, ne fournissant initialement que les données d'identification des bénéficiaires, sans (ou presque sans) renseignements sur les sommes reçues dans le cadre du programme. Après l'appariement de ces données et de celles de la SIPP, les responsables du projet renvoient à l'organisme une liste des codes d'identification s'appliquant aux personnes ayant fait l'objet de l'appariement (plus un nombre suffisant de codes de cas non appariés afin d'assurer la confidentialité de l'échantillon SIPP). L'organisme extrait les données sur les paiements reçus par ces personnes et les fait parvenir au Census Bureau.

Comme il a été mentionné précédemment, les erreurs contenues dans les dossiers peuvent nuire aux études de vérification des enregistrements. Bien que plusieurs des fichiers de données administratives obtenus pour ce projet comportent quelques légères failles (par exemple, l'initiale du deuxième prénom n'est pas indiqué, le sexe n'est pas précisé, l'âge est donné plutôt que la date de naissance, etc.), seulement trois d'entre eux semblent pouvoir poser de sérieux problèmes sur le plan analytique. Deux fichiers présentent un relevé incomplet des bénéficiaires: le fichier de l'indemnisation des accidents du travail de l'État de New York et le fichier des indemnités et pensions des anciens combattants dont les données portent sur les quatre États. Le premier fichier ne tient pas compte d'un nombre inconnu de cas considérés comme "réglés" (c.-à-d. de cas sur lesquels on s'était déjà prononcé et qui avaient déjà commencé à recevoir des paiements versés par une entreprise d'assurance privée) au moment où la base de données a été établie il y a plusieurs années. Le deuxième fichier ne contient pas de données sur les bénéficiaires dont les prestations ont été envoyées à un établissement financier ou autre, ces bénéficiaires représentant environ un pour cent de l'ensemble des bénéficiaires. Les autres fichiers ne semblent pas présenter de problèmes de cet ordre. Le troisième fichier assure un dénombrement complet, mais ne contient pas les renseignements relatifs à l'adresse des bénéficiaires, ces données s'avèrant très utiles au moment de l'appariement.

L'écart entre la date d'émission du chèque et la date de réception constitue un problème inévitable qui touche tous les fichiers administratifs dans une certaine mesure.

De toute évidence, le répondant SIPP déclare la date de réception du paiement et n'est pas au courant de la date d'émission tandis que l'inverse se produit dans le cas des dossiers relatifs au programme. Lorsque la date d'émission du chèque se situe vers la fin d'un mois, il peut être difficile de distinguer une erreur de téléscopage croissant d'un écart légitime entre le mois de l'émission et le mois du paiement. Lorsqu'il y a des différences dans les définitions, par exemple comme pour cette question des dates d'émission/paiement, nous tenterons d'en présenter des modèles explicites dans nos analyses.

4. APPARIEMENT

4.1 Introduction

La qualité de l'appariement a une incidence considérable sur certaines des estimations des erreurs de réponse les plus importantes telles que la variance de l'erreur de réponse. Idéalement, les variables utilisées pour apparier les données de l'enquête et celles des dossiers seraient mesurées sans erreur et permettraient d'identifier un particulier. Il est entendu qu'un tel idéal n'est jamais atteint.

Toutefois, les variables dont nous disposons pour l'appariement des données de l'enquête et des dossiers devraient contribuer à réduire considérablement les erreurs d'appariement. Certaines d'entre elles, par exemple le numéro de sécurité sociale (SSN), permettent d'identifier de façon unique un particulier, même si d'autres renseignements comme l'adresse sont périmés, illisibles, effacés ou manquants. Pour des raisons qui n'ont pas de rapport direct avec la présente étude (mais qui peuvent certainement lui être profitables), le Census Bureau a pris des mesures spéciales pour s'assurer que le SSN déclaré lors de la SIPP est complet et valide. On a vérifié les déclarations de toutes les personnes échantillonnées des vagues 1 et 2 qui ont fourni un SSN ou ont dit ne pas avoir de SNN et, au besoin, ces déclarations ont été corrigées par la Social Security Administration (administration de la sécurité sociale). À la suite de cette opération, Sater (1986) calcule que les données sur le SSN contenues dans le fichier SIPP sont valides pour environ 95 pour cent des personnes de l'échantillon SIPP qui en ont effectivement un.

La profusion d'autres données, nom de famille, prénom, numéro de voirie, nom de rue, nom de l'immeuble d'appartements, ville, code postal, sexe et date de naissance, suffit à assurer un appariement de grande qualité, même en l'absence d'un code d'identification unique tel que le SSN. En outre, pour nous aider à évaluer l'incidence de toute autre erreur d'appariement, le responsable de l'appariement du Census Bureau produit une mesure ordinale de la valeur de l'appariement ou du non-appariement de chaque observation de l'enquête et de la donnée correspondante des dossiers administratifs.

4.2 Méthodes d'appariement automatisé du Census Bureau

L'étude de vérification des enregistrements est effectuée à l'aide de méthodes d'appariement statistique automatisé qui sont fondées sur l'ouvrage théorique de Fellegi et de Sunter (1969). Ces méthodes ont été élaborées au Census Bureau, principalement pour les besoins de l'estimation du sous-dénombrement relatif au recensement.

L'appariement statistique automatisé consiste à examiner deux fichiers informatiques et à trouver des paires d'enregistrements, soit un de chaque fichier, qui concordent (pas nécessairement de façon exacte) pour certaines combinaisons de variables. Le processus comporte des mesures discontinues multiples, mais fondamentalement, on en compte quatre: uniformiser les zones communes de données des deux fichiers que le programme d'appariement examinera afin de déterminer si on peut apparier ou non une paire d'enregistrements; trier les deux fichiers de façon à établir de petits sous-ensembles

d'enregistrements (ou "blocs" d'enregistrements) qui sont constitués d'un nombre adéquat de paires que le programme d'appariement doit vérifier; déterminer et quantifier l'utilité de chaque zone de données qui pourrait faire l'objet de l'appariement visant à relever les concordances parfaites; et appliquer les algorithmes informatiques qui permettent d'effectuer l'appariement proprement dit des enregistrements.

4.2.1 Uniformisation

Tous les fichiers de données de l'étude de vérification des enregistrements, c'est-à-dire les fichiers SIPP et ceux des dossiers administratifs, sont traités à l'aide d'un programme d'uniformisation des données sur l'adresse qui uniformise la structure des diverses composantes de l'adresse (par ex., le nom de la rue, le type de rue et son orientation, le nom de la ville, l'abréviation utilisée pour l'État, etc.) et analyse chaque composante d'une zone fixe de données. Plusieurs programmes ont été établis à cette fin. Nous utilisons actuellement le programme ZIPSTAN qui a été élaboré au Census Bureau, mais il est possible que nous adoptions bientôt un nouveau produit mis au point par notre Division de la géographie.

En plus des méthodes d'uniformisation qui s'appliquent à tous les fichiers de données, les zones individuelles de données de bon nombre de fichiers doivent être modifiées de façon que les fichiers présentent une même structure pour les besoins de l'appariement. Parmi les variables qui présentent un problème sur ce plan, nous retrouvons le sexe (qui peut être indiqué par un code alphabétique, "m" ou "f", ou numérique, "1" ou "2"); la date de naissance (qui peut être indiquée de différentes façons: "mm-jj-aa", "ss-aa-mm-jj" ou selon la date julienne); et le nom (il peut y avoir une seule zone ou des zones distinctes pour chaque composante). Actuellement, nous établissons des programmes sur demande pour faire ce genre d'uniformisation, mais il est possible que nous ayons bientôt recours à une nouvelle version du programme général d'uniformisation des données du Census Bureau (GENSTAN) pour effectuer cette tâche.

4.2.2 Groupage d'enregistrements

Le groupage d'enregistrements, qui consiste à créer des sous-ensembles d'enregistrements que le programme d'appariement examinera en vue d'assortir des paires d'enregistrements (par ex., Jaro, 1985), est une stratégie nécessaire lorsqu'il faut apparier des fichiers qui contiennent un grand nombre d'enregistrements. De toute évidence, il y aurait une probabilité maximale de faire des appariements parfaits si pour chaque enregistrement d'un fichier, on parcourait l'autre fichier en entier pour trouver l'enregistrement correspondant. Toutefois, il est impossible de procéder à de telles recherches dans des fichiers de cette taille. Grâce aux sous-ensembles d'enregistrements qui sont établis pour chaque fichier, il est possible de faire l'appariement; toutefois, certains enregistrements sont alors exclus des sous-ensembles, ce qui accroît la probabilité que certains rapprochements exacts ne seront pas faits. Par conséquent, il faut s'assurer que les éléments groupés présentent suffisamment de variation pour permettre la répartition des fichiers en de nombreux (et plus petits) blocs, et que ces éléments permettent de discerner de façon rapide s'il y a ou non appariement, c'est-à-dire qu'ils concordent presque toujours lorsqu'une paire d'enregistrements est bien assortie et ne concordent presque jamais lorsque l'inverse se produit. (Dans ce dernier cas, on suppose qu'un élément idéal pour le groupage ne contient d'erreur dans ni l'une ni l'autre des deux zones.)

Le premier critère, celui de la variation suffisante, est facile à respecter, mais le second l'est moins. La principale stratégie de groupage adoptée pour l'étude de vérification des enregistrements consiste à utiliser les trois premiers chiffres du code postal de cinq chiffres et un code SOUNDEX de quatre caractères tiré du nom de famille de la personne qui fait partie de l'échantillon ou du bénéficiaire. Le premier code est un

indicateur géographique d'une division régionale d'un État qui, selon les experts en appariement du Census Bureau, ne comporte habituellement pas d'erreur. Le deuxième code est un algorithme fréquemment utilisé pour créer un code de même longueur et de même structure à partir de chaînes de caractères en entrée de longueurs variables. Le code est composé de la première lettre de la chaîne (ici, le nom de famille), puis d'un code numérique utilisant seulement certaines lettres qui sont dans le reste de la chaîne. Les codes établis pour les besoins du groupage permettent de réduire le nombre d'erreurs de groupage attribuables aux erreurs d'orthographe, mais pas de les supprimer complètement.

Comme le succès de l'appariement dépend en grande partie de la stratégie de groupage, on utilisera dans l'étude au moins deux et peut-être trois stratégies de groupage distinctes -- et, dans chaque cas, il n'y aura aucun rapport entre les éléments groupés -- pour chaque paire de fichiers à apparier. De cette façon, il y a moins de risque qu'une paire d'enregistrements réellement assortis ne soit pas repérée à cause du groupage. Ces plans de groupage subséquents ne seront pas uniformes pour tous les appariements (en raison de la disponibilité variable de certaines zones de données ou à cause de problèmes connus d'ordre qualitatif), mais comprendront probablement une certaine combinaison de variables telles que le sexe, le mois de naissance, le jour de naissance, le code SOUNDEX pour le nom de la ville ou de la rue ou le SSN partiel.

4.2.3 Poids d'appariement des zones de données

Compte tenu de certaines variations, les zones de données utilisées pour effectuer l'appariement des fichiers SIPP et des données administratives comprendront le numéro de voirie, le nom de la rue, le numéro d'appartement, la ville, le code postal, le SSN, le sexe, la date de naissance, le nom de famille et le prénom. Nous savons intuitivement que ces zones ne sont pas équivalentes lorsqu'il faut déterminer si une paire donnée d'enregistrements est assortie ou non, la concordance de la variable "sexe" ne constituant pas un indice d'appariement réel aussi valable que la concordance de la variable "SSN", par exemple. Dans leur présentation d'une théorie générale de couplage des enregistrements, Fellegi et Sunter (1969) discutent du calcul des poids qui tiennent compte des différents pouvoirs de discrimination des diverses zones de données et de quelle façon les règles de décision optimale utilisent les poids. Le personnel de recherche en matière de couplage des enregistrements du Census Bureau a élaboré des programmes en se fondant sur la méthode d'établissement des systèmes non linéaires de Newton (voir Luenberger, 1984) pour résoudre les équations de Fellegi-Sunter, et ces programmes sont utilisés dans le cadre de l'étude de vérification des enregistrements SIPP pour calculer les derniers poids d'appariement.

4.2.4 Programme d'appariement automatisé

Le Census Bureau travaille actuellement à l'élaboration d'un programme d'appariement automatisé (CENMATCH) qui sera exploité sur les ordinateurs individuels de IBM, sur l'unité centrale 4361 de IBM et sur d'autre matériel, lequel applique les calculs de Fellegi-Sunter à un ensemble défini par l'utilisateur de zones de données tirées de fichiers triés (groupés) selon les exigences de l'utilisateur. L'utilisateur introduit les poids d'appariement initiaux pour chaque zone, détermine les critères de concordance pour la comparaison de chaque zone (c.-à-d. si les zones doivent être parfaitement identiques pour que le programme d'appariement détermine qu'il y a concordance ou si une comparabilité approximative suffit), relève les inscriptions manquantes et précise comment il faut procéder dans ces cas (les inclure ou ne pas en tenir compte dans le calcul d'un poids d'appariement composite), et détermine les valeurs limites des poids composites pour les paires appariées et non appariées. L'utilisateur produit les codes appropriés du programme en COBOL en vue d'effectuer un appariement conformément à

ces instructions à l'aide de GENLINK, le générateur de programmes de couplage des enregistrements du Census Bureau (LaPlant, 1987).

Autrement dit, le programme d'appariement: 1) parcourt chaque fichier de données afin de trouver des blocs d'enregistrements comparables, c'est-à-dire des enregistrements dont les éléments choisis du groupage concordent parfaitement; 2) compte le nombre d'enregistrements dans les blocs relevés pour s'assurer que la longueur des blocs de l'un ou l'autre des fichiers ne dépasse pas la longueur maximale prédéterminée; 3) calcule le poids composite relatif à toutes les paires d'enregistrements possibles dans le bloc; 4) fait correspondre à chaque enregistrement du plus petit bloc un enregistrement apparié du plus grand bloc selon une formule qui maximise le poids composite total pour toutes les paires du bloc; 5) applique la technique décisionnelle de Fellegi-Sunter pour déterminer si une paire d'enregistrements est assortie, non assortie ou nécessite un autre examen; et 6) établit un fichier de correspondance se rapportant aux enregistrements sautés (c.-à-d. les enregistrements d'un bloc d'un fichier qui ne concordent pas avec ceux du bloc correspondant de l'autre fichier) et aux enregistrements appariés (appariés/examen/non appariés) dans chaque fichier.

5. ANALYSE

L'objet de l'étude de la vérification des enregistrements est d'estimer certains paramètres des erreurs de mesure applicables aux échantillons de personnes, au contenu des données et aux mois d'enquête, et d'évaluer le rapport qui existe entre les erreurs mêmes et entre les erreurs et les variables qui reflètent les caractéristiques du plan de sondage. L'objectif général de l'étude est d'utiliser les données appariées en vue d'estimer pour chaque variable dichotomique relative à la participation:

- l'erreur systématique de réponse (à l'aide du résultat de la différence de valeur des données de l'enquête et des enregistrements);
- les variables explicatives de l'erreur systématique de réponse (à l'aide des techniques de régression logistique ou de régression des probits, ou peut-être des techniques LISREL fondées sur des matrices présentant des coefficients d'association à séries multiples et tétrachoriques (Joreskog et Sorbom, 1984);
- la variance des erreurs de réponse (par ex., établie à partir des résidus de la régression);
- les conditions ou les groupes qui présentent des variances des erreurs de réponse très importantes et très petites; et
- le genre et le degré de confusion relative aux programmes de transfert qui engendre des erreurs de réponse (à l'aide de méthodes d'analyse de la structure de covariance comme LISREL).

Nous prévoyons estimer les mêmes paramètres pour les déclarations des sommes d'argent reçues dans le cadre de chaque programme de transfert, mais nous n'avons pas encore choisi la méthode d'estimation de base.

Les questions concernant les erreurs de mesure sont réparties en deux catégories: les questions qui s'appliquent à tous les mois d'enquête et celles pour lesquelles on compare les erreurs d'un mois d'enquête à un autre. La première catégorie comprend les estimations du nombre d'erreurs de réponse imputées aux répondants visés et aux enquêtés-substituts et celles qui sont attribuées aux intervieweurs. La deuxième catégorie comprend les erreurs qui résultent des enquêtes par panel et que l'on connaît bien, par exemple, les erreurs de télescopage, le biais dû à la durée de la présence dans l'échantillon, les défaillances de mémoire, le biais attribuable au groupe de renouvellement, etc., c'est-à-dire celles qui sous-entendent que les erreurs de mesure

différeront d'un mois d'enquête à un autre, toute autre chose demeurant constante par ailleurs. Nous ajoutons à cette liste ce que Hill (1987) a qualifié de biais du "point de jonction" des enquêtes longitudinales et dont nous reparlerons ci-dessous.

Pour mieux comprendre les questions que nous désirons aborder au sujet des différentes périodes de déclaration, il faut se reporter à la figure 2 qui présente le calendrier des mois d'interview et des périodes de référence pour un groupe de renouvellement de répondants SIPP.

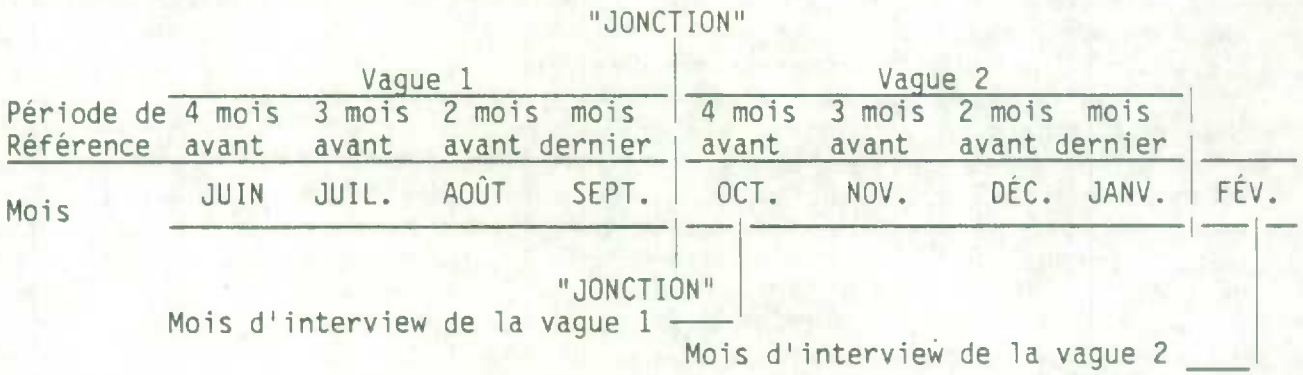


Figure 2: Mois d'enquête prévus pour le groupe de renouvellement 1 dans le cadre de la SIPP

La figure fait état de deux interviews. La première a lieu au début d'octobre et recueille des renseignements sur des faits survenus en septembre (le mois précédent), en août (deux mois avant), en juillet (trois mois avant) et en juin (quatre mois avant). De même, au cours de la deuxième interview qui est effectuée quatre mois plus tard, les données recueillies portent sur janvier, décembre, novembre et octobre. Nous nommons "jonction" la période de transition entre septembre et octobre parce que cette période se situe entre les périodes de référence visées par les deux interviews.

Pour connaître l'applicabilité de l'hypothèse de télescopage interne (selon laquelle les personnes n'oublient pas certains faits, mais croient que ces faits sont survenus à une date moins éloignée que la date réelle), nous déterminerons si l'erreur systématique de réponse relative aux premiers mois de la période de référence (juin et juillet pour la vague 1 et octobre et novembre pour la vague 2) est négative et si elle est positive pour les derniers mois (août et septembre, décembre et janvier), et si la somme de ces deux erreurs correspond à zéro.

Nous avons l'intention de vérifier l'hypothèse qui s'applique aux interviews portant sur une période limitée et selon laquelle les répondants déclarent les faits qui se sont passés il y a longtemps comme ayant eu lieu au cours d'une période de référence non limitée par une interview précédente (juin à septembre), mais qu'une telle chose ne se produit pas lorsque les périodes de référence sont délimitées par une interview antérieure (dans le cas présent, octobre à janvier).

Pour examiner l'hypothèse au sujet des défaillances de mémoire (voulant que la probabilité d'oublier un fait augmente avec le temps), nous déterminerons si l'erreur systématique de réponse a une valeur négative plus élevée au cours des premiers mois de chaque période de référence qu'au cours des derniers mois.

Les hypothèses relatives à la durée du temps passé dans l'échantillon et aux groupes de renouvellement supposent que les erreurs de réponse seront plus importantes à la deuxième interview qu'à la première, après avoir corrigé les effets résultant des

variations saisonnières. Nous examinerons cette hypothèse et si nous constatons qu'elle est juste, nous examinerons certaines des théories avancées par des spécialistes qui expliquent pourquoi un tel phénomène peut se produire. Est-ce que les éléments de réponse de la première et de la deuxième interview sont différents, comme le laissent entendre Stasny et Fienberg (1985), ou est-ce que la qualité des déclarations des répondants diminue à la deuxième interview, comme le prévoit l'hypothèse du conditionnement de Neter et Waksberg (1966)?

Nous n'avons pas encore déterminé dans quelle mesure ces problèmes classiques des enquêtes longitudinales se posent pour la SIPP. Toutefois, un des problèmes connu a trait à l'estimation des changements observés d'un mois à un autre dans la participation aux programmes (Burkhead et Coder, 1985). Plus précisément, un plus grand nombre de changements dans la participation aux programmes surviennent au "point de jonction" entre les interviews (entre septembre et octobre, figure 2) qu'entre les mois visés par une interview (par ex., entre juin et juillet, ou entre juillet et août, ou entre août et septembre). Le Census Bureau n'a pas encore publié les estimations mensuelles de la transition relative à la participation aux programmes parce que ces estimations révèlent un modèle sur lequel l'erreur de mesure semble avoir influé considérablement. Moore et Kasprzyk (1984) ainsi que Hill (1987) se sont interrogés sur les types d'erreurs de réponse, de non-réponse ou d'application qui pourraient être à l'origine de ce modèle et quel ensemble d'estimations de la transition est le plus approprié. En analysant le problème en fonction des données administratives, nous espérons pouvoir déterminer de façon plus sûre dans quelle mesure les erreurs de réponse et de non-réponse contribuent à créer le modèle observé.

Il est possible qu'il y ait un lien entre le biais attribuable au point de jonction et le phénomène mieux connu selon lequel la variance des erreurs de mesure a tendance à gonfler les estimations de changement brut ou à sous-estimer la stabilité. Selon des ouvrages récents (Fuller, 1986), il existe plusieurs solutions à ce problème. Nous prévoyons entreprendre une étude empirique des effets des erreurs de mesure sur les estimations de la transition afin de déterminer s'il est possible, par exemple, de corriger les erreurs de réponse en nous fondant sur les estimations établies à la suite des réinterviews.

Enfin, nous avons mentionné précédemment qu'il pourrait être difficile d'obtenir des estimations non faussées des erreurs si les dossiers contiennent aussi des erreurs. Nous voulons estimer, à l'aide des mesures de réinterview (qui identifient l'estimation de var e), la variance des erreurs dans les dossiers (var u). Toutefois, nous continuons de supposer que les dossiers sont sans biais.

6. PREMIÈRES CONCLUSIONS

Pour illustrer notre approche, examinons la question du "point de jonction" à l'aide de certaines données d'essai dont nous nous servons pour nous familiariser avec les méthodes de traitement des données. Rappelons que le problème du point de jonction résulte du fait que les rapports mensuels d'interview présentent un plus grand nombre de changements relatifs à la participation aux programmes entre les mois visés par des interviews distinctes qu'entre les autres mois (visés par la même interview).

Voici quelques questions au sujet des données d'enquête sur lesquelles les données des dossiers administratifs pourraient fournir des éclaircissements:

1. Y a-t-il trop de changements déclarés au point de jonction?
2. Y a-t-il trop peu de changements déclarés pour les autres mois?

3. Est-ce que les différentes sources déclarent le même nombre de changements pour tous les mois d'enquête, mais répartissent ces changements différemment?

Nous verrons ensuite ce que nous appelons les données de "comparaison globale" qui peuvent aider à répondre à ces questions. Il faut toutefois préciser que ces données s'appliquent à un échantillon de commodité et ne représentent pas nécessairement une population qui pourrait nous intéresser. Il faut aussi signaler qu'il s'agit d'un petit nombre de cas si l'on se fonde sur les normes relatives aux enquêtes gouvernementales. Pour ces raisons, nous nous en tiendrons aux paramètres descriptifs des observations.

Les comparaisons globales ne nécessitent pas l'appariement individuel des données de l'enquête et de celles des dossiers. Toutefois, dans l'exemple que nous donnons, nous utilisons le même échantillon de 1,536 personnes tant pour les valeurs de l'enquête que celles des dossiers. Cette façon de procéder permet d'éliminer les différences de définition des champs d'observation que présente souvent cette méthode.

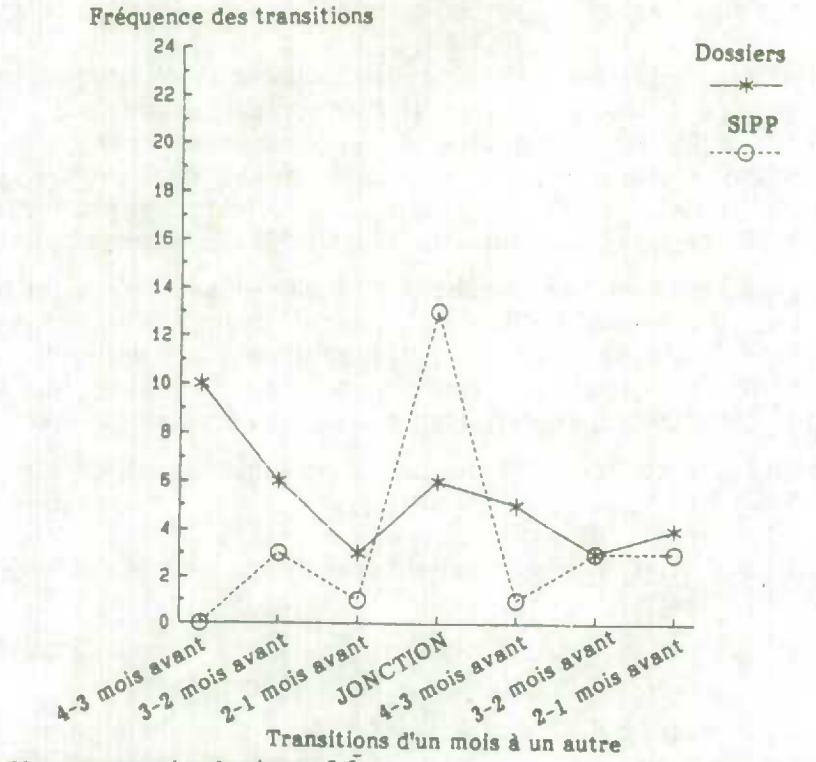
Si l'on suppose que les données des dossiers sont exactes, le graphique de l'AFDC (figure 3) sous-entend:

1. Qu'on déduit qu'un trop grand nombre de transitions sont imputées aux périodes de jonction de l'enquête.
2. Qu'on déduit qu'un trop petit nombre de transactions sont imputées aux autres mois.
3. Qu'un trop petit nombre de transitions sont déclarées globalement à l'enquête, ce qui représente un problème réel de déclaration par défaut ainsi qu'un problème de répartition en fonction du temps.

Lorsqu'on regarde le graphique des bons alimentaires (figure 4), nous constatons des tendances semblables, mais non identiques:

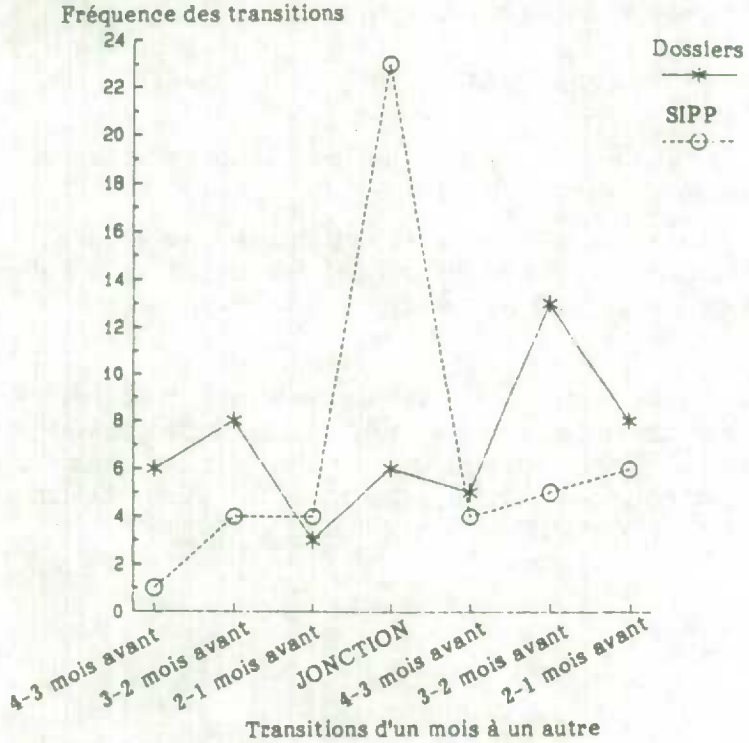
1. On déduit encore qu'un trop grand nombre de transitions se produisent au point de jonction.
2. Le biais attribuable à la déclaration par défaut établi pour les autres mois ne constitue pas un problème sérieux.
3. Les données de l'enquête et celles des dossiers présentent à peu près le même nombre de transitions totales, ce qui permet de croire qu'il s'agit d'un problème de répartition en fonction du temps, et non d'un biais partiel.

Il y a plusieurs autres tests à faire et de nombreuses autres hypothèses à explorer avant de pouvoir tirer des conclusions au sujet de la nature des erreurs de mesure et de leurs causes probables. Nous estimons que les données des dossiers administratifs nous permettront de mieux comprendre l'importance et la nature de ces erreurs d'enquête et nous aideront peut-être à déterminer leur cause.



Moyenne pour les dossiers = 5.3
Moyenne pour l'enquête = 3.4

Figure 3: Transitions relatives à l'AFDC, selon les données de la SIPP et selon les dossiers



Moyenne pour les dossiers = 7.0
Moyenne pour la SIPP = 6.7

Figure 4: Transitions relatives aux bons alimentaires, selon les données de la SIPP et selon les dossiers.

REMERCIEMENTS

Bon nombre de personnes ont contribué à la réalisation de l'étude de vérification des dossiers SIPP. Bien que nous ne puissions pas nommer ici toutes celles qui ont pris part au projet, nous désirons témoigner notre reconnaissance à Jeannette Robinson, qui a préparé la multitude de fichiers des dossiers administratifs en vue de l'appariement, à Bill LaPlant, que nous avons consulté en raison de ses vastes connaissances techniques concernant le programme d'appariement du Census Bureau et les programmes qui s'y rattachent, à Chris Dyke, qui s'est montré infatigable dans ses efforts pour assurer le bon déroulement du programme d'appariement dans le nouveau système informatique, et à Dan Kasprzyk, pour l'appui constant et patient qu'il a témoigné à l'égard de notre projet.

BIBLIOGRAPHIE

Bailar, B. (1968). "Recent Research in Reinterview Procedures", *Journal of the American Statistical Association*, vol. 63, 41-63.

Burkhead, D., et Coder, J. (1985). "Gross Changes in Income Recipiency from the Survey of Income and Program Participation", *Proceedings of the Social Statistics Section*, American Statistical Association, Washington, DC.

David, M. (1983). *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program* New York: Social Science Research Council.

Feather, J. (1972). *A Response Record Discrepancy Study*, University of Saskatchewan, Saskatoon.

Fellegi, I., et Sunter, A. (1969). "A Theory for Record Linkage", *Journal of the American Statistical Association*, vol. 64, 1183-1210.

Fuller, W., et Tin C.C. (1986). "Response Error Models for Changes in Multinomial Variables", *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 425-441.

Hill, D. (1987). "Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods", communication présentée aux réunions annuelles de l'American Statistical Association, San Francisco, CA, 13 août.

Jaro, M. (1985). "Current Record Linkage Research", communication présentée au Census Advisory Committee de l'American Statistical Association, U.S. Bureau of the Census, 25 avril 1985.

Jöreskog, K., et Sörbom, D. (1984). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likehood, Instrumental Variables, and Least Squares* Methods, Mooresville, Indiana: Scientific Software, Inc.

Koons, D. (1973). "Quality Control and Measurement of Nonsampling Error in the Health Interview Survey", *Vital and Health Statistics*, série 2, n° 54, U.S. Public Health Service, Washington, DC.

Laplant, W. (1987). "Maintenance Manual for the Generalized Record Linkage Program Generator (GENLINK) SRD Program Generator System", Statistical Research Division, document interne, Washington, DC: U.S. Bureau of the Census.

Luenberger, D. (1984). *Linear and Nonlinear Programming*, Reading, MA: Addison Wesley.

Marquis, K. (1986). "Discussion of 'Correlates of Reinterview Inconsistency in the Current Population Survey", *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 235-240.

Marquis, K. (1978). *Record Check Validity of Survey Responses: A Reassessment of Bias in Reports of Hospitalizations*, The Rand Corporation, Santa Monica, CA. R-2319-HEW.

Marquis, K., Marquis, S., et Polich, M. (1986). "Response Bias and Reliability in Sensitive Topic Surveys", *Journal of the American Statistical Association*, vol. 81, 381-389.

Moore, J., et Kasprzyk, D. (1984). "Month-to-Month Recipiency Turnover in the ISDP", *Proceedings of the Survey Research Methods Section, American Statistical Association*, 726-731.

Nelson, D., McMillen D., et Kasprzyk (1985). "An Overview of the Survey of Income and Program Participation, Update 1", *SIPP Working Paper Series*, n° 8401, Washington, DC: U.S. Bureau of the Census.

Neter, J., Maynes, S., et Ramanathan, R. (1965). "The Effect of Mismatching on the Measurement of Response Errors", *Journal of the American Statistical Association*, vol. 60, 1005-1027.

Neter, J., et Waksberg, J. (1966). "A Study of Response Errors in Expenditures Data from Household Interviews", *Journal of the American Statistical Association*, vol. 59, 18-55.

Sater, D. (1986). "SSN Response Rates and Results of SSN Validation/Improvement Operation", note de service du U.S. Bureau of the Census à l'intention de R. Herriot, 11 mars 1986.

Stasny, E., et Fienberg, S. (1985). "Some Stochastic Models for Estimating Gross Flows in the Presence of Nonrandom Nonresponse", *Proceedings of the Conference on Gross Flows in the Labor Force Statistics*, Department of Commerce and Department of Labor, Washington, DC, 25-39.

SESSION V: COMMUNICATIONS SOLLICITÉES

LES DOSSIERS ADMINISTRATIFS COMME AUTRE SOURCE DE DONNÉES

Président: F. Scheuren, U.S. Internal Revenue Service

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

LES DONNÉES ADMINISTRATIVES COMME SUBSTITUTION AUX DONNÉES DU RECENSEMENT

J. PODOLUK[1]

RÉSUMÉ

Il existe au Canada plusieurs bases de données administratives qui couvrent d'importants segments de la population. Ainsi, en raison de l'existence de programmes universels de sécurité sociales, il existe des bases de données qui recouvrent pratiquement l'intégralité de la population âgée de 65 ans et plus et les enfants âgés de moins de 18 ans. Ces ensembles de données viennent compléter les dossiers de l'impôt sur le revenu informatisés qui recouvrent plus des trois quarts de la population adulte. Le document examine le champ de ces dossiers, compare les concepts de la fiscalité et les concepts de l'enquête et décrit le genre de données démographiques que l'on peut élaborer à partir des dossiers fiscaux. Il propose un programme possible d'évaluation de la relation qui existe entre les améliorations possibles aux dossiers fiscaux afin d'en relever l'utilité statistique.

INTRODUCTION

Contrairement à d'autres pays industrialisés, le Canada ne possède pas de registre de population. Il y existe cependant depuis longtemps de gros fichiers de données administratives créés pour les besoins de programmes fiscaux et sociaux. Plus ça va et plus on trouve dans ces fichiers des renseignements se rapportant à la presque totalité de la population ou à certains segments de la population comme les personnes âgées ou les enfants. Mais avant l'arrivée des ordinateurs, ces fichiers étaient tenus à jour à la main, de sorte qu'il était difficile d'y avoir accès à des fins statistiques.

Lorsqu'on a commencé à informatiser la tenue des dossiers dans les années soixante, Statistique Canada a fait quelques premières tentatives d'exploitation statistique des données administratives, mais les résultats n'ont pas été concluants, pour diverses raisons. Certains fichiers n'étaient pas encore informatisés, le système de codes postaux était plutôt primitif et les ordinateurs n'avaient pas la capacité nécessaire pour traiter des millions d'enregistrements. Par ailleurs, même si Statistique Canada avait, en vertu de la Loi sur la statistique, reçu le mandat d'élaborer un système statistique national, son droit d'accès aux fichiers de données administratives comme les fichiers de l'impôt sur le revenu n'était pas bien défini. Des modifications apportées à la loi par la suite ont corrigé cette situation.

[1] Jenny R. Podoluk, expert-conseil, 626, avenue Gainsborough, Ottawa (Ontario) K1A 2Y8 Canada.

Les difficultés rencontrées au début ont été en grande partie surmontées et, compte tenu de la prolifération des données administratives et de la croissance du coût du recensement, le moment est venu de se demander dans quelle mesure les statistiques administratives peuvent compléter ou remplacer les données de recensement. Nous allons essayer de répondre à cette question et en indiquer quels aspects méritent, à notre avis, une étude plus poussée.

1. QUALITÉS DES FICHIERS DE DONNÉES ADMINISTRATIVES

Il n'est pas encore reconnu de façon absolue que les données administratives peuvent entièrement remplacer les données provenant d'autres sources. Le fait qu'on soit peu enclin à accepter que les fichiers de données administratives prennent la place des résultats d'enquête ou de recensement tient en partie à ce que ces fichiers sont créés dans le cadre de programmes fiscaux ou sociaux : ces programmes pouvant être modifiés ou éliminés, les fichiers de données, qui sont des sous-produits de l'appareil administratif, sont susceptibles de l'être également.

Il existe toutefois certains fichiers de données administratives qui sont, en fait, des fichiers presque exhaustifs de l'univers visé. Il convient de souligner que même si l'on a amélioré la qualité et la représentativité de certains fichiers de données depuis une vingtaine d'années, aucun nouveau programme d'importance exigeant la création de gros fichiers de données n'a été lancé pendant cette période, de sorte que les fichiers qu'on pourrait songer à exploiter sont essentiellement les fichiers de données qui existaient déjà dans les années soixante, à l'époque où les premiers efforts ont été faits pour incorporer les données administratives dans le système statistique.

2. DESCRIPTION DES FICHIERS DE DONNÉES

Parmi ces fichiers de données, les plus importants sont les suivants:

2.1 Fichiers fiscaux

Fait intéressant, les statistiques de l'impôt sur le revenu existent depuis plus de soixante ans et, jusqu'à la Deuxième Guerre mondiale, c'était le Bureau fédéral de la statistique qui était chargé de les produire. Pendant la guerre, cette responsabilité a été attribuée au ministère du Revenu national. Jusqu'à l'arrivée de l'ordinateur, les rapports statistiques du ministère du Revenu national reposaient sur certains échantillons de déclarations d'impôt dont on transcrivait des données très détaillées. Ces données ne renfermaient cependant aucune information personnelle permettant d'identifier les personnes auxquelles elles se rapportaient et étaient utilisées strictement à des fins statistiques. Les dossiers des déclarants étaient, pour leur part, tenus à la main.

Aujourd'hui, les fichiers de l'impôt sur le revenu des particuliers sont complètement informatisés et presque toute la population adulte y est représentée. Non seulement les fichiers contiennent des données fiscales courantes, mais on tient aussi à jour des fichiers longitudinaux portant sur plusieurs années, de sorte qu'on peut maintenant étudier les variations dans le temps de la situation d'un déclarant. Le ministère du Revenu national, aujourd'hui Revenu Canada, choisit encore un échantillon statistique désigné sous le nom d'échantillon de **Statistique fiscale** pour procéder à une analyse plus détaillée encore des déclarants, et c'est cet échantillon qui sert à la production du rapport statistique annuel. C'est ainsi que Revenu Canada a maintenant des séries de dossiers informatisés qui peuvent être utilisées pour produire des statistiques. Outre le traitement informatisé des dossiers, les autres facteurs importants qui ont contribué à augmenter l'utilité des données sont l'utilisation presque universelle du numéro d'assurance sociale (N.A.S.) après

l'introduction du Régime de pensions du Canada (RPC) dans les années soixante et l'adoption de l'actuel système de codes postaux.

Au Canada, il existe un certain nombre de programmes de sécurité sociale universels; les principaux sont le régime de sécurité de la vieillesse (SV), les allocations familiales, l'assurance-chômage (AC) et le Régime de pensions du Canada. Les allocations familiales et l'assurance-chômage sont des programmes qui remontent à l'avant-guerre, la SV, au début des années cinquante et le RPC, aux années soixante. Au départ, les allocations familiales et les prestations d'AC n'étaient pas imposables, de sorte qu'on ne disposait de données fiscales ni sur les bénéficiaires ni sur les cotisants. Aujourd'hui, elles le sont et on peut donc utiliser les fichiers fiscaux pour étudier la population qui les perçoit, ou du moins une bonne partie de cette population. Les prestations de la SV sont également imposables, mais les fichiers relatifs à leurs bénéficiaires ne sont pas aussi complets. Quant aux fichiers du RPC, étant donné qu'on se sert des dossiers fiscaux pour les tenir à jour, on y trouve des données sur presque toutes les personnes qui travaillent.

On peut donc dire qu'en général les bases de données administratives autres que les fichiers fiscaux peuvent uniquement compléter ces derniers et que, prises individuellement, elles sont pour la plupart d'une utilité statistique limitée. Nous allons à présent les décrire brièvement.

2.2 Fichiers des allocations familiales

Toutes les familles ayant des enfants de moins de 16 ans ou ayant des enfants de 16 et 17 ans qui fréquentent l'école reçoivent des allocations familiales. Par conséquent, de ces fichiers, il est possible de tirer des données sur le nombre de familles et sur le nombre d'enfants selon l'âge.

2.3 Fichiers de la sécurité de la vieillesse

Il y a dans ces fichiers des enregistrements concernant presque toutes les personnes âgées de 65 ans et plus ainsi que certaines veuves et femmes mariées âgées de 60 à 64 ans dont le revenu est faible. Les pensionnés à faible revenu peuvent avoir droit au supplément de revenu garanti, et les bénéficiaires de ce supplément doivent faire une déclaration d'impôt. Il se peut qu'à l'avenir ces fichiers soient de moins en moins représentatifs si les critères d'admissibilité au régime excluent les immigrants d'âge avancé.

2.4 Régime de pensions du Canada

Le Régime de pensions du Canada a été créé il y a une vingtaine d'années. Dès le début, on s'est servi du régime fiscal pour recueillir et enregistrer les cotisations de la population active au régime de pensions.

Les prestations du RPC étant imposables, les dossiers fiscaux, comme dans le cas des allocations familiales et des prestations de SV, sont une source de renseignements sur les caractéristiques des bénéficiaires. Bien qu'on tienne des fichiers distincts sur les cotisants pour l'administration du régime de pensions, la plupart des données de base sont extraites des dossiers fiscaux. Ces fichiers ont une lacune: du fait que le Québec administre son propre régime de pensions, on y trouve seulement des données sur la population active des neuf autres provinces. En outre, comme au delà d'un certain montant de gains annuels, les cotisations à verser au régime n'augmentent plus, il est difficile de savoir exactement si les données que contiennent les fichiers des cotisations se rapportent aux gains réels ou au montant plafond des gains en fonction duquel les cotisations sont déterminées.

2.5 Fichiers de l'assurance-chômage

Au début, les prestations d'assurance-chômage n'étaient pas imposables, mais comme le nombre de personnes ayant droit à ces prestations a considérablement augmenté depuis une quinzaine d'années environ, Revenu Canada s'est mis à jouer un rôle important dans l'administration de ce programme. Les prestations d'assurance-chômage sont devenues imposables et Revenu Canada a commencé à percevoir les cotisations. Par conséquent, les dossiers fiscaux sont encore une fois une source de renseignements, et sur les cotisants et sur les bénéficiaires.

2.6 Résumé

Aujourd'hui, les statistiques fiscales sont le pivot des bases de données administratives nationales. Les autres séries de données administratives peuvent être considérées comme un complément des données fiscales. Dans certains cas, par exemple pour les données du RPC sur les cotisations, les données sont dérivées des données fiscales.

Nous avons déjà fait remarquer que les bases de données administratives sont par définition instables du fait que les modifications apportées aux programmes ou aux lois peuvent influer sur la portée et la nature des données qui y sont introduites. Et, comme nous l'avons dit plus haut, la plupart des grandes séries de données actuelles existent depuis plusieurs dizaines d'années et elles ont pratiquement toutes évolué au cours de cette période.

Le gouvernement a toutefois indiqué que la réforme fiscale était une des grandes priorités des prochaines années et qu'il pourrait même procéder à une restructuration du régime d'imposition avant le prochain recensement décennal. Le ministre des Finances a dernièrement rendu publique une série de recommandations concernant les changements qu'il voudrait apporter à la structure de l'impôt sur le revenu. Il recommande d'éliminer les exemptions et de les remplacer par divers crédits d'impôt et par des décharges sur les taxes de vente dont bénéficieraient les non-contribuables. Il est peu probable que les nouvelles propositions entraînent une diminution du nombre de déclarants: en fait, si de nouvelles taxes de ventes sont introduites en même temps que de plus grandes décharges, la proportion de la population qui produit une déclaration d'impôt devrait plutôt continuer d'augmenter.

Dans les prochaines sections, nous étudierons plus en détail les caractéristiques particulières des fichiers de données administratives, la nature de leur utilisation actuelle, les évaluations qu'il faudrait faire pour 1986 des données de recensement par rapport aux données administratives, d'autres utilisations possibles des données administratives, les limites des données administratives et enfin les améliorations qu'il faudrait y apporter pour pouvoir les utiliser à la place des données du recensement de 1991.

3. LES DOSSIERS FISCAUX

3.1 Unité d'imposition

Contrairement au régime fiscal en vigueur dans certains pays, comme les États-Unis, le régime fiscal canadien n'impose pas les revenus combinés des conjoints dans le cas des couples mariés; chacun des conjoints doit plutôt remplir sa propre déclaration d'impôt soit pour déclarer la partie de ses revenus assujettis à l'impôt soit pour demander certains crédits comme le crédit d'impôt pour enfants. Le régime fiscal canadien repose donc principalement sur le revenu individuel.

Cependant, on recommande périodiquement l'adoption du système de déclaration commune pour les couples en raison des inégalités engendrées par la déclaration

individuelle et du fait que, de plus en plus, l'obtention de certains avantages nécessite la divulgation des deux revenus; mais rien n'a encore été fait dans ce sens. Étant donné la tendance actuelle dans la société voulant que les femmes soient considérées comme des citoyens à part entière, notamment en ce qui a trait aux pensions, il est possible que, pour des raisons politiques, le gouvernement ne puisse pas adopter un régime de déclaration d'impôt conjointe.

Il demeure que, dans les conditions actuelles, le régime fiscal et le régime de sécurité sociale en vigueur comportent toutes sortes d'anomalies pour ce qui est du traitement des conjoints. Par exemple, un couple marié dont le revenu combiné de $25,000 se compose d'un revenu de $15,000 et d'un revenu de $10,000 paie moins d'impôts qu'un couple marié ayant un revenu unique de $25,000 bien que le conjoint qui touche le revenu dans le deuxième cas ait droit à une exemption de personne mariée. Par ailleurs, les allocations familiales sont payables aux mères mais constituent des sources de revenu imposable et doivent être déclarées par le conjoint qui gagne le plus, habituellement le mari. De même, celui qui gagne le plus doit déduire les dépenses de garde d'enfants même si ces dépenses sont partagées par les deux conjoints. Par contre, pour demander un SRG, les conjoints doivent déclarer leurs revenus combinés. Dans les unions libres, les conjoints de fait ont droit à certains avantages comme les pensions, mais un de ces conjoints n'a pas le droit de compter son partenaire en union libre comme personne à charge dans sa déclaration d'impôt. Dans d'autres cas, le conjoint qui produit une déclaration d'impôt peut, si ça l'avantage, ajouter à son revenu certains revenus que son conjoint, qui ne produit pas de déclaration, a touchés; par exemple, des dividendes.

Par conséquent, Revenu Canada ne cherche pas à produire de statistiques sur les couples mariés (ni sur les familles) et, pour produire des données sur les familles, il faut avoir recours au couplage des enregistrements correspondant aux déclarations. Les quatre principales variables pouvant être utilisées pour le couplage des enregistrements sont le numéro d'assurance sociale du conjoint, l'état matrimonial, le nom et l'adresse. Tous ces renseignements nécessaires au couplage peuvent être obtenus pour la plupart des conjoints vivant à la même adresse, mais même dans ces cas-là les données peuvent être imparfaites si les deux conjoints n'ont pas le même nom. Cela arrive si l'épouse a choisi de garder son nom de jeune fille pour des raisons d'ordre professionnel ou si le couple vit en union libre[2]. Lorsque les conjoints n'habitent pas le même domicile parce qu'ils travaillent dans des villes différentes ou pour d'autres raisons, le couplage des enregistrements peut devenir plus difficile.

---

[2] Au Québec, une loi a été adoptée il y a plusieurs années en vertu de laquelle les femmes qui se marient gardent leur nom de jeune fille. Dans certains services de l'administration publique de cette province, il semble qu'on remplace le nom de femme mariée par le nom de jeune fille, par exemple sur les cartes d'assurance-maladie, même quand le mariage remonte à longtemps. Dans les déclarations d'impôt envoyées à Revenu Canada, certaines femmes mariées indiquent encore leur nom de femme mariée, tandis que d'autres inscrivent leur nom de jeune fille.

Comme, dans un couple marié, il n'est pas obligatoire de déclarer le nom complet de son conjoint mais seulement son prénom, il devient de plus en plus difficile de procéder au couplage des déclarations des conjoints mariés depuis peu. (En ce qui concerne les couples qui produisent des déclarations depuis cinq ou six ans, les plus anciennes contiennent l'information requise.) La situation se complique davantage lorsqu'il y a des enfants, parce que les enfants peuvent porter soit le nom d'un de leurs parents (les filles portant le nom de leur mère et les garçons celui de leur père), soit une combinaison des noms de leurs parents. Apparemment, les parents devront s'entendre à l'avenir sur le nom de famille que porteront leurs enfants. Cela donne à penser qu'il faudrait demander à Revenu Canada d'exiger des personnes mariées qu'elles déclarent le nom complet de leur conjoint et pas seulement son prénom.

Pour les données sur le revenu recueillies au moyen d'enquêtes et de recensements par Statistique Canada, l'unité d'observation est également l'individu, mais comme ces enquêtes et ces recensements recueillent invariablement des données sur chaque membre d'un ménage et sur les liens entre tous les membres d'un ménage, ces données peuvent facilement être agrégées selon diverses combinaisons comme le revenu des couples mariés, le revenu de la famille de recensement, le revenu de la famille économique et le revenu du ménage familial. Toutes ces unités sont définies en fonction des occupants d'un même logement. La famille de recensement se compose d'un mari et d'une épouse ou d'une mère ou d'un père et d'enfants non mariés, tandis que la famille économique comprend toutes les personnes apparentées vivant sous le même toit. Un ménage est défini comme comprenant tous les occupants d'un même logement, lequel est une unité de résidence physiquement déterminée.

À part les données sur les particuliers, les données tirées des dossiers fiscaux qu'on peut le mieux regrouper sont celles qui permettent de constituer le revenu des familles époux-épouse. On peut également essayer de produire des données agrégées sur le revenu des familles de recensement, mais, pour des raisons dont nous reparlerons un peu plus loin, les résultats ne seraient pas aussi bons. Quant aux données agrégées sur le revenu des familles économiques, la possibilité de les produire est encore plus lointaine.

Les données sur le revenu des ménages sont en demande, mais les utilisateurs qui les préfèrent à celles sur le revenu des familles évaluent sans doute mal les différences entre ces deux types de revenu. Le concept de revenu du ménage doit être flou dans leur esprit, ou alors ils en perçoivent mal les limites.

La seule base sur laquelle on pourrait se fonder pour produire des données agrégées sur le revenu des ménages à partir des dossiers fiscaux serait l'adresse du déclarant. Cependant, il n'est pas garanti que les résultats cadreraient avec le concept de revenu du ménage utilisé par Statistique Canada. Par exemple, il peut y avoir deux appartements à la même adresse sans que cela paraisse dans les déclarations d'impôt, de sorte que deux ménages au sens de la définition de Statistique Canada seraient considérés comme un seul selon les déclarations d'impôt; le vrai revenu des ménages s'en trouverait ainsi exagéré. Par ailleurs, il se peut que certains membres d'un ménage utilisent l'adresse d'un escompteur dans leur déclaration d'impôt au lieu d'indiquer leur propre adresse; leur revenu ne pourra alors pas être associé au lieu de résidence. Il en résulte une sous-évaluation du revenu des ménages.

Pour la plupart des études qu'on veut faire, à l'exception peut-être de l'analyse des caractéristiques du logement, le revenu des familles est sans doute un concept plus utile que le revenu des ménages. En général, le revenu moyen des ménages d'une région donnée est supérieur au revenu moyen des familles, mais l'inverse peut être vrai, dans une région géographique qui comprend une forte proportion de ménages d'une seule personne, par exemple. Les ménages d'une seule personne ont habituellement un revenu inférieur à celui des ménages familiaux; par conséquent, si l'on veut comparer le revenu des ménages au revenu des familles, il faut au moins distinguer les ménages d'une seule personne des ménages de deux personnes ou plus.

Le concept canadien de ménage est une variante du concept utilisé aux États-Unis et il est propre aux statistiques produites dans ces deux pays. Ailleurs, ce concept n'est pas lié à l'occupation d'une unité de logement donnée; le ménage est plutôt défini comme un groupe de personnes occupant la même unité de logement et mettant en commun des revenus et des dépenses. Cette notion correspond essentiellement à la définition utilisée dans l'enquête sur les dépenses des familles et aussi sans doute à celle que les utilisateurs préfèreraient à la définition canadienne actuelle.

En résumé, les nouvelles séries de données devraient, si possible, porter, en premier lieu, sur le revenu des conjoints et, en second lieu, sur le revenu des familles de recensement. Le revenu des ménages semble avoir une utilité limitée, car il est difficile

de le calculer selon le concept de Statistique Canada. Comme le revenu des familles est sans doute l'indicateur le plus significatif du niveau de vie, il ne vaut probablement pas la peine à l'heure actuelle de prendre des ressources destinées à autre chose pour produire des données sur le revenu des ménages.

3.4 Degré de représentativité des données fiscales

Le tableau 1 présente des données sur le pourcentage de la population qui a produit une déclaration d'impôt en 1983, selon le sexe et l'âge[3]. Les statistiques fiscales portent sur une plus grande proportion d'hommes que de femmes, quoique les femmes qui ne produisent pas de déclaration d'impôt soient souvent des femmes mariées ayant peu ou pas de revenus. Les groupes les moins représentés sont ceux des moins de 20 ans. On suppose que les déclarants de moins de 20 ans se situent dans le groupe des 15 à 19 ans et qu'ils sont pour la plupart des étudiants dont le revenu est assez faible pour les exempter de remplir une déclaration d'impôt. Les personnes âgées aussi produisent peu de déclarations d'impôt, en particulier les femmes de 60 ans et plus et les hommes de 70 ans et plus. Dans tous les groupes d'âge, le taux d'activité des femmes est plus faible et les femmes âgées sont plus susceptibles de dépendre des pensions du gouvernement pour la plus grande partie ou la totalité de leur revenu.

Il est intéressant de noter que la proportion de femmes de 20 à 29 ans qui produisent une déclaration d'impôt est un peu plus élevée que celle des hommes de la même tranche d'âge, tandis qu'on observe le contraire chez les 30 ans et plus. La production d'une déclaration par les mères qui désirent obtenir un crédit d'impôt pour enfants explique probablement cette variation. Malgré cela, 91% des hommes et 84% des femmes âgés de 20 à 64 ans sont directement représentés dans les fichiers fiscaux. Ces proportions sont de 83 et 74% respectivement pour l'ensemble de la population de 15 ans et plus. Ce sont les plus jeunes qui, en étant proportionnellement les plus nombreux à ne pas produire de déclaration d'impôt, font le plus baisser le degré de représentativité des données fiscales.

Le fait que les femmes soient moins représentées a probablement deux explications: a) le taux d'activité des femmes est moins élevé, même s'il ne cesse d'augmenter et que celui des hommes des tranches d'âge supérieures diminue et b) les femmes âgées dépendent plus des pensions de l'État, lesquelles constituent des sources de revenu entièrement ou partiellement non imposables.

Les données sur l'activité annuelle établies d'après les résultats de l'enquête mensuelle sur la population active donnent également une indication des lacunes des fichiers fiscaux. L'enquête sur l'activité annuelle permet d'estimer le nombre de personnes actives en 1983. Le tableau 2 compare les statistiques tirées des deux sources. En exprimant le nombre de déclarations d'impôt produites par sexe et par grand groupe d'âge en pourcentage du nombre de personnes actives en 1983, on obtient les résultats suivants.

---

[3] Les comparaisons figurant dans les sections qui suivent sont faites à partir des données présentées dans **Statistique fiscale.** Étant donné la méthode de tirage de l'échantillon, il est possible que les mêmes comparaisons faites avec des données portant sur l'ensemble de la population produisent des résultats légèrement différents. Les totaux pour 1983 n'ont pu être obtenus à partir des fichiers principaux, mais il serait utile de comparer les taux de représentativité en utilisant les données portant sur l'ensemble de la population. Il convient également de signaler qu'un grand nombre de personnes qui ne produisent pas de déclaration d'impôt sont des personnes à charge de déclarants. On estime donc que les déclarations d'impôt rendent compte directement ou indirectement de 90% de la population.

Table 1

Pourcentage de la population produisant une déclaration d'impôt, selon l'âge et le sexe, 1983*

Âge	Hommes (pourcentage)	Femmes	Total
Moins de 20 ans	35.3	31.6	33.5
20-24	86.7	89.1	87.9
25-29	93.0	93.7	93.3
30-34	92.9	92.6	92.7
35-39	92.3	90.0	91.6
40-44	91.3	88.1	89.7
45-59	92.9	81.6	87.3
50-54	92.1	73.5	82.8
55-59	93.3	63.0	77.8
60-64	89.8	56.4	72.1
65-69	86.2	53.0	68.2
70 ans et plus	69.0	45.5	55.1
Total	83.4	73.7	78.5
Total 20 ans et plus	89.3	78.0	83.4
Total 20-64 ans	91.3	83.8	87.5

* Le nombre de déclarants exprimé en pourcentage de la population selon l'âge en décembre 1983 a été estimé en interpolant les estimations de la population de 1983-1984 selon l'âge. Les données sont produites à partir des statistiques fiscales de 1985 qui sont fondées sur un échantillon représentant 2.9% de la population totale. Il est possible que les fichiers portant sur l'ensemble de la population soient plus complets. Il reste à savoir si les données incluent les déclarations remplies au nom de personnes décédées en 1983 ou au début de 1984. Certains faits tendent à prouver que oui. Dans ce cas, les personnes des tranches d'âge supérieures pourraient être surreprésentées.

Tableau 2

Pourcentage de la population active ayant produit une déclaration d'impôt, selon le sexe, 1983

Âge	Hommes	Femmes
15-24	64.7	63.8
25-44	94.8	92.5
45 et plus	89.8	63.9
Total	85.9	75.1

Comme nous l'avons vu un peu plus haut, certaines prestations de sécurité sociale sont désormais assujetties à l'impôt. On peut comparer les données fiscales aux données administratives pour calculer la proportion de bénéficiaires qui produisent une déclaration d'impôt et la proportion de ces prestations que les bénéficiaires indiquent dans leur déclaration d'impôt. Les trois types de prestations sociales assujetties à l'impôt sont les allocations familiales (y compris celles versées par le Québec), les prestations du régime de sécurité de la vieillesse (mais non le supplément de revenu garanti), les prestations d'assurance-chômage et les rentes versées par le Régime de pensions du Canada et le Régime des rentes du Québec.

En décembre 1983, environ 3,634,811 familles recevaient des allocations familiales. Les montants payés cette année-là, incluant les montants versés par le Québec, ont

totalisé $2,487 millions. Le nombre de déclarants ayant indiqué avoir reçu des allocations familiales dans leur déclaration d'impôt pour 1983, soit 3,722,799, était supérieur au chiffre mentionné plus haut; l'écart est probablement attribuable au fait que certains déclarants ayant touché des allocations familiales en 1983 n'en recevaient plus en décembre. Du reste, à cause de la baisse du taux de natalité, le nombre de familles qui reçoivent des allocations familales diminue au fil des ans. Par ailleurs, le montant des prestations reçues par les personnes prises en compte dans les données fiscales représentait 90% du montant total des prestations versées. Cela signifie que 10% des prestations n'étaient pas indiquées dans les déclarations d'impôt. Il y a de quoi être surpris quand on considère que les bénéficiaires d'allocations familiales qui ont un faible revenu ont tout intérêt à produire des déclarations d'impôt pour obtenir les crédits d'impôt pour enfants[4].

Comme les personnes âgées sont sous-représentées dans les fichiers fiscaux, ces derniers ne rendent probablement pas compte de l'ensemble des prestations du régime de sécurité de la vieillesse versées. En 1983, la proportion de personnes qui ont déclaré des prestations de SV au fisc par rapport au nombre de personnes qui recevaient de telles prestations en décembre de la même année était de 59%, et les montants déclarés représentaient 58% du montant total des prestations versées[5].

Les données concernant le nombre approximatif de bénéficiaires de l'assurance-chômage, du RPC et du RRQ en 1983 comportent toutes un risque de double compte. On peut toutefois calculer la proportion de ces prestations qui a été déclarée au fisc. En ce qui concerne les prestations d'assurance-chômage, 94.3% du montant total des prestations versées ont été déclarées, ce qui confirme l'impression qu'une faible proportion seulement de la population active ne produit pas de déclaration d'impôt. Quant aux prestations du RPC et du RRQ, 81.3% de leur montant total a été déclaré. On peut donc supposer qu'à la longue, à mesure que l'augmentation du nombre de bénéficiaires du RPC et du RRQ fera diminuer le nombre de personnes âgées qui dépendent pour leur revenu de la SV ou du SRG, les données fiscales se rapportant aux personnes âgées seront plus représentatives.

3.5 Résumé

Le champ d'observation des dossiers fiscaux comprend une très forte proportion de la population des 20 à 64 ans qui perçoivent un revenu. Il est moins complet pour ce qui est des personnes âgées, quoiqu'il devrait s'améliorer à la longue; quant aux jeunes, il y sont le plus sous-représentés. Il se peut que certains d'entre eux n'aient pas de revenu.

Cependant, compte tenu du fort taux d'activité des jeunes d'après les résultats de l'enquête sur la population active, il est plutôt probable qu'ils reçoivent un revenu, mais que celui-ci ne soit pas assez élevé pour être imposable. Il serait intéressant de se pencher sur les données du recensement de 1986 pour déterminer les caractéristiques des jeunes non-déclarants. Par ailleurs, pour combler certaines lacunes des données fiscales,

---

[4] Si l'on compare les montants d'allocation familiale versés aux montants déclarés selon les fichiers portant sur l'ensemble de la population, il semble que la proportion des montants déclarés soit plutôt de 98%. Il faudrait vérifier ces calculs pour voir si les paiements du Québec sont inclus dans les comparaisons. Les montants versés par le gouvernement fédéral représentaient à eux seuls 97% des chiffres cités dans **Statistique fiscale** pour 1983, mais c'est la somme des paiements du fédéral et du Québec qu'il faut comparer aux chiffres de Revenu Canada.

[5] Environ $100 millions en prestations de SV et $34 millions en prestations du RPC et du RRQ auraient été versés à des personnes vivant à l'étranger. Les comparaisons ci-dessus tiennent compte des paiements effectués à des non-résidents.

en l'absence d'un recensement, celles-ci pourraient être combinées aux données des allocations familiales et de la SV. Cette possibilité mérite d'être étudiée.

4. LIENS CONCEPTUELS AVEC LES DONNÉES DE STATISTIQUE CANADA

Les données de base communes aux dossiers fiscaux, aux recensements aux enquêtes sont celles sur l'âge, le sexe et l'état matrimonial[6]. Il y a aussi les données sur le revenu, qui sont à certains égards différentes du point de vue conceptuel, et les données sur le lieu de résidence, qui sont moins complètes et moins précises dans les fichiers de l'impôt. Les statistiques fiscales donnent très peu de renseignements sur les catégories professionnelles et n'en donnent aucun sur un certain nombre de caractéristiques comme le statut en regard de l'immigration, le lieu de naissance, l'origine ethnique ou le niveau de scolarité. Nous reparlerons davantage de ces aspects un peu plus loin.

4.1 Concept de revenu

Le recensement et l'enquête sur les finances des consommateurs mesurent le revenu conformément au concept de revenu monétaire du secteur du revenu personnel des Comptes nationaux, à quelques exceptions près. Ce concept exclut notamment les héritages, les gains en capital, les recettes fortuites comme les prestations d'assurance et les gains à la loterie et enfin les gains ou pertes de jeu.

Certains revenus mesurés dans les séries de Statistique Canada ne sont pas imposables, tandis que les revenus provenant de certaines sources sont mesurés artificiellement à des fins fiscales. Par ailleurs, Revenu Canada prélève un impôt sur certaines rentrées de fonds qui ne sont pas considérées comme des revenus par Statistique Canada. Voici une liste des éléments touchés.

a) Revenu non imposable - Certains paiements de transfert ne sont pas assujettis à l'impôt et ne sont donc pas déclarés au fisc; les principaux sont le supplément de revenu garanti (payable aux personnes âgées des tranches de revenu inférieures), les suppléments de revenu provinciaux, les allocations d'ancien combattant, les pensions d'ancien combattant, et les prestations d'assistance sociale et de bien-être social.

b) Revenu artificiel - Les dividendes font l'objet d'un traitement fiscal spécial. Jusqu'à ces derniers temps, le revenu total ne comprenait pas le montant réel des dividendes, mais ce montant augmenté de 50%. C'est le résultat de cette modification qui figure comme composante du revenu dans le fichier fiscal, le revenu total étant ainsi gonflé artificiellement d'un montant égal à la moitié des dividendes reçus. Le facteur de majoration a maintenant été réduit à un tiers, de sorte qu'il peut y avoir une diminution apparente, mais non réelle, des revenus sous forme de dividendes pour 1986. La diminution du facteur de majoration signifie que la surestimation des revenus sous forme de dividendes ne sera plus que d'un tiers. Les statistiques fiscales pourraient être corrigées pour permettre de calculer le montant réel du revenu sous forme de dividendes. (En vertu de la réforme fiscale proposée, le facteur de majoration pourrait être réduit encore davantage.)

c) Gains imposables non mesurés dans les statistiques du recensement et des enquêtes de Statistique Canada - Revenu Canada impose certains revenus en nature et certains avantages sociaux reçus en supplément des salaires et traitements, par exemple les

---

[6] Les enquêtes permanentes par sondage, comme l'enquête sur la population active, et le recensement utilisent des concepts uniformisés, de sorte que toute comparaison des concepts de l'impôt sur le revenu et des concepts de Statistique Canada s'applique aussi bien au recensement qu'aux enquêtes.

primes d'assurance-maladie versées par l'employeur, les avantages liés à l'utilisation d'une voiture de fonction et l'hébergement gratuit. Ces postes sont considérés comme faisant partie des salaires et traitements, lesquels, selon la définition servant au calcul de l'impôt, incluent non seulement les salaires et traitements en espèces, mais aussi une partie de ce qu'on définit comme des revenus supplémentaires du travail dans les Comptes nationaux. Il est impossible de ne pas tenir compte de ces revenus du travail dans le calcul du revenu.

Les revenus d'intérêt touchés sur les obligations constituent un deuxième exemple de cas où le traitement du revenu diffère de celui de Statistique Canada. Dans le passé, Statistique Canada demandait aux répondants de déclarer les intérêts en espèces effectivement reçus[7]. Pour sa part, Revenu Canada donnait aux déclarants le choix de déclarer soit les intérêts encaissés soit les intérêts à recevoir, c'est-à-dire non encore encaissés. Un fort pourcentage de déclarants optaient pour la méthode de caisse, même s'ils détenaient des obligations dont les intérêts ne seraient pas encaissés avant plusieurs années. La politique fiscale a été modifiée l'an dernier. Elle exige dorénavant la déclaration des intérêts composés au moins une fois tous les trois ans même si les intérêts n'ont pas été convertis en espèces. À la longue, cela aura pour effet de rendre les certificats de dépôt et les obligations à intérêts composés moins intéressants, mais il sera désormais difficile de savoir si les intérêts déclarés au fisc représentent les intérêts reçus ou les intérêts courus mais non encore reçus[8].

Les gains en capital sont considérés comme des revenus imposables, mais l'impôt n'est prélevé que sur la moitié de ces gains. Comme tout contribuable peut actuellement déduire de son revenu imposable jusqu'à concurrence de $500,000 de gains en capital pour la durée de sa vie (ce montant va être réduit à $100,000) les montants déclarés dans cette catégorie de revenu vont diminuer à l'avenir. Il est possible de ne pas tenir compte de ces gains pour effectuer des comparaisons avec les données de Statistique Canada, qui ne les incluent pas dans le revenu.

Au nombre des autres revenus qui sont imposables, mais qui peuvent ne pas être considérés comme des revenus dans les enquêtes sur les ménages, il y a les bourses d'études, les bourses d'entretien, les bourses de perfectionnement, les subventions à l'économie d'énergie ainsi que les revenus provenant d'activités illicites comme le jeu et la prostitution. Par ailleurs, le traitement des régimes enregistrés d'épargne-retraite (REÉR) peut entraîner le double compte de certains revenus pour une année donnée. Par exemple, si un contribuable cotise à un REÉR au cours d'une certaine année et en retire de l'argent la même année, Revenu Canada considère le montant retiré comme un revenu même si ce montant correspond à la cotisation versée cette année-là, laquelle a déjà été comptée dans le revenu brut déclaré.

En résumé, le concept de revenu utilisé pour les statistiques fiscales peut être modifié de manière à améliorer la comparabilité des données fiscales avec celles de Statistique Canada, mais il est impossible de rendre les deux séries de données parfaitement compatibles.

4.2 Classification géographique

La Division des données régionales et administratives a maintenant une telle connaissance des problèmes que pose l'attribution d'un code géographique dans les dossiers fiscaux qu'il n'est pas nécessaire de s'y attarder ici. Il semblerait que tous les dossiers, à

---

[7] Cette pratique n'était pas conforme aux concepts des Comptes nationaux, qui mesurent les revenus d'intérêt courus ou gagnés tous les ans.

[8] On suppose que les responsables des prochains recensements et de l'enquête sur les finances des consommateurs (EFC) devront décider s'ils adoptent ou non la façon de procéder de Revenu Canada.

quelques exceptions près, contiennent à présent le code postal, mais, pour toutes sortes de raisons, il est encore difficile d'apparier les données de Revenu Canada sur les déclarants aux données tirées du recensement. Certaines déclarations sont remplies par des escompteurs (comme H.R. Block), des avocats, des comptables ou des fiduciaires au nom du contribuable, de sorte que l'adresse qui figure sur la déclaration est celle de l'agent plutôt que celle de l'intéressé. Revenu Canada est apparemment en train de prendre les mesures nécessaires pour obtenir l'adresse personnelle des contribuables qui font faire leur déclaration d'impôt par un escompteur.

Revenu Canada attribue un code selon la localité, et c'est un système auquel on ne peut semble-t-il pas se fier, car ce ministère s'en tient à ce que le déclarant inscrit sans vérifier si c'est exact. Si, par exemple, un déclarant donne Ottawa comme lieu de résidence plutôt que Nepean, ou Toronto plutôt que Scarborough, le code qu'on lui attribue est celui d'Ottawa dans le premier cas ou de Toronto, dans le second. Par ailleurs, Revenu Canada n'essaie plus d'attribuer de code selon la région métropolitaine telle que définie par Statistique Canada. Il faut donc avoir recours à un fichier de conversion des codes postaux pour corriger les mauvais codes ou faire en sorte qu'un code corresponde aux secteurs géographiques délimités par le Bureau.

La conversion est possible si le secteur postal est bien défini, mais pas s'il faut établir la correspondance avec un secteur géographique peu peuplé, comme lorsqu'un déclarant donne pour adresse une route rurale qui traverse plusieurs municipalités[9]. On peut donc établir l'agglomération urbaine à partir du code postal, mais il est difficile d'en faire autant avec les régions rurales.

L'incidence que risque d'avoir la distribution du courrier dans des boîtes aux lettres communautaires plutôt qu'à chaque adresse dans les nouveaux quartiers constitue une nouvelle préoccupation. Il va falloir bien vérifier la précision des adresses pour ce genre de courrier.

4.3 Composition des familles

Des travaux de nature expérimentale visant à produire des données sur le revenu des familles à partir des fichiers fiscaux sont en cours, et les principaux problèmes que cela présente ont déjà été décrits. En bref, il est possible d'apparier presque toutes les déclarations de couples mariés, mais il est encore difficile de déterminer d'après la déclaration d'impôt si une personne vit en union libre avec quelqu'un ou bien s'il s'agit d'un parent unique ou d'une personne seule.

On demande aux couples mariés qui produisent une déclaration d'impôt d'indiquer le numéro d'assurance sociale de leur conjoint. La plupart le font, de sorte que les renseignements susceptibles de permettre l'appariement des données sont le nom (en général), le numéro d'assurance sociale (en général), l'état matrimonial et l'adresse (en général). Il arrive que le nom de famille des conjoints ne soit pas le même parce que la femme a préféré garder son nom de jeune fille ou le nom qu'elle avait acquis lors d'un mariage antérieur[10]. L'adresse n'est pas nécessairement un bon indicateur, pour plusieurs raisons; ainsi, certains époux ne demeurent pas au même endroit, ou alors ils font faire leur déclaration par quelqu'un d'autre, ou encore ils utilisent l'adresse de leur travail. Selon les résultats des travaux accomplis jusqu'à présent, le numéro d'assurance sociale permettrait d'apparier les déclarations des conjoints dans presque tous les cas. Il convient cependant de signaler que cette méthode risque d'avoir pour conséquence l'appariement de

[9] Par exemple, l'adresse postale du village d'Appleton, situé dans le comté de Lanark, est R.R. 3, Almonte, alors que ce village n'est qu'à 3 milles de Carleton Place et à 5 ou 6 milles d'Almonte.

[10] Il a déjà été question du problème qui existe au Québec depuis peu.

déclarations se rapportant à des conjoints qui ne partagent pas le même logement en permanence et ne forment donc pas une unité familiale d'après la définition de Statistique Canada. Il serait peut-être utile, lorsqu'on apparie les déclarations de couples mariés, de déterminer s'ils demeurent à la même adresse ou à des adresses différentes dans une même localité ou dans des localités différentes.

La famille de recensement est définie comme étant composée d'un époux et d'une épouse ou d'un parent seul, avec ou sans enfants jamais mariés, vivant dans le même logement. Il devient difficile de savoir qu'un couple a des enfants lorsque ces derniers font une déclaration d'impôt. Si le nom de famille et l'adresse d'une personne sont les mêmes que ceux d'un couple donné et si l'état matrimonial indiqué est "célibataire", on peut supposer qu'il s'agit de l'enfant du couple, quoiqu'il faille comparer son âge avec celui des parents présumés afin de ne pas prendre des frères, soeurs, neveux ou nièces pour des enfants. Parmi les autres problèmes qu'on rencontre, il y a celui des enfants issus de mariages différents qui ne portent pas le même nom que l'un des parents ou les deux, et celui des ménages d'immigrants où les noms de famille ne sont pas établis selon les mêmes règles que celles en usage au Canada.

Les données qui figurent sur la déclaration des parents devraient permettre de connaître le nombre d'enfants que ces derniers ont et donc d'établir la taille des familles. Là encore, un problème risque de se présenter si un enfant fait une déclaration d'impôt tout en étant compté comme personne à charge dans la déclaration d'un de ses parents du fait que son revenu n'est pas élevé. Il faut donc vérifier la situation chaque fois qu'un enfant produit une déclaration d'impôt. En 1983, un enfant pouvait être considéré comme une personne à charge si son revenu était inférieur à $3,870[11].

Le principal problème que pose la reconstitution des familles à partir des dossiers fiscaux survient lorsqu'on essaie de distinguer les parents uniques, les conjoints de fait et les personnes seules. Aux fins de l'impôt, un déclarant ne peut pas demander l'exemption de personne mariée s'il vit en union libre. Par conséquent, deux conjoints de fait qui n'ont pas d'enfants doivent faire chacun une déclaration d'impôt comme s'ils vivaient seuls. Par contre, un parent qui subvient aux besoins de ses enfants peut se servir de l'un d'eux pour demander l'équivalent de l'exemption de personne mariée. Il est donc possible de reconnaître le dossier d'un parent qui a des enfants à charge, encore que la famille ainsi reconstituée risque de ne pas répondre à la définition de Statistique Canada, car les enfants n'habitent pas nécessairement avec le parent qui les déclare.

Dans les faits, une famille qui semble être monoparentale pourrait bien être une famille dans laquelle les parents vivent en union libre (si les deux conjoints de fait ont chacun des enfants, leur famille pourrait passer pour deux familles monoparentales). Le personnel chargé de l'exploitation des données administratives étudie actuellement la possibilité de distinguer les familles dans lesquelles les parents vivent en union libre des familles monoparentales. Il s'agit pour cela de déterminer qui habite à la même adresse et, lorsqu'on a affaire à une famille monoparentale, de supposer que la personne de sexe opposé âgée de 0 à 12 ans de plus ou de moins que le parent unique et qui vit avec lui pourrait en réalité être son conjoint de fait. Dans une situation semblable sans enfant, on peut également supposer que les deux adultes sont des conjoints de fait. Grâce à ce raisonnement, il est possible de réduire le nombre estimé de familles monoparentales et de personnes seules. Comparativement aux statistiques du recensement, le nombre estimé de familles monoparentales semble encore trop élevé et le nombre d'unions libres, trop faible, mais les efforts qui ont été faits en vue de produire de meilleures estimations relativement à ces deux types de famille ont permis de faire baisser le nombre estimé de personnes seules obtenu d'après les dossiers fiscaux.

[11] On fait une vérification par recoupement afin d'éviter le double compte des enfants qui font une déclaration d'impôt et qui figurent en outre sur celle d'un de leur parent à titre de personne à charge.

Si l'on éliminait certains critères de recherche, on pourrait améliorer encore davantage les estimations relatives aux familles monoparentales, particulièrement celles qui sont entièrement composées d'enfants adultes et d'un parent âgé. Pour les statistiques du recensement, un parent et des enfants jamais mariés qui vivent ensemble constituent une unité familiale quel que soit l'âge de ces derniers. Mais pour les travaux qui visent la reconstitution des familles à partir des déclarations d'impôt, on limite à 29 ans l'âge d'une personne jamais mariée qu'on peut considérer comme l'enfant d'un certain parent. Étant donné le vieillissement de la population et l'accroissement du nombre de parents très âgés, il est fort possible que certains déclarants considérés comme des personnes seules dans les études qui se servent des données fiscales soient en réalité des enfants qui habitent le même logement que leur père ou que leur mère. On pourrait ainsi avoir une famille dans laquelle une mère de 80 ans vivrait avec ses enfants, lesquels seraient âgés de 50 à 55 ans[12].

4.4 Comparaison entre l'échantillon statistique et le fichier principal

Les principales statistiques publiées par Revenu Canada sont celles qui sont établies à partir des déclarations des particuliers et qui paraissent chaque année dans la publication **Statistique fiscale.** Elles sont plus détaillées que les données qui se trouvent dans le fichier principal et sont fondées sur un échantillon de déclarations d'impôt. En 1983, l'échantillon était constitué de 440,000 déclarations, ce qui représente un taux de sondage moyen de 2.9%. L'échantillon n'est pas un simple échantillon aléatoire; en 1983, il comportait 588 strates normales établies selon cinq caractéristiques : 1) la source du revenu (3 catégories); 2) le secteur géographique urbain (les villes à population semblable dans chaque région ont été regroupées pour un total de 15 groupes urbains); 3) le secteur géographique rural (chaque secteur comprend les régions exclues des secteurs urbains pour un total de 12 secteurs au Canada et un groupe de non-résidents); 4) le type de revenu (imposable ou non); 5) la tranche de revenu (4 tranches de revenu imposable et 3 tranches de revenu non imposable). Il y a deux autres catégories: les déclarants qui échappent à cette classification et ceux dont le revenu ou les déductions sont exceptionnellement élevés. Le taux de sondage varie entre 2 et 100%. L'échantillon a été choisi par ordinateur selon un mode de comptage séquentiel. La catégorie du revenu d'emploi dans les régions à forte densité de population des provinces de l'Ontario et du Québec a les taux de sondage les plus bas tandis que les régions et catégories à faible population ont les taux de sondage les plus élevés. Les catégories de revenu élevé et de revenu d'un travail autonome sont fortement représentées.

4.5 Occupation

Dans l'échantillon de **Statistique fiscale,** les déclarations sont classées selon l'occupation. Cette dénomination induit en erreur, car elle ne désigne pas la profession mais la principale source de revenu. Ainsi, un déclarant dont le revenu de placement est plus élevé que le revenu gagné est classé dans la catégorie des investisseurs plutôt que dans celle des salariés ou des travailleurs autonomes. Quant aux salariés, ils sont classés non pas selon la nature du travail qu'ils font mais selon le genre d'employeur pour lequel ils travaillent : entreprise, établissement (enseignants et professeurs), secteur public (administration fédérale ou provinciale, forces armées, société d'État provinciale ou fédérale) et salariés non classés. La classification est plus précise en ce qui a trait aux travailleurs autonomes, que ce soient les propriétaires d'une entreprise ou les professionnels; les premiers sont classés selon le genre d'entreprise (bâtiment, services publics, etc.) et les seconds, selon un nombre limité de professions (médecins, comptables, etc.).

[12] Il se peut que la limite de 29 ans soit éliminée à l'avenir.

Le fondement logique de la classification n'est pas évident. D'une part, certains formulaires ne contiennent de question ni sur l'occupation ni sur le nom de l'employeur, d'autre part, on suppose que les personnes sans emploi au moment où elles font leur déclaration n'ont pas d'employeur. Les questions sont si vagues que les déclarants n'y répondent probablement pas tous de la même façon. Enfin, il est difficile de dire si le classement des travailleurs autonomes se fonde sur les réponses données dans le formulaire T1 général ou sur les renseignements contenus dans l'état des résultats de l'entreprise ou de la profession.

En principe, si on prenait les renseignements relatifs aux sources de revenu d'emploi que contiennent les déclarations d'impôt, on pourrait produire à l'aide des données fiscales des statistiques presque conformes au concept de population active annuelle brute (ne comprenant pas les travailleurs familiaux non rémunérés) de Statistique Canada. Cependant, les données publiées ne permettent pas de le faire, car le nombre de personnes ayant déclaré un revenu d'emploi selon chacune des huit sources y est indiqué séparément; comme les actifs peuvent avoir un revenu provenant de plus d'une source, il est impossible d'éviter les doubles comptes. Ainsi, bien que 11,196,000 personnes aient déclaré un traitement ou un salaire en 1983, seulement 9,940,000 d'entre elles figurent dans la catégorie des salariés. Les 1.2 million de déclarants qui ne figurent pas dans cette catégorie ont touché seulement 3% du montant total des traitements et salaires déclarés.

5. ÉVALUATION DES DONNÉES ADMINISTRATIVES COMME SOURCE DE DONNÉES SOCIO-DÉMOGRAPHIQUES

Les prochaines sections traitent du genre de données qu'on pourrait tirer des dossiers administratifs pour remplacer celles du recensement si celui de 1991 n'avait pas lieu ou s'il se limitait à un simple dénombrement de la population.

Comme nous l'avons déjà vu, les concepts utilisés dans les fichiers administratifs sont assez différents des concepts retenus à Statistique Canada, de sorte que les données administratives peuvent être de nature semblable à celles produites au Bureau sans être nécessairement comparables. Nous allons énumérer dans la section qui suit les séries de données annuelles susceptibles d'être tirées des fichiers de données administratives tels qu'ils existent actuellement. Nous parlerons plus loin des modifications qu'il faudrait apporter aux formulaires de déclaration d'impôt pour augmenter les possibilités statistiques des données fiscales.

On pourrait se servir des données fiscales actuelles et les compléter avec celles que contiennent notamment les dossiers du régime de sécurité de la vieillesse (SV) et ceux des allocations familiales pour produire les séries suivantes par secteur géographique.

5. Estimations de la population selon l'âge, le sexe et l'état matrimonial.

Actuellement, Statistique Canada produit chaque année des estimations de la population selon le sexe, l'âge et la province. Celles-ci sont corrigées en fonction des données du recensement et, dans la période intercensitaire, établies à partir des statistiques de l'état civil et des données sur la migration provenant de diverses sources comme les dossiers de la sécurité sociale et les dossiers fiscaux.

Le recensement fournit des données semblables au niveau infraprovincial. Quant aux données fiscales, on peut en tirer directement ou indirectement des chiffres de population selon le sexe et, dans la plupart des cas, l'âge. Comme nous l'avons vu plus haut, leur univers présente des lacunes en ce qui a trait aux jeunes et aux personnes âgées. Les dossiers fiscaux permettent d'estimer le nombre de jeunes à partir des exemptions pour enfants entièrement à charge que les parents demandent. Mais comme le montant de l'exemption ne fournit qu'une vague indication de l'âge des enfants à charge, il manque

encore le sexe et l'âge de ces derniers. Les déclarations d'impôt contiennent ces renseignements, mais ceux-ci ne figurent pas dans les fichiers informatiques[13]. Les dossiers des allocations familiales peuvent compléter les dossiers fiscaux pour ce qui est du nombre d'enfants par région et de leur âge, mais apparemment eux non plus n'indiquent pas le sexe. Toutefois, si l'on utilisait les données du recensement comme données de référence, on pourrait se servir des dossiers fiscaux et des dossiers des allocations familiales pour effectuer des projections au niveau infraprovincial entre les recensements.

Les personnes âgées constituent l'autre groupe manquant. Lorsque l'un des deux conjoints fait une déclaration d'impôt, il est possible d'estimer l'âge de l'autre à partir des dossiers fiscaux. Lorsque aucun des conjoints ne fait de déclaration d'impôt, mais que l'un des deux reçoit des prestations de la SV ainsi que le SRG, il est possible de déterminer l'âge de l'autre. Ce sont principalement les bénéficiaires de la SV et du SRG qui ne font pas de déclaration d'impôt, et la plupart d'entre eux sont des femmes qui vivent seules. Depuis que le RPC et le RRQ existent, la plupart des hommes à la retraite ont un revenu, et ce revenu est dans un nombre croissant de cas imposable. Chez les déclarants, 88% des hommes touchaient une rente du RPC ou du RRQ en 1983, contre 59% des femmes seulement. La proportion des personnes âgées qui produisent une déclaration d'impôt va augmenter avec le temps.

Au mois de juin 1985, seulement 36% des bénéficiaires de la SV et du SRG, étaient des hommes, dont un peu plus des deux tiers étaient mariés. Ce qui veut dire que près de 64% des bénéficiaires de la SV et du SRG étaient des femmes, dont les deux tiers environ étaient classées dans la catégorie des célibataires (catégorie qui comprend probablement les veuves). La plupart des personnes âgées qui ne font pas de déclaration d'impôt sont donc des femmes seules et, lorsqu'il s'agit de bénéficiaires de la SV et du SRG qui sont mariés, nous pouvons connaître l'âge du conjoint.

Il fut un temps où l'une des lacunes des données fiscales était l'absence de données sur les personnes pour lesquelles les prestations d'assistance sociale et de bien-être social représentaient l'unique source de revenu. La forte proportion de la population visée par les allocations familiales a changé cette situation. Aujourd'hui, un grand nombre de femmes produisent une déclaration d'impôt alors que leur revenu imposable est faible ou nul, ce qui semble indiquer qu'il s'agit de mères de famille qui, pour la plupart, font cette déclaration parce qu'elles veulent toucher le crédit d'impôt pour enfants. Celui-ci est versé principalement aux parents à faible revenu, et les assistés sociaux ayant des enfants à charge y sont donc admissibles.

Il convient d'apporter certaines précisions. Grâce au code postal, il est possible d'agréger les estimations de population relatives aux secteurs urbains effectuées à partir des données administratives, mais il est plus difficile de le faire pour les secteurs ruraux. Même en ce qui concerne les secteurs urbains, comme l'adresse postale diffère parfois de celle du domicile, les estimations relatives aux petits secteurs comme les secteurs de recensement sont moins fiables que celles qui se rapportent aux grandes agglomérations. Finalement, les problèmes sont les mêmes pour ce qui est des dossiers administratifs que pour les dossiers fiscaux; ainsi, 18% des personnes qui ne touchaient que la SV comme revenu se faisaient envoyer leur chèque chez un fiduciaire, à la banque ou au bureau de poste.

Tout changement apporté aux programmes universels de sécurité sociale au cours des prochaines années risque de diminuer l'utilité des dossiers des allocations familiales et de la SV comme compléments des dossiers fiscaux pour les personnes qui ne font pas de déclaration d'impôt. Cette perte sera en toute probabilité compensée par une augmentation du taux de déclaration.

[13] À l'avenir, les enregistrements de Revenu Canada indiqueront l'âge (mais pas le sexe) des enfants à charge.

5.2 Estimations de la mobilité

Le recensement recueille des données sur la mobilité au cours de la période intercensitaire, c'est-à-dire sur les changements éventuels de domicile d'un recensement à l'autre. Les personnes sont classées dans des catégories différentes selon qu'elles occupent le même logement qu'au recensement précédent, qu'elles ont déménagé mais à l'intérieur de la même municipalité, qu'elles ont déménagé dans une autre municipalité ou qu'elles arrivent de l'extérieur du Canada.

Les dossiers fiscaux contiennent deux éléments d'information se rapportant à la mobilité annuelle : la province de résidence au mois de décembre et l'adresse au moment où la déclaration est faite. Ils ont un avantage sur le recensement du fait que les données sont recueillies chaque année, et donc qu'elles renseignent sur la mobilité pendant la période intercensitaire. Par ailleurs elles permettent d'obtenir presque les mêmes informations que celles produites grâce au recensement : pas de changement de domicile, déménagement à l'intérieur d'une municipalité (au moyen du code postal) et déménagement d'une municipalité à une autre. Il est en outre possible d'analyser les caractéristiques des migrants selon d'autres variables comme le sexe (sauf pour les enfants), l'âge et le revenu.

5.3 Population active annuelle brute, chômeurs et gains

Les dossiers fiscaux semblent porter sur la quasi-totalité de la population active, à l'exception des très jeunes travailleurs comme les étudiants qui ont un emploi à temps partiel. Du fait que les employeurs sont tenus de produire des feuillets de renseignements indiquant les retenues à la source, les données sur les traitements et salaires devraient en principe être très précises. Pour ce qui est des personnes âgées de plus de 20 ou 24 ans, les données fiscales sur le sexe, l'âge et les gains de la population active annuelle brute sont probablement de qualité égale ou supérieure aux données du recensement, ces dernières étant tributaires de la mémoire des recensés qui peut, avec le passage du temps, faire défaut. Les données fiscales se rapprocheraient des données concernant les personnes actives sur le marché du travail l'année précédente, si l'on ne tient pas compte des travailleurs familiaux non rémunérés (une espèce en voie de disparition). Elles permettraient de calculer le taux d'activité annuel et le taux annuel brut de chômage[14]. Les données fiscales fourniraient également des renseignements sur le revenu des déclarants en chômage et, en appariant les déclarations de ces derniers à celles de leur conjoint, on pourrait analyser le lien éventuel avec les gains du conjoint. Les données sur les sources des gains permettraient de reproduire approximativement les catégories de travailleurs, en l'occurrence : travailleur rémunéré seulement; travailleur autonome seulement; les deux, mais principalement travailleur rémunéré; les deux, mais principalement travailleur autonome.

Le recensement permet de mesurer l'activité sur le marché du travail au moment où il a lieu et de savoir si les inactifs étaient actifs plus tôt dans l'année ou au cours de l'année précédente. Il fournit donc des renseignements sur l'activité au cours des 17 mois qui l'ont précédé ainsi que sur la profession, la branche d'activité et la catégorie de travailleurs. Les personnes actives au moment du recensement constituent la "population active actuelle" tandis que les personnes actives au cours de l'année civile précédente représentent la "population active annuelle brute". L'enquête sur la population active (EPA) mesure la population active actuelle, mais permet également de connaître la dernière période d'activité des personnes qui ne font pas partie de cette population. L'enquête sur l'activité annuelle, supplément annuel de l'EPA, mesure la population active annuelle brute.

[14] Les personnes malades ou en congé de maternité ont droit, elles aussi, à l'assurance-chômage.

On utilise beaucoup ces deux concepts pour présenter les données du recensement sur la population active. Le concept de population active annuelle brute permet d'estimer la proportion de la population qui est active sur le marché du travail pendant une période donnée, et il se rapporte à un plus grand nombre de personnes que le concept de population active actuelle. C'est en outre celui des deux qui se prête le mieux à l'étude des données sur les gains.

Les données tirées des dossiers fiscaux indiquent, sans double compte, le nombre de personnes en chômage au cours de l'année, et elles se rapprochent de celles qu'on obtient au moyen de l'enquête sur l'activité annuelle. Le recensement, quant à lui, ne produit pas de tels renseignements. Par contre, les données fiscales sous-estiment le nombre de chômeurs, car certains d'entre eux ne font pas de déclaration d'impôt et d'autres ne touchent pas de prestations d'assurance-chômage. Mais il y a, en revanche, des personnes qui reçoivent des prestations d'assurance-chômage parce qu'elles sont malades ou enceintes, et non parce qu'elles sont en chômage.

À Statistique Canada, la population active est classée selon la catégorie de travailleurs. Autrement dit, les répondants sont classés selon leur principale occupation au moment de l'enquête. Dans les statistiques fiscales, le classement est fait en fonction de la source du revenu d'emploi.

5.4 Répartition du revenu

Étant donné la nature des concepts utilisés, les dossiers fiscaux ne contiennent pas de données sur le revenu d'une partie de la population qui reçoit des paiements de transfert non imposables et de certains groupes d'âge, notamment les jeunes. Le champ d'observation semble excellent pour ce qui est des personnes âgées de 20 à 64 ans, mais il ne tient pas compte des jeunes travailleurs et il est inadéquat en ce qui concerne les personnes dont le revenu est constitué de prestations du bien-être social. Quant aux personnes âgées qui ne font pas de déclaration d'impôt, on peut estimer leur revenu d'après les dossiers administratifs de la SV et du SRG.

Le fait qu'une faible proportion de jeunes seulement déclarent leur revenu au fisc ne représente pas un très grand obstacle à l'estimation du revenu des familles, car les parents qui déclarent des enfants à charge doivent indiquer le revenu de ces derniers. Le problème est que ces données ne semblent pas être saisies : elles existent mais ne sont pas exploitables par une machine. Si les jeunes produisaient une déclaration d'impôt, on aurait accès à ces données. En ce qui concerne les personnes âgées qui ne font pas de déclaration (celles qui touchent des prestations de la SV et le SRG), on peut estimer leur revenu d'après d'autres données administratives. La seule partie de la population dont le revenu (individuel ou familial) ne peut être connu est celle qui est composée de familles monoparentales vivant de l'assistance sociale.

Les personnes âgées qui ne font pas de déclaration d'impôt semblent être en grande partie des personnes qui demandent le SRG. Celles qui y ont droit doivent déclarer le montant et la source des revenus qu'elles ont eu au cours de l'année précédente. Les seules sources exclues sont certains paiements de transfert non imposables comme les allocations d'ancien combattant. Les deux conjoints doivent faire une déclaration d'impôt commune. On pourrait se servir des dossiers de la SV pour estimer le revenu des personnes âgées qui ne produisent pas de déclaration d'impôt.

Les femmes qui bénéficient de l'assistance sociale présentent un problème particulier, car elles font une déclaration d'impôt pour pouvoir recevoir le crédit d'impôt pour enfants alors que leurs principales sources de revenu sont non imposables et ne figurent donc pas sur la déclaration. Nous reparlerons de ce problème plus loin.

5.5 Statistiques relatives aux familles

Nous avons déjà parlé de ce qui peut se faire en matière de reconstitution des familles à partir des statistiques fiscales. Il est possible de produire les séries suivantes : statistiques sur les particuliers ainsi que données sur les couples mariés et sur les familles de recensement (mais pas les familles économiques). J'ai déjà indiqué mes réserves au sujet de la production de statistiques sur les ménages.

Il est bien sûr possible de faire des classements recoupés à l'aide de toutes les séries de données que nous venons d'énumérer et de construire des tableaux illustrant par exemple le revenu ou le revenu d'emploi selon le sexe, l'âge et l'état matrimonial ou encore le revenu familial selon l'âge des conjoints ou la taille de la famille. Ainsi qu'on a pu le voir dans le résumé, les dossiers fiscaux et autres dossiers administratifs ne contiennent pas toute la gamme de données que les recensement de 1981 et de 1986 ont recueillies et donc ne sauraient remplacer ces importantes enquêtes, mais ils peuvent fournir chaque années des données régionales équivalentes ou complémentaires. Il faut cependant que l'unité géographique de base soit le code postal.

6. ÉVALUATIONS PROPOSÉES DES DONNÉES ADMINISTRATIVES

Si les ressources le permettent, il faudrait faire des évaluations postcensitaires comparant les statistiques du recensement aux statistiques administratives. Nous suggérons trois sortes d'évaluations : évaluations à un macroniveau, appariements à un microniveau et recherches visant l'amélioration des données fiscales en vue de leur utilisation à des fins statistiques.

Les évaluations à un macroniveau seraient les plus faciles à faire, car elles nécessiteraient uniquement, comme outils, des totalisations et, comme ressources, des analystes. Pour les appariements à un microniveau, il faudrait des ressources et un budget spéciaux, et là encore il se peut qu'on ne puisse pas les terminer à temps pour changer quoi que ce soit au recensement de 1991. Ces appariements permettraient de mieux évaluer le lien qui existe entre les concepts utilisés pour le recensement et ceux utilisés pour les données administratives et ils nous renseigneraient sur les améliorations qu'il faudrait apporter aux données administratives. Les recherches visant l'amélioration des dossiers fiscaux en vue de leur utilisation à des fins statistiques exigeraient davantage de ressources encore et ne pourraient être terminées avant que les plans relatifs au recensement de 1991 ne soient arrêtés; pourtant elles nous apprendraient des choses essentielles, à savoir dans quelle mesure on pourrait améliorer les déclarations d'impôt pour que les dossiers fiscaux puissent remplacer les recensements ultérieurs, comme celui de 1996.

L'étude dont cette communication s'inspire contenait des recommandations plus détaillées à propos des évaluations et des recherches susceptibles de permettre une meilleure interprétation des données de recensement et des données fiscales. Les sections qui suivent résument ces recommandations.

6.1 Évaluations à un macroniveau

Il faudrait faire des totalisations de base à divers niveaux régionaux au moyen à la fois des statistiques fiscales et des statistiques du recensement afin d'évaluer le champ d'observation des dossiers fiscaux. Les concepts devraient être modifiés afin que les données deviennent comparables. Les séries à comparer pourraient porter, par exemple, sur le revenu total selon le sexe, l'âge, l'état matrimonial et la taille du revenu. Il faudrait calculer, selon le sexe et l'âge, le pourcentage de la population qui reçoit un revenu et celui de la population qui fait une déclaration d'impôt, également selon le sexe et l'âge.

Une autre évaluation devrait comparer, selon une classification semblable, l'univers de la population active qui est représenté dans les déclarations d'impôt à celui qui est représenté dans le recensement.

Les totalisations mentionnées ci-dessus se rapportent aux individus. Il faudrait également comparer le revenu des couples mariés (l'ensemble des couples mariés et les couples mariés faisant partie de la population active) selon les caractéristiques qui leur sont propres.

Nous avons déjà parlé de la difficulté de savoir si on a affaire à une famille monoparentale ou à un couple vivant en union libre d'après les déclarations d'impôt; on est en train d'essayer de trouver un moyen de reconstituer ce genre de familles à partir des données fiscales. Il est encore plus difficile de comparer les données de recensement sur ces familles à un macroniveau à cause des problèmes conceptuels que cela présente. Pour faire ces comparaisons, il faudrait ventiler les données selon les caractéristiques des familles, comme l'âge des membres et les sources de revenu.

Dans la classification de l'occupation utilisée pour les statistiques fiscales, un grand nombre de déclarants sont mis dans la catégorie des "déclarants non classés" ou des "salariés non classés". Il faudrait pousser plus loin l'analyse des caractéristiques de ces groupes.

6.2 Appariement à un microniveau

Il s'agirait de procéder, d'une part, à l'appariement des enregistrements en vue de déterminer quelles populations, dans les dossiers fiscaux, posent des problèmes et, d'autre part, à certaines études visant à établir si d'autres données, comme celles qui proviennent des dossiers de la SV et du SRG, pourraient compléter les données fiscales. On pourrait prendre un échantillon de déclarations produites par des femmes pour lesquelles on ne trouve pas de déclarations de conjoint et les apparier avec les données de recensement se rapportant à ces mêmes femmes afin de voir si les données fiscales permettent effectivement de déterminer quand on a affaire à un parent seul, à une personne vivant en union libre ou à une personne seule. Par ailleurs, de nombreuses femmes dont le revenu est faible ou nul font une déclaration d'impôt dans le seul but de toucher le crédit d'impôt pour enfants. Un bon nombre d'entre elles sont des assistées sociales dont le revenu n'est pas imposable. L'appariement à un microniveau des données fiscales et des données de recensement les concernant permettrait de calculer le nombre de cas pour lesquels les données fiscales sur le revenu sont incomplètes ou inadéquates.

Nous venons de suggérer l'appariement d'un échantillon de déclarations d'impôt et des données de recensement concernant les mêmes personnes en vue de cerner les problèmes auxquels on se heurte lorsqu'on essaie de mettre en évidence la structure d'une famille. Comme nous l'avons vu plus haut, les groupes qui sont le plus sous-représentés sont les jeunes et les personnes âgées. Pour ce qui est des jeunes, il faudrait prendre un échantillon de données de recensement s'y rapportant et les apparier aux données fiscales concernant les membres de l'échantillon afin d'étudier les caractéristiques de l'univers des non-déclarants, notamment les liens à l'intérieur du ménage, la fréquentation scolaire, l'activité, les gains et la dépendance à l'égard du revenu des parents.

Quant aux personnes âgées qui ne sont pas représentées dans les fichiers fiscaux, pour être plus renseigné à leur égard, il faudrait apparier les données fiscales à un échantillon de dossiers tirés des fichiers de la SV et du SRG. Les données de la SV et du SRG laissent croire que les personnes qui ne font pas de déclaration d'impôt sont principalement des personnes seules à faible revenu, lesquelles devraient presque toutes recevoir le supplément de revenu garanti. Or, étant donné qu'il faut faire au moins une déclaration d'impôt pour avoir droit au SRG, il devrait être possible de connaître par déduction la répartition du revenu parmi les non-déclarants.

6.3 Améliorations à apporter aux déclarations d'impôt en vue de leur utilisation à des fins statistiques

À l'heure actuelle, les déclarations d'impôt sont conçues en vue de répondre à des besoins administratifs et ce n'est qu'indirectement qu'elles servent à des fins statistiques. Cependant, elles pourraient éventuellement permettre d'obtenir un plus grand nombre de données plus utiles sur la profession des déclarants et la branche d'activité dans laquelle ils l'exercent à condition notamment d'avoir la collaboration de Revenu Canada. Pour obtenir ces données, on pourrait s'y prendre de trois façons: 1) en augmentant l'espace réservé à la participation au marché du travail dans les déclarations d'impôt, 2) en demandant aux déclarants de cocher la catégorie professionnelle à laquelle ils appartiennent et 3) en se servant des fichiers fiscaux comme base de sondage pour effectuer des enquêtes postales sur les caractéristiques de la population active. L'échantillon pourrait être stratifié, entre autres, selon le niveau de revenu ou la région. Rien ne garantit toutefois qu'on arriverait ainsi à améliorer les données sur la population active, car le public pourrait s'opposer à ce qu'on se serve des déclarations d'impôt pour recueillir des données purement statistiques.

7. CONCLUSION

Actuellement, il est possible de produire à l'aide des déclarations d'impôt et de certaines données supplémentaires des données socio-démographiques de base très importantes pour la quasi-totalité de la population, et ce au niveau régional, quoique les secteurs ruraux soient plus difficiles à délimiter que les secteurs urbains. Le potentiel statistique des déclarations d'impôt pourrait être exploité encore davantage si on y mettait les ressources nécessaires et qu'on obtenait la collaboration de Revenu Canada. Cependant, les changements que ce ministère a commencé à introduire à partir de l'année d'imposition 1986 ont atténué certains problèmes qui existaient au moment où l'étude sur laquelle se base cette communication a été faite et qui y sont mentionnés. Ainsi, pour avoir droit à certains crédits, les personnes qui produisent une déclaration d'impôt devront y indiquer le montant des revenus non imposables qu'ils auront touché, par exemple les prestations d'aide sociale. Ce changement touchera principalement les personnes à faible revenu, et il aura pour conséquence de rendre les données fiscales sur le revenu plus complètes. Il éliminera également certaines différences qui existent au niveau des concepts entre les données fiscales et celles de Statistique Canada. En outre, la réforme fiscale aura probablement pour autre conséquence d'augmenter la proportion de la population qui produit une déclaration d'impôt. Les estimations effectuées relativement à l'ensemble de la population en seraient alors plus exactes.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

LA QUALITÉ DES DONNÉES ADMINISTRATIVES D'UN POINT DE VUE STATISTIQUE L'EXPÉRIENCE DU DANMARK

POUL JENSEN[1]

RÉSUMÉ

Cet article décrit sous divers aspects (utilité générale des données, caractéristiques techniques et qualité des résultats) l'expérience du Danemark dans l'utilisation de données administratives comme source statistique de premier plan. Il traite aussi brièvement les problèmes que soulève la comparaison de la qualité de données de diverses sources et expose dans une perspective plus vaste le problème de la substitution des données administratives aux données de recensement classiques.

1. REMARQUES PRÉLIMINAIRES

L'utilisation des dossiers administratifs comme source de données statistiques n'est pas un phénomène nouveau et dans la pratique, c'est souvent la seule façon raisonnable de produire des statistiques.

Les avantages et les inconvénients de l'utilisation statistique des dossiers administratifs au sens classique sont donc bien connus. Toutefois, l'implantation des techniques modernes dans l'administration publique depuis vingt-cinq ans a multiplié les possibilités d'utilisation des registres administratifs comme source de données statistiques et ce, pour les raisons suivantes.

Premièrement, les données sont classées pour la plupart dans des registres informatisés, ce qui signifie en principe qu'elles sont plus faciles à obtenir et à traiter et qu'elles sont vraisemblablement de meilleure qualité car l'informatisation exige des méthodes d'enregistrement systématiques.

Deuxièmement, la création simultanée de registres administratifs et de systèmes généraux d'identification fondés essentiellement sur les personnes (à tous le moins dans certains pays comme le Danemark) a ouvert la voie au couplage d'enregistrements, par lequel on combine des données de différentes sources. Les pays qui n'avaient pas de système général d'identification fondé sur les personnes ont ébaloré des méthodes de couplage qui se sont avérées suffisamment fiables pour les besoins statistiques.

Troisièmement, grâce à la puissance des ordinateurs actuels, la collecte de données à des fins administratives a maintenant une portée beaucoup plus grande.

Quatrièmement, dans certains pays, les registres de personnes, d'entités commerciales, etc., constituent efficacement la base pour la collecte de données statistiques selon des méthodes classiques.

[1] Poul Jensen, directeur, General Economic Statistics, Danmarks Statistik, Sejrøgade 11, 2100 København, Danmark.

Nous tentons ci-dessous de dégager les conséquences de cette évolution sur le plan statistique pour le Danemark.

L'adoption de la loi créant Danmarks Statistik en 1966 laissait présager l'informatisation de l'administration publique et l'exécution de programmes statistiques qui auraient été impensables sans cette informatisation. De fait, les événements qui ont découlé de l'adoption de cette loi allaient se succéder très rapidement. Il y a eu tout d'abord la création du système de la taxe sur la valeur ajoutée en 1967, puis la création du registre central de la population (CPR) l'année suivante; en 1970, on a mis sur pied un système de retenue fiscale (retenue à la source) et dans la première moitié des années 1970, on a appliqué un certain nombre de réformes sociales qui nécessitaient l'utilisation de registres informatisés. En 1975, on créait le registre central des entreprises et des établissements (Det centrale erhvervsregister), dont la gestion devait être confiée plus tard à Danmarks Statistik, et, deux ans plus tard, on créait le registre central des immeubles et des logements (BBR).

Dès le départ, Danmarks Statistik avait pour objectif d'exploiter les possibilités statistiques qu'offraient ces registres et d'autres du même genre. C'est pourquoi une forte proportion des statistiques produites par l'organisme public sont aujourd'hui fondées sur des données de registres administratifs placés sous la responsabilité des autorités publiques; de plus, les données de registres en sont venues à remplacer les données de recensements à grande échelle comme les recensements d'entreprises et les recensements de la population et du logement.

Le principe des registres administratifs a été appliqué à d'autres statistiques de première importance dont le dépouillement se faisait à partir de formules administratives. Parmi ces statistiques, qui reposent maintenant sur des données extraites de bases ordinolingues, mentionnons les statistiques du commerce extérieur et les statistiques du chômage. En outre, de nouvelles statistiques ont fait leur apparition, par exemple les statistiques des ventes en général (selon les données sur la TVA) et celles de l'emploi en général (selon les données sur l'ATP, c'est-à-dire les données relatives au régime de retraite supplémentaire du marché du travail). Les secteurs plus traditionnels comme ceux des statistiques démographiques et des statistiques sociales ont eux aussi profité des nouvelles techniques par un élargissement et une amélioration notables de leurs bases de données, de sorte que nous disposons maintenant de très grandes quantités de données détaillées. Dans presque tous les autres secteurs, les registres administratifs servent d'une manière ou d'une autre de base de données.

Il ne fait aucun doute que les progrès décrits ci-dessus ont amené une hausse sensible de la production de statistiques. Il est moins sûr que la qualité de ces statistiques soit satisfaisante. Le présent article se prête mal à une description complète du problème qui, de fait, est très étendu, mais les thèmes suivants reflètent quelques-uns des problèmes de qualité.

- quelle a été l'incidence des progrès décrits ci-dessus sur l'utilité générale des statistiques (section 2);
- quelles sont les caractéristiques techniques des données extraites de fichiers administratifs (section 3);
- quelles sont les caractéristiques des statistiques fondées sur des registres lorsqu'on applique des critères courants de qualité comme la fiabilité, la continuité, l'actualité, etc. (section 4).

Les sections 2, 3 et 4 ont uniquement pour thème l'utilisation de données administratives au Danemark. Le section 5 contient une brève comparaison de la qualité des données administratives et de celle d'autres sources de données statistiques; enfin, le section 6 expose certaines opinions inspirées par le débat international sur le problème du recensement.

2. L'utilité générale des statistiques

Il est clair que les données se rattachant à un secteur particulier de l'administration publique s'imposent comme des données primaires pour décrire les activités de ce secteur. Il n'est pas sûr toutefois -- et les nombreuses discussions à ce sujet le prouvent -- que ces données puissent se prêter à un usage général puisque leur utilité est déterminée et circonscrite par les procédés administratifs auxquels elles correspondent.

Il ne fait pas de doute que certaines catégories de données ne se prêtent pas à un usage général. Toutefois, si l'on envisage le problème globalement, l'analyse est quelque peu différente.

Premièrement, l'argument ne tient pas pour ce qui a trait aux **registres de base** comme le registre central de la population, le registre central des entreprises et des établissements, etc., qui ont pour but de fournir de l'information destinée à plusieurs usages et à plusieurs directions administratives. Les registres de ce genre ne devraient pas être influencés de façon catégorique par leur utilisation à des fins spécifiques. Le contenu des registres de base devrait donc répondre en partie aux mêmes exigences que les statistiques générales.

Deuxièmement, il est possible de combiner des données de différents systèmes administratifs grâce au couplage d'enregistrements. Le couplage d'enregistrements vise principalement à produire des combinaisons de données qui se rattachent à divers secteurs statistiques mais -- aspect important du couplage -- qui ne renferment pas nécessairement une trop forte proportion de données d'un seul service. Le secteur statistique du Danemark qui repose le plus sur des données administratives renferme des données provenant de plus d'un registre et notamment des données provenant d'au moins un registre de base.

Troisièmement, il est possible dans certains cas de compléter les données administratives, lorsque cela est nécessaire, par des données qui visent à les rendre plus conformes à des usages statistiques généraux. Le principal exemple du genre au Danemark est le projet "du lieu de travail", qui permet de classer les activités économiques en fonction de l'établissement plutôt qu'en fonction de l'unité (quasi-entreprise) qui se trouve dans le registre administratif en question.

Quatrièmement, il est plus ou moins possible (à tout le moins au Danemark, grâce à une disposition spéciale de la loi créant Danmarks Statistik) de modifier le contenu des registres administratifs en vue de son utilisation statistique.

Il faut convenir que l'effet global de ces modifications est si prononcé que la thèse de l'utilité restreinte des données administratives ne peut être considérée comme un principe général. Il faut néanmoins reconnaître deux choses: (1) parmi les nombreuses données en jeu, il y en aura toujours qui seront si étroitement liées à un processus administratif donné que leur utilité pourrait être légèrement restreinte dans la perspective de l'utilisation de statistiques générales; (2) les principes appliqués pour l'utilisation de registres administratifs peuvent être différents de ceux qui sous-tendent les méthodes de collecte classiques; par exemple, le principe de la commutation appliqué aux statistiques fondées sur des registres au Danemark diffère quelque peu de celui adopté pour les recensements classiques de la population et du logement.

Seule une analyse détaillée des données et de leur utilisation à des fins diverses permettra de déterminer dans quelle mesure les réserves qui précèdent amoindrissent la valeur des statistiques fondées sur des registres par rapport à celle des statistiques dépouillées suivant des méthodes classiques. Néanmoins, on peut affirmer que l'harmonisation avec les définitions internationales sera vraisemblablement plus difficile s'il s'agit de données administratives au lieu de données "purement" statistiques.

En ce qui concerne la valeur statistique générale des données administratives, il convient, en dernier lieu, de faire les remarques suivantes:

- L'utilisation d'identificateurs généraux signifie que les mêmes éléments d'information peuvent servir dans différents secteurs de la statistique. Par exemple, comme la classification industrielle utilisée pour le registre central des entreprises et des établissements se retrouve maintenant dans presque tous les secteurs de la statistique, on est assuré de la **cohérence** des données.
- Le couplage d'enregistrements est aussi un outil d'analyse précieux en permettant de réaliser **des analyses transversales** de micro-données de même que des analyses **longitudinales.**

Ces caractéristiques, qui sont bien sûr très importantes, ne se retrouvent pas dans les statistiques établies selon des méthodes classiques: la cohérence transversale accroît la valeur analytique des statistiques générales. Les enquêtes générales réalisées par interview, qui sont un substitut aux analyses transversales, sont extrêmement coûteuses. Par ailleurs, les enquêtes à échantillon constant réalisées uniquement dans le but de créer une banque de données pour des enquêtes longitudinales futures posent énormément de problèmes au point de vue des coûts et à d'autres égards. Enfin, la collecte rétrospective de données pour des périodes plus longues donne des résultats de qualité douteuse.

3. Caractéristiques techniques de données extraites de fichiers administratifs

À l'aube de l'ère informatique, beaucoup croyaient que les données traitées par ordinateur constitueraient un excellent outil statistique. On supposait que les fournisseurs de données de même que les autorités qui allaient traiter et utiliser ces données admettraient peu d'erreur dans les données administratives. Après tout, la présence d'erreurs dans les données administratives pouvait avoir des conséquences embarrassantes pour les citoyens et nuisibles pour les procédés administratifs. Beaucoup d'optimistes croyaient qu'une fois les systèmes mis sur pied et les problèmes d'exécution résolus, il ne suffisait plus que de "presser le bouton" pour obtenir rapidement des statistiques fiables. Ils espéraient donc que les nouvelles techniques mises à la disposition des producteurs de statistiques leur permettraient de résoudre plus facilement les problèmes qui caractérisent depuis longtemps les statistiques, soit le manque d'exactitude et la longueur indue des délais de production.

Il y avait aussi de nombreux sceptiques qui étaient surtout préoccupés par la teneur des données et les inconvénients censés amoindrir la valeur statistique de ces données.

Comme nous l'avons vu au section 2, les craintes des pessimistes ont été démenties tout comme les espoirs des optimistes. L'expérience danoise recèle plusieurs exemples de données erronées et de perturbations dans les procédés administratifs - et, en l'occurrence, dans les procédés statistiques -malgré les efforts considérables qui ont été faits dans certains cas pour résoudre ce genre de problèmes. Par ailleurs, d'autres exemples montrent qu'après des années de recherche intense, il est possible de créer des conditions très proches des conditions originelles optimales - par exemple, en ce qui concerne les statistiques démographiques fondamentales du Danemark - mais d'autres exemples encore illustrent le contraire.

On suppose habituellement que le genre de problèmes mentionnés ci-dessus ne sont pas inhérents aux fichiers administratifs et qu'ils peuvent être résolus dans chaque secteur après une période d'expérimentation, du moins s'il n'y a pas une pénurie de spécialistes de l'informatique.

Toutefois, à la lumière de l'expérience du Danemark, il est juste de se demander si les problèmes que nous venons d'évoquer sont à ce point fondamentaux qu'ils ne peuvent être résolus autrement que par l'élaboration systématique de solutions méthodiques.

Le présent article n'a pas pour but d'approfondir cette question mais l'expérience danoise nous permet d'exposer certains problèmes universels touchant l'utilisation des registres administratifs comme source de données statistiques:

1. Problèmes de communication

Il est reconnu que les activités informatiques à grande échelle exigent une planification et une supervision efficaces. C'est particulièrement le cas lorsque plusieurs parties dépendent d'un même système informatisé. On peut au moins dire qu'il y a des informaticiens qui exercent des fonctions différentes et des utilisateurs qui se situent à des niveaux différents. Le processus se complique d'autant plus lorsqu'on étend ce réseau complexe de coopération à l'utilisation statistique. Il devient donc indispensable d'avoir une documentation complète et entièrement mise à jour. Malheureusement, les systèmes administratifs informatisés ne sont pas à l'abri des changements de dernière minute (ni les systèmes statistiques d'ailleurs!). Des ordinateurs centraux différents ont des normes différentes en ce qui a trait aux formes de documentation et cela crée des imperfections, des erreurs ou des confusions.

2. Problème des erreurs de procédure

L'extraction de données des registres administratifs à des fins statistiques est bien sûr exposée à des erreurs de programmation, à des erreurs de procédure, etc. Dans le cas des données administratives, les erreurs ne tardent habituellement pas à paraître, ce qui n'est pas le cas des données statistiques. De plus, les autorités administratives intéressées n'accordent pas nécessairement une grande priorité aux données purement statistiques et leurs informaticiens ne sont pas particulièrement versés en statistique.

3. Problème du traitement des erreurs

Le traitement statistique des données a toujours comporté un mécanisme de détection des erreurs. Il n'existe pas encore de mécanisme semblable pour le traitement administratif des données ni pour l'usage à des fins administratives de données ayant fait l'objet d'un contrôle.

Au moment du traitement statistique des données, celles-ci devraient évidemment faire l'objet d'une vérification, mais il est peut-être déjà trop tard pour relever les erreurs et les corriger. De toutes façons, la correction d'erreurs entre établissements est une tache difficile et fastidieuse. De plus, les autorités chargées de la supervision des registres au Danemark hésitent à approuver de telles mesures venant des autorités statistiques parce que les lois concernant le couplage d'enregistrements, l'accès libre aux registres, etc., sont moins sévères dans le cas des registres statistiques que dans le cas des registres administratifs. Cela incite le comité de surveillance des données à appliquer très vigoureusement le "principe de la circulation à sens unique".

4. Problème du contrôle et de la combinaison

Les méthodes qui servent à convertir les données administratives primaires en données statistiques sont très complexes. Elles supposent une planification et une supervision du traitement des données tellement soignées qu'en pratique, ces opérations peuvent être difficiles à exécuter.

5. Absence de contrôle visuel

Les problèmes énumérés ci-dessus sont aggravés par le fait que les données traitées par ordinateur sont invisibles par définition. Tandis que les données primaires classiques se prêtent au contrôle visuel, qui permet de repérer sur le champ les erreurs les plus élémentaires, des méthodes spéciales sont nécessaires pour soumettre les données informatisées aux opérations de contrôle courantes. De fait, dans certains cas les erreurs ne sont pas décelées avant que les résultats finals soient connus et même, parfois, avant qu'ils aient été publiés. Cela crée évidemment une situation plutôt embarrassante.

La liste des problèmes énumérés ci-dessus montre qu'il n'est pas facile de surmonter les problèmes "techniques" liés à l'utilisation des registres administratifs comme sources statistiques.

Les points suivants semblent mettre en lumière les causes fondamentales de ces problèmes:

a. Les méthodes de traitement utilisées dans un contexte statistique ne tiennent pas compte de toutes les différences fondamentales entre les données primaires classiques - questionnaires visibles et systématisés à l'avance de manière à répondre aux exigences statistiques - et les données administratives informatisées.
b. Dans le traitement administratif des données informatisées, on peut difficilement faire une planification rigoureuse ou exercer une supervision soignée ou encore produire une documentation abondante puisque la condition préalable implicite pour toutes ces étapes - des délais suffisants -n'est pas satisfaite à cause de changements engendrés par des facteurs externes, par exemple:
   - modification à bref délai des lois et des règles administratives;
   - nécessité de modifier les systèmes de temps à autre pour obtenir une plus grande efficacité;
   - évolution rapide des moyens techniques -matériel et logiciel - qui force l'introduction de modifications fastidieuses;
c. Le niveau de tolérance pour les erreurs contenues dans les données extraites de systèmes administratifs informatisés n'est pas aussi faible qu'on aurait pu le prévoir. Les nombreux problèmes d'expérimentation liés à l'implantation de l'informatique dans le secteur public aussi bien que dans le secteur privé semblent faire croire au public et aux médias que les erreurs informatiques sont un élément de la vie courante avec lequel il faut apprendre à composer.

Étant donné l'importance des données administratives informatisées dans la production de statistiques, il est nécessaire d'élaborer une stratégie visant à réduire au minimum l'effet des inconvénients.

Une stratégie de ce genre n'a pas encore vu le jour au Danemark mais voici quelques éléments dont elle pourrait être constituée:

- vérification systématique (le plus tôt possible dans le processus) de la conformité des données avec la documentation reçue;
- établissement de procédures visant à remplacer le contrôle visuel du contenu des données par des totalisations de données clés tout le long du traitement;
- élargissement des méthodes de vérification dans le but précis de reconnaître plus rapidement les erreurs de systèmes et les autres erreurs fondamentales, cela pouvant se faire par exemple au moyen d'une comparaison systématique avec des données provenant de la même source mais ayant trait à des périodes antérieures ou avec des données correspondantes tirées d'autres sources.

4. QUALITÉ DES RÉSULTATS STATISTIQUES

Comme nous l'avons mentionné plus haut, le niveau de qualité final des statistiques fondées entièrement ou partiellement sur des données administratives dépend d'un certain nombre de facteurs, dont l'efficacité des mesures prises pour résoudre les problèmes exposés ci-dessus. Nous faisons ci-dessous quelques observations générales sur le sujet.

1) Même si quelques-uns des problèmes peuvent être fondamentaux, il est souvent possible de hausser le niveau de qualité une fois que les "difficultés de croissance" du système ont été surmontées.

2) Dans les secteurs considérés, les données administratives garantissent une **couverture** complète et éliminent par le fait même les erreurs d'échantillonage. On peut ainsi obtenir un plus haut niveau de définition que dans le cas des enquêtes par sondage et, dans la plupart des cas, les données administratives sont à l'abri des problèmes de non-réponse.

3) En ce qui concerne la **fiabilité** statistique et les caractéristiques générales des données administratives, il est nécessaire de distinguer divers genres de données.

   - En règle générale, les données des registres de base devraient être fiables, du moins les plus importantes ou les plus fréquemment utilisées. Comme nous l'avons déjà indiqué, ces données ne sont pas influencées par les différences de systèmes administratifs et sont ordinairement définies au point de vue opérationnel selon des principes statistiques.
   - Les données administratives qui servent de base aux décisions et aux calculs qui ont des conséquences administratives directes doivent avoir un degré de précision élevé mais ne sont pas pour autant exemptes de faiblesses. Par exemple, les données sur le revenu produites par l'administration fiscale peuvent ne concerner que les revenus qui ont été déclarés et évalués.
   - Les données administratives qui servent principalement de données de base peuvent être incomplètes ou de qualité inférieure.
   - Les données recueillies par l'intermédiaire de canaux administratifs mais à des fins statistiques précises peuvent aussi être d'une qualité inférieure peut-être parce que leur ordre de priorité dans le contexte administratif est peu élevé.

4) L'**actualité** des données varie. Elle dépend principalement de la formule de mise à jour, continue ou annuellement, des registres administratifs.. La mise à jour annuelle de données peut signifier un délai trop long pour les indicateurs à court terme. Par exemple, le délai de production des statistiques de la population active du Danemark fondées sur des registres se situe encore autour de dix-huit mois et peut, à la rigueur, être ramené à environ douze mois.

   Dans d'autres secteurs, les données sont produites dans des délais très satisfaisants. Par exemple, les données préliminaires du commerce extérieur du Danemark sont normalement disponibles trois semaines et demie après la fin du mois en question et les statistiques détaillées sur la population du Danemark sont produites à peine deux mois suivant la fin de l'année.

5) Nous avons déjà souligné que les changements administratifs ou législatifs pouvaient influer sur la **continuité** des données, par exemple les statistiques sur le revenu qui sont tributaires de la révision des lois fiscales. Cependant, la question de la continuité n'a pas vraiment posée de problème jusqu'à maintenant au Danemark.

Lorsque le pays est passé des statistiques classiques aux statistiques fondées sur des registres, le problème de la continuité s'est fait sentir de façon particulière mais, évidemment, cette situation n'a été que temporaire.

5. COMPARAISON DE LA QUALITÉ DES DONNÉES ADMINISTRATIVES ET DE LA QUALITÉ DES DONNÉES D'AUTRES SOURCES STATISTIQUES

La section précédente a montré clairement que l'utilisation statistique de données administratives soulevait de nombreux problèmes de qualité. Lorsque nous avons le choix entre des données administratives et d'autres genres de données, nous opterons pour les premières si notre principale préoccupation est de maximiser le rapport efficacité-coût et de réduire le fardeau de réponse.

Cependant, il peut nous arriver d'oublier que les autres sources statistiques soulèvent également d'énormes problèmes de qualité. Ainsi, de nombreuses études statistiques ont montré que les recensements classiques pouvaient comporter d'importantes erreurs de réponse ou de codage. De plus, les recensements dans certains pays sont exposés à des problèmes de sous-dénombrement.

La fiabilité statistique des enquêtes par sondage dépend de la taille de l'échantillon. Cette source probable d'erreur fait l'objet d'une théorie statistique très poussée et il est possible d'estimer les limites de ses effets probables.

Cependant, il n'existe pas de méthode qui nous permette de vérifier dans quelle mesure les écarts observés entre les résultats de deux enquêtes consécutives sont attribuables à la tendance fondamentale que nous voulons mesurer ou à l'erreur statistique.

En outre, le taux de non-réponse, et plus particulièrement les variations de ce taux, influent sensiblement sur la validité des résultats et il est évidemment très difficile de résoudre ce problème.

Il serait très difficile et coûteux de mesurer avec plus de précision les effets des sources d'erreur inhérentes aux diverses méthodes de calcul. Des comparaisons directes sont rarement réalisables puisqu'il est peu courant de produire des statistiques sur le même sujet à l'aide de méthodes différentes.

Au Danemark, il existe toutefois deux catégories de données sur la population active: la première repose sur le couplage d'enregistrements issus de divers registres administratifs publics (surtout ceux de l'administration fiscale) et la seconde, sur une enquête classique par interview téléphonique. La première catégorie de données constitue un dénombrement complet et les résultats peuvent être répartis selon une classification géographique détaillée. La seconde repose sur un échantillon de quelque 15,000 familles et porte sur environ 25,000 personnes de 16 à 74 ans.

Une étude comparative indique que le nombre de personnes occupées (groupes d'âges: 16-74 ans) selon l'enquête fondée sur des registres de novembre 1983 était de 2 489 000 alors que selon l'enquête par interview du printemps 1984, il était de 2 458 000.

Environ la moitié de cet écart de 1.3% s'explique par des facteurs saisonniers qui ont provoqué une baisse de l'emploi entre le quatrième trimestre de 1983 et le second trimestre de 1984. Comme cet écart est observable dans la plupart des autres groupes comparables, nous pouvons en conclure que les deux catégories de données reflètent l'une comme l'autre le niveau global de l'emploi au Danemark.

La comparaison est tout à fait différente lorsqu'il s'agit d'évaluer le nombre de personnes en chômage; les écarts sont en effet plus prononcés. Selon l'enquête par interview, il y avait 231 000 personnes en chômage au printemps de 1984; 196 000 d'entre elles étaient inscrites à des bureaux publics d'emploi et étaient par conséquent incluses

dans les statistiques courantes du chômage. Toutefois, ces dernières statistiques indiquent 243 000 personnes en chômage pour la même période. Environ la moitié de cet écart s'explique par un taux de réponse moins élevé pour les personnes en chômage que pour les personnes actives. Du reste, cela semble être une tendance générale au Danemark. Ainsi, une enquête de suivi menée auprès des personnes qui n'avaient pas répondu au questionnaire d'enquête en 1979 a permis de constater un taux de non-réponse de 40% parmi les personnes en chômage, comparativement à 23% pour la population en général.

6. UTILISATION DE DONNÉES ADMINISTRATIVES EN REMPLACEMENT DES DONNÉES DES RECENSEMENTS TRADITIONNELS

Le recensement est le secteur statistique qui a soulevé le plus d'intérêt au niveau international en ce qui a trait à l'utilisation de données administratives.

De toute évidence, les données de registres démographiques peuvent remplacer les données recueillies par recensement classique dans les pays où il existe de tels registres. De plus, dans les pays où il existe un système de numérotation unique à plus ou moins grande portée, on peut utiliser la méthode de "l'appariement exact" pour obtenir des données que l'on recueille normalement par recensement; ces données sont obtenues au moyen du couplage d'enregistrements. Au Danemark, le système des registres a pris une telle ampleur qu'il est maintenant possible d'en extraire toutes les données que l'on recueille normalement par recensement. Toutefois, les données administratives ne suffisent pas. Ainsi, pour connaître le niveau d'instruction de la population avant 1970, on se reporte au dernier recensement classique de la population, réalisé cette année-là. Les données sur la relation entre le domicile et le lieu de travail des personnes occupées (pour les statistiques sur les migrations journalières) sont fournies en partie par Danmarks Statistik et en partie par l'administration fiscale, qui se sert d'un prolongement statistique d'un de ses systèmes, notamment le projet du lieu de travail, évoqué plus haut.

Pour ce secteur et d'autres, Danmarks Statistik collabore avec les autorités administratives selon le principe de la "circulation à sens unique". Cela semble acceptable aux yeux du public en général.

Comme les données primaires contenues dans les registres sont mises à jour annuellement, il serait possible en principe de réaliser des recensements complets à intervalles rapprochés mais en pratique, cela n'est pas réalisable. Chaque année, une partie du contenu des registres est publiée avec les statistiques courantes des divers secteurs. D'autres données - par exemple, les statistiques relatives aux sous-districts des municipalités - sont diffusées par l'intermédiaire de systèmes de service spéciaux adaptés aux besoins de groupes d'utilisateurs et elles peuvent servir à des totalisations ad hoc définies par les utilisateurs.

Ces formes de diffusion ont réussi à satisfaire une telle proportion des demandes de statistiques de recensement que l'on a pu annuler un recensement fondé sur des registres qui était prévu pour 1986 sans provoquer de réactions majeures chez les utilisateurs. Cette annulation s'inscrivait dans le cadre d'une politique de réduction du programme de la statistique du Danemark, rendue nécessaire par de fortes compressions budgétaires. Dans la plupart des autres secteurs, la politique de réduction a été mal accueillie par les utilisateurs.

À la suite de ces observations, il est permis de se demander si les recensements classiques répondent encore aux besoins d'aujourd'hui. Dans certains pays, les recensements suscitent une vive hostilité dans le public, leurs résultats ne sont pas publiés dans des délais raisonnables et il est difficile d'en assurer le déroulement harmonieux puisqu'à chaque fois, il faut mettre sur pied une nouvelle superstructure. Par surcroît, les recensements classiques présentent des lacunes sur le plan de la qualité.

Devant ces arguments, on pourrait prétexter que la plupart des pays n'ont pas de solution de rechange pour les recensements classiques et qu'il leur est tout simplement impossible de trouver d'autres façons de produire des statistiques sur la taille de la population et sa structure démographique, géographique et sociale.

Or, dans le seul secteur statistique qui peut être comparé au secteur du recensement (celui de la comptabilité nationale), les conditions sont aussi sévères puisqu'il est impossible de réaliser des "recensements des comptes nationaux". Néanmoins, on produit des statistiques de la comptabilité nationale en regroupant des données de diverses sources selon des méthodes complexes d'imputation, d'estimation et de compensation.

La question est de savoir si l'on peut procéder de la même façon dans le secteur du recensement. Les problèmes notés ci-dessus sont traités dans divers articles de Philip Redfern et ont aussi fait l'objet de discussions aux conférences de l'IIS; malheureusement, à la conférence de 1987 tenue à Tokyo, les membres n'ont pas accordé beaucoup d'attention à ces problèmes. Il va sans dire que le Danemark, avec ses nombreuses possibilités d'utilisation de données administratives, ne peut faire avancer beaucoup plus la recherche dans ce secteur mais il convient de souligner que la majeure partie des données sur lesquelles reposent les statistiques de recensement du Danemark sont issues des systèmes de l'administration fiscale et que dans la plupart des pays, des systèmes semblables renferment des données qui sont de même nature que les données de recensement.

S'il était possible d'extraire, de systématiser et d'utiliser ces données à des fins statistiques, la plupart des pays industrialisés pourraient vraisemblablement créer une base de données semblables aux données de recensement et y ajouteraient des données d'enquête par sondage afin d'en faire une base acceptable de données de référence sur les personnes. La recherche qui se fait présentement au bureau central de statistiques de Norvège concernant l'intégration de données de registres et de données d'enquête en vue du recensement de 1990 pourrait avoir une importance capitale sur ce plan.

À ce propos, il convient de souligner que l'existence de clés d'identification spécifiques (par exemple, code de la personne) n'est pas une condition préalable pour le couplage de données de sources différentes. D'autres formes de couplage d'enregistrements ("appariement statistique", etc.) peuvent être utiles dans beaucoup de cas.

Il semblerait donc que certains pays n'ont pas écarté cette solution pour venir à bout de leurs problèmes de recensement mais les recherches seront longues et ardues et les organismes statistiques centraux pourraient ne pas être en mesure de mener à terme cette mission sans l'aide des autorités politiques et administratives.

7. CONCLUSION

L'utilisation des registres administratifs comme principale source de données présente des avantages considérables mais soulève aussi d'énormes problèmes.

Des recherches méthodologiques sérieuses devront être entreprises non seulement dans des pays comme le Danemark, où existent déjà les conditions nécessaires à l'utilisation de données administratives, mais aussi dans les pays où il n'existe pas encore de telles conditions mais où le souci de réduire le fardeau de réponse et de trouver une solution de rechange aux méthodes classiques, de moins en moins efficaces, peut faire ressortir la nécessité de recourir aux fichiers administratifs comme principale source de statistiques. À notre avis, ces questions poseront un défi de taille aux organismes statistiques centraux d'ici la fin du siècle.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

ÉVALUATION DE L'EFFET DE LA RÉFORME FISCALE SUR LES PROGRAMMES DU CENSUS BUREAU

GERALD GATES[1]

RÉSUMÉ

La législation sur la réforme fiscale de 1986 a entrainé des changements spectaculaires dans le système de déclaration fiscale aux États-Unis. Un grand nombre des formulaires utilisés pour la déclaration du revenu et des dépenses au "Internal Revenue Service" sont en cours de révision; des procédures sont en voie de préparation pour l'interprétation de la nouvelle législation et pour guider les contribuables. Enfin, le système de traitement fiscal est en cours de remaniement afin de prendre en compte tous ces changements. Puisque le Census Bureau utilise des données administratives provenant des fichiers fiscaux des programmes du recensement économique et des estimations démographiques et de différents programmes d'évaluation de la couverture et du contenu, nous sommes directement visés par tout changement dans la façon dont les impôts sont recueillis et traités. De plus, il y a des conséquences indirectes à la suite des modifications correspondantes apportées aux pratiques de tenue des dossiers des entreprises et des particuliers qui font une déclaration au Census Bureau lors de nos enquêtes et de nos recensements. Ce document décrit comment le Census Bureau compte s'adapter à l'adoption de la législation fiscale, et quelle mesure il sera contraint de prendre pour minimiser les perturbations causées par la perte d'information, les changements de définition et les retards de collecte et de traitement.

La réforme fiscale de 1986 a non seulement profondément transformé le régime d'imposition des États-Unis mais elle a aussi créé des défis de taille et des possibilités intéressantes pour les statisticiens qui utilisent les données administratives des fichiers fiscaux. Elle a réduit de façon globale les taux d'impôt pour les particuliers comme pour les entreprises, elle a transféré une partie du fardeau fiscal des particuliers aux sociétés, elle a réduit l'échelle des taux d'impôt progressif, mais elle a maintenu les recettes courantes en supprimant les privilèges fiscaux dont bénéficiaient les contribuables à revenu élevé et en réduisant la charge fiscale des contribuables à plus faible revenu. Elle a aussi contribué à élargir l'assiette de l'impôt en limitant ou en éliminant les divers crédits, déductions ou traitements spéciaux qui visaient à encourager certains types d'investisements. Par suite de ces changements importants, l'IRS (Internal Revenue Service/service du revenu interne) a révisé bon nombre des formules utilisées pour la déclaration d'impôt. En outre, les règlements et les procédures sont en train d'être modifiés, le système de traitement est remanié en fonction des nouvelles formules et de

[1] Gerald W. Gates, Program and Policy Development Office, U.S. Bureau of the Census, Washington, D.C. 20233, É.-U.

celles qui ont été révisées, et une vaste campagne de publicité se déroule actuellement pour informer les contribuables sur les dispositions de la nouvelle loi.

Les avocats et les conseillers en matière fiscale sont aussi prêts à aider leurs clients à tirer profit de la nouvelle loi. Des décisions financières seront vraisemblablement prises en fonction des nouvelles dispositions régissant le taux d'impôt sur le revenu et les déductions. Il y a aussi la question de la mise en vigueur progressive de la nouvelle loi. Les principales dispositions de la loi seront appliquées au cours de l'année d'imposition 1987, mais les taux d'impôt diminueront davantage en 1988 et plusieurs crédits et déductions seront réduits ou supprimés sur une période de deux à cinq ans. "Deux des mesures les plus importantes prises par les contribuables par suite de l'adoption de cette loi est le report du revenu et l'anticipation des déductions. Le fait que la réduction des taux d'impôt sera étalée sur une période de deux ans et que de nombreux éléments déductibles seront limités ou supprimés incitera une grande partie des contribuables à reporter leur revenu et (ou) à anticiper leurs déductions de façon à minimiser les impôts qu'ils auront à payer en 1986 et en 1987" (Wakefield 1987).

Dans ce contexte, le Census Bureau doit déterminer dans quelle mesure les fichiers de données établis à partir des renseignements fournis par l'IRS seront touchés et comment traiter efficacement les nouvelles données. Nous sommes autorisés à recueillir ces données en vertu des intitulés 26 et 13 du United States Code qui définissent les fonctions de l'IRS et du Census Bureau respectivement. Le paragraphe 6103(j) de l'intitulé 26 permet au Census Bureau d'avoir accès "... aux déclarations ou aux données des déclarations (...) conformément au règlement (...) dans le but d'établir la structure des recensements et des comptes économiques nationaux, et de mener les activités statistiques connexes autorisées par la loi, mais seulement dans la mesure nécessaire à la réalisation de ces activités." (traduction) De même, le paragraphe I.6a) de l'intitulé 13 autorise le Census Bureau à "...s'adresser à tout autre département, organisme ou établissement fédéral ... en vue d'obtenir les renseignements nécessaires aux activités précisées dans le présent intitulé." (traduction) Pour toute nouvelle application des données fiscales qui n'est pas prévue par la loi, le Secrétaire au Commerce (pour le Census Bureau) doit adresser une demande au Secrétaire au Trésor (pour l'IRS), et il faut une modification au Code of Federal Regulations (CRF) indiquant la nature et l'utilisation prévue des données requises.

La suite du présent document porte sur l'utilisation que le Census Bureau fait des données fiscales dans ses programmes de statistique économique et démographique et sur l'incidence de la réforme fiscale sur cette utilisation. Une évaluation de cette incidence est en cours et a fourni des résultats sur les données qui sont ajoutées, perdues, éventuellement retardées ou traitées différemment. Ce document examine aussi les réactions possibles des contribuables, qui pourront modifier les habitudes de déclaration et, conséquemment, influer sur les données qui nous sont fournies. Enfin, le document traite des conflits qui opposent le mandat de l'organisme statistique et celui de l'organisme administratif et des mesures prises par le Census Bureau pour participer plus activement à la collecte des données administratives et assurer que ces données soient toujours utiles.

1. UTILISATION DES DONNÉES FISCALES PAR LE CENSUS BUREAU

Depuis 1954, le Census Bureau reçoit de l'IRS des fichiers fiscaux qui sont utilisés pour effectuer ses recensements et ses enquêtes économiques et agricoles. Nous avons utilisé ces données plutôt que de procéder par enquête directe pour obtenir des données auprès de nombreuses petites entreprises. Ces données ont aussi servi, entre autres choses, à déterminer les nouvelles entreprises, à mettre à jour et à compléter les listes postales d'entreprises et d'exploitations agricoles et à classer certaines entreprises dans une activité économique.

Pour les recensements économiques de 1982, tous les renseignements sur approximativement 1.3 million des 3.5 millions d'entreprises à établissement unique visées provenaient des déclarations d'impôt. En outre, les fichiers fiscaux ont fourni les données relatives à l'ensemble des 4.7 millions de sociétés sans employés rémunérés. Bien que ces données fiscales ne correspondent qu'à 5 à 10 pour cent de l'ensemble des rentrées de fonds (toutes les entreprises à établissements multiples ainsi que les grandes entreprises à établissement unique reçoivent un questionnaire du recensement par la poste), elles représentent une réduction considérable du fardeau de réponse des petites entreprises.

Pour les recensements économiques de 1987, qui constituent le prochain cycle de notre programme quinquennal de recensements économiques, le Census Bureau tirera ses données de quatorze formules d'impôt de l'IRS qui ont rapport aux entreprises (tableau 1). Ces fichiers fiscaux fournissent des données sur les rentrées de fonds pour toutes les entreprises auxquelles on ne fait pas parvenir un questionnaire du recensement par la poste. Nous les utiliserons pour définir l'univers global des sociétés sans employé rémunérés visées par les recensements économiques (c.-à-d. les entreprises qui ont des données sur les rentrées de fonds, mais pas de feuille de paye correspondante). Nous obtiendrons les adresses pour faire le codage géographique et déterminer l'univers initial des enquêtes sur les entreprises qui appartiennent à des groupes minoritaires ou à des femmes. Enfin, nous utiliserons des renseignements tels que l'indicateur de fin d'année et le nombre de mois d'exploitation lors de la sélection et de la vérification des échantillons pour les diverses enquêtes en cours. Dans l'ensemble, les données de plus de 75 millions de dossiers d'impôt sur les entreprises de 1987 seront utilisées pour planifier les recensements économiques de 1987.

En plus de ces utilisations pour nos programmes économiques, les données des déclarations des particuliers sont utilisées dans nos programmes démographiques pour établir des estimations démographiques intercensitaires et pour effectuer des recherches connexes, pour faire des estimations bisannuelles du revenu par personne, pour évaluer la qualité de la couverture des recensements décennaux et pour effectuer des recherches statistiques et mener des projets de développement concernant les enquêtes actuelles du Census Bureau (par exemple l'évaluation de la qualité des données sur le revenu déclaré lors de l'enquête sur la population). L'utilisation de ces formules de déclaration d'impôt sur le revenu des particuliers (formules 1040) pour les questions démographiques et pour les activités de recherche et d'évaluation associées aux recensements décennaux remonte aux années 60. Des échantillons de personnes dont le nom figure sur une déclaration d'impôt sur le revenu des particuliers ont été comparés aux données des dossiers des recensements décennaux afin d'évaluer la qualité de la couverture. Bien que le programme d'évaluation de la couverture de 1990 n'ait pas été pleinement élaboré, il est fort possible qu'il prévoie des appariements de dossiers de l'IRS et du Census Bureau.

Nous utilisons les données des déclarations des particuliers pour établir nos estimations démographiques depuis 1972. Pour les besoins du programme des estimations démographiques, nous utilisons les déclarations d'impôt simultanées pour produire des matrices des migrations représentant les changements démographiques pour des secteurs gouvernementaux donnés d'une année à une autre. Ces estimations de l'évolution démographique sont utilisées chaque année avec les registres des naissances et des décès afin de mettre à jour les estimations démographiques pour les administrations des États, des comtés et des municipalités.

Les estimations bisannuelles du revenu par personne au niveau des États, des comtés et des municipalités sont établies en même temps que les estimations de la population. Les estimations du revenu par personne sont calculées à l'aide de cinq postes de la déclaration d'impôt: revenu brut rajusté, dividendes, intérêts, salaires et traitements, redevances et loyers bruts. Les estimations des différences dans le revenu de deux années d'imposition consécutives servent à mettre à jour les estimations du revenu par personne.

Tableau 1

Déclarations d'impôt sur le revenu des entreprises en 1987 à utiliser pour les recensements économiques de 1987

Formule de l'IRS	Titre
SS-4	Application for Employer Identification Number (Demande de numéro d'identification de l'employeur)
941	Employer's Quarterly Federal Tax Return (Déclaration d'impôt fédérale et trimestrielle de l'employeur)
990	Return of Organization Exempt from Income Tax (Déclaration d'un organisme dispensé d'impôts sur le revenu)
990-PF	Return of Private Foundation (Déclaration d'une fondation privée)
990-T	Exempt Organization Business Income Return (Déclaration de revenu d'entreprise des organismes dispensés)
1040 Annexe C	Profit or Loss from Business or Profession (Bénéfices ou pertes provenant de l'exploitation d'une entreprise ou de l'exercice d'une profession)
1040 Annexe SE	Computation of Social Security Self Employment Tax (Calcul de l'impôt de sécurité sociale - Travail indépendant)
1065	U.S. Partnership Return of Income (Déclaration du revenu d'une société américaine de personnes)
1065 Annexe K1	Partner's Share of Income, Credits, Deductions, etc. (Part du revenu, des crédits, des déductions, etc., de l'associé)
1120	U.S. Corporation Income Tax Return (Déclaration d'impôt sur le revenu des sociétés américaines)
1120-A	U.S. Short Form Corporation Income Tax Return (Déclaration d'impôt sur le revenu des sociétés américaines - Formule abrégée)
1120-F	Partner's Share of Income Tax Return of a Foreign Corporation (Déclaration d'impôt sur le revenu d'une société étrangère aux É.-U.)
1120-S	U.S. Income Tax Return for a S-Corporation (Déclaration d'impôt sur le revenu d'une petite société aux É.-U.)
1120-S Annexe K1	Shareholder's Share of Income, Credits, Deductions, etc. (Part de revenu, de crédits, de déductions, etc., de l'actionnaire)

2. PROBLÈMES RÉSULTANT DE L'UTILISATION DES DONNÉES FISCALES

Depuis que nous avons commencé à utiliser les données de l'IRS, il est arrivé que les fichiers que l'IRS nous a fait parvenir n'aient pas répondu à nos attentes par suite d'un malentendu, de communications insuffisantes ou de méthodes de vérification inadéquates. Ainsi, lors des recensements économiques de 1982, le Census Bureau a demandé à l'IRS d'ajouter plusieurs questions aux formules d'impôt des entreprises. Une de ces questions n'a jamais été ajoutée aux formules. De plus, le Census Bureau voulait avoir accès aux données fiscales de 1982 et avait demandé que le règlement soit modifié en conséquence, mais on n'a donné suite à cette demande que presque trois ans plus tard à cause d'une mauvaise coordination des efforts. Enfin, le codage des activités économiques laisse beaucoup à désirer parce que les centres de traitement régionaux ont coupé au plus court afin de satisfaire à des quotas de production rigoureux. Toutefois, ces problèmes ne

constituaient pas la norme et n'ont pas eu une forte incidence sur le programme des recensements économiques de 1982. Néanmoins, la leçon tirée de cette expérience s'est avérée coûteuse. L'IRS et le Census Bureau partagent la responsabilité de ces erreurs et s'engagent à assurer un déroulement plus efficace des recensements économiques de 1987. À cette fin, des normes relatives à la qualité des données et à l'établissement des délais ont été établies, des réunions ont lieu régulièrement pour faire le compte rendu de la situation et un système automatisé de vérification a été élaboré pour suivre le déroulement des activités et fournir des rapports réguliers sur l'état d'avancement (Jonas, Hanczaryk 1987). L'importante révision du système fiscal, qui entre en vigueur au cours de l'année d'imposition 1987, rend ce défi encore plus intéressant à relever.

3. RÉSULTATS DE NOTRE ÉVALUATION

Étant donné que le Census Bureau compte sur les données fiscales, que les recensements économiques et le recensement décennal sont proches et que l'utilisation des dossiers administratifs occasionne des problèmes particuliers, l'évaluation des effets de la législation sur la réforme fiscale de 1986 n'en est que plus pressante. Pour avoir la situation en main et assurer la coordination des efforts, le Program and Policy Development Office (bureau de l'élaboration des programmes et politiques) du Census Bureau a commencé, vers la fin de 1986, à étudier l'incidence de la loi sur la réforme fiscale de 1986 sur les programmes du Census Bureau. Dans le cadre de cette étude, le PPDO a pris connaissance de la législation et de ses effets sur les programmes du Census Bureau, discuté de ces effets avec la SOI (Statistics of Income Division/Divisoin des statistiques du revenu) de l'IRS et les divisions du Census Bureau qui utilisent ces données, vérifié les changements apportés aux procédures et aux formules d'impôt, examiné l'incidence sur le système de traitement en fonction du contenu et de la qualité des fichiers transmis envoyés au Census Bureau, et fait le suivi des opérations de traitements afin de déterminer les arriérés de travail. Cette étude est effectuée en collaboration avec la Division SOI, qui répond à nos questions sur la mise en application de la réforme fiscale et nous fournit des ébauches de toutes les formules et rapports de traitement courants. Comme mesure d'appui, nous passons en revue tous les avis du **Federal Register** afin de relever les demandes faites par l'IRS à l'Office of Management and Budget (bureau de la gestion et du budget) pour corriger les nouvelles formules d'impôt pour 1987 et celles qui ont été révisées de même que les règlements et procédures concernant l'application de la nouvelle loi. En outre, nous avons parcouru les renseignements publiés sur la loi sur la réforme fiscale de 1986 afin de déceler tout lien qui pourrait exister avec les programmes du Census Bureau et ses besoins futurs. Compte tenu de nos efforts et des résultats de nos programmes permanents de contrôle de la qualité, nous sommes de plus en plus convaincus que les erreurs de 1982 ne se répèteront pas et que nous serons en mesure de tirer pleinement avantage des possibilités qu'offre la loi sur la réforme fiscale de 1986 sur le plan de la statistique tout en minimisant les effets négatifs que cette loi pourrait avoir sur nos programmes.

Le processus d'évaluation des effets de la réforme fiscale est permanent. L'élaboration des formules d'impôt de 1987 est achevée ou en voie de l'être, et le système de traitement est modifié de façon à tenir compte des changements. Nous avons déterminé les avantages et les inconvénients que la réforme présenterait pour nos programmes et tentons actuellement d'évaluer les conséquences moins évidentes résultant des modifications apportées aux pratiques de tenue des dossiers et des réactions des contribuables à tous ces changements.

3.1 Nouvelles données

En ce qui a trait aux avantages, la loi sur la réforme fiscale de 1986 offre plusieurs possibilités intéressantes pour le Census Bureau. Par exemple, l'article 1524 stipule que les contribuables doivent maintenant obtenir et déclarer un numéro de sécurité sociale pour chaque enfant à charge de cinq ans et plus. Ce genre de renseignement serait utile pour toute évaluation du champ d'observation du recensement décennal comportant l'appariement direct des dossiers de l'IRS et du Census Bureau. Plus précisément, il est fort probable que l'on retrouve les adolescents sur plus d'une déclaration (la leur et celle de leurs parents). La déclaration du numéro de sécurité sociale de ce groupe de contribuables permettrait au Census Bureau de relever avant l'appariement les cas de doubles comptes.

Un autre avantage de la réforme fiscale est l'exigence prévue à l'article 1521 de la loi et selon laquelle les produits bruts des transactions immobilières doivent être déclarés à l'IRS. Une des parties contractantes doit fournir sur une nouvelle formule (1099-S) une description du bien immobilier (adresse ou description conforme à la loi), le produit brut de la transaction moins la valeur du bien ou des services obtenus, et la date de la transaction.

Dans le cadre de chaque recensement quinquennal des gouvernements, le Census Bureau a mené une TPVS ou Taxable Property Values Survey (enquête sur les valeurs immobilières imposables). Ces enquêtes sont effectuées depuis 1957 et ont deux composantes principales. La première est une étude du rapport évaluation/prix de vente, qui est la seule source de données sur les niveaux réels d'évaluation de l'impôt foncier à l'échelle nationale. La deuxième composante est une estimation de l'importance et de la composition de l'assiette de l'impôt sur les biens immeubles d'après la valeur imposable et le nombre de parcelles de terrain.

En 1987, les restrictions budgétaires nous ont obligés à laisser tomber l'étude des rapports et à nous concentrer uniquement sur la composition de l'assiette fiscale. Dans ce contexte, les nouvelles exigences de l'IRS concernant la déclaration du prix de vente des biens immobiliers présentent des possibilités intéressantes. Après quelques modifications, par exemple l'inclusion d'une exigence stipulant que la valeur imposable et l'utilisation actuelle du même bien soient aussi déclarées, la formule 1099-S pourrait bien devenir un outil de collecte des données nécessaires à l'établissement du rapport évaluation/prix de vente. On conçoit que cette nouvelle formule pourrait faciliter les activités de recouvrement de l'impôt de l'IRS et servir à de nombreuses autres fins. Le Census Bureau pourrait utiliser les données recueillies pour effectuer des études nationales des rapports évaluation/prix de vente chaque année plutôt qu'à tous les cinq ans. De même, les nombreux États qui font leurs propres études de ce rapport pourraient éventuellement utiliser les renseignements des formules 1099-S et ne seraient plus obligés de faire leur propre collecte de données. À l'heure actuelle, nous tentons de déterminer avec l'IRS s'il serait souhaitable et possible de modifier la formule 1099-S de façon à accroître son utilité sur le plan administratif et statistique.

Enfin, "...les contribuables devront (maintenant) déclarer tous les intérêts exempts d'impôt reçus. Par suite de cette stipulation et d'autres dispositions, il sera possible d'établir des estimations plus fidèles des concepts économiques types du revenu en se fondant uniquement sur les renseignements déclarés sur les déclarations d'impôt" (Jabine 1987). Comme on l'a vu précédemment, le Census Bureau reçoit de l'IRS des données sur cinq éléments de revenu dont il se sert pour faire des études d'évaluation et pour produire des estimations du revenu par personne. Si les données sur les intérêts fournies par l'IRS englobaient aussi les intérêts non imposables, il serait possible d'établir de meilleures comparaisons entre le revenu en intérêts déclaré lors de l'enquête supplémentaire de mars se rattachant à la Current Population Survey (enquête sur la population actuelle) et le revenu déclaré dans le cadre de la SIPP (Survey of Income and Program

Participation/enquête sur le revenu et la participation aux programmes), et les résultats obtenus présenteraient des avantages à long terme pour l'établissement des estimations du revenu par personne.

3.2 Les données que nous n'obtiendrons plus

Pour ce qui est des inconvénients, nous savons qu'à cause de la réforme fiscale certains renseignements que nous utilisons pour nos programmes de statistiques démographiques ne seront plus recueillis. Ainsi, on ne tient pas compte des dossiers d'impôt pour les personnes de 65 ans et plus lors de l'établissement de nos estimations démographiques au niveau des comtés parce que la couverture des fichiers fiscaux pour ce groupe d'âge est jugée insatisfaisante du fait qu'un trop grand nombre de personnes de 65 ans et plus ne produisent pas de déclaration. D'autres sources, par exemple les fichiers de Medicare, donnent une meilleure couverture. Étant donné qu'il n'est pas nécessaire d'indiquer l'âge sur la déclaration d'impôt, il n'est pas possible d'identifier directement les gens de ce groupe d'âge en vue de les exclure de nos estimations. Toutefois, avant la loi sur la réforme fiscale de 1986, les personnes de 65 ans et plus avaient droit à une exemption personnelle supplémentaire. Par conséquent, nous pouvons déterminer si le contribuable a plus de 65 ans en nous fondant sur la déclaration d'une telle exemption. Avec la réforme fiscale, cette exemption est supprimée et remplacée par une déduction forfaitaire supplémentaire ne s'appliquant qu'aux déclarants de 65 ans et plus qui ne produisent pas une déclaration détaillée. Dans le cas de ceux qui produisent une déclaration détaillée, il n'est pas possible de déterminer l'âge directement. Pour ce groupe, la déclaration d'un revenu de sécurité sociale sur la formule d'impôt peut servir de variable corrélée pour l'attribution de l'âge. Toutefois, cette donnée permet seulement de déterminer que le contribuable a moins de 65 ans, puisque des personnes de moins de 62 ans peuvent toucher un revenu de sécurité sociale si elles sont invalides, tout comme des personnes de 62 à 64 ans. Les renseignements fournis à l'annexe R (crédit pour les personnes âgées) peuvent aussi être utilisés à cette fin.

Les résultats de recherches préliminaires effectuées par le Census Bureau à l'aide du fichier-échantillon de 1984 de la Division SOI ont indiqué qu'il était possible d'identifier 64 pour cent des personnes de 65 ans et plus en se fondant sur la déclaration d'une exemption forfaitaire supplémentaire pour personnes âgées. En établissant aussi que si les déclarants principaux et secondaires reçoivent un revenu de sécurité sociale, ils ont 65 ans et plus, nous pourrions repérer une proportion additionnelle de 29 pour cent, mais aussi compter par erreur une proportion supplémentaire de 17 pour cent. Par ailleurs, si nous déterminons que seuls les déclarants principaux qui reçoivent un revenu de sécurité sociale ont 65 ans et plus, nous obtenons une proportion additionnelle de 21 pour cent et une proportion erronée de 8 pour cent. L'annexe R permet seulement d'identifier moins de 1 pour cent des personnes de 65 ans et plus. Les recherches se poursuivent sur ces méthodes de détermination de l'âge et sur d'autres facteurs qui permettraient de compenser les données sur l'âge perdues à cause de la réforme fiscale (Sater 1987).

Nous subirons une perte d'information encore plus importante du fait que de nombreux contribuables ne sont plus tenus de payer des impôts par suite de l'augmentation du nombre d'exemptions personnelles et du montant des déductions forfaitaires. Selon les calculs, presque 5 millions de personnes à faible revenu ne figureront plus sur le rôle d'imposition (Pechman 1987). L'élimination de ces personnes à faible revenu pourrait influer sur les données relatives aux migrations établies dans le cadre du programme des estimations démographiques. Nous n'ignorons pas que l'exclusion des non-déclarants a une certaine incidence sur nos données relatives aux migrations. Si l'on suppose que les personnes à faible revenu ne se déplacent pas autant que les personnes ayant un revenu plus élevé, il est possible que les données sur les migrations établies à partir des déclarations d'impôt ne donnent pas un juste aperçu des mouvements migratoires réels de l'ensemble de la population. Par exemple, nous pourrions surestimer l'importance des

migrations externes pour certaines grandes villes. Une augmentation considérable du nombre de non-déclarants résultant de la réforme fiscale aggraverait le problème.

D'après les résultats d'une étude effectuée par le Census Bureau à l'aide de trois algorithmes de simulation se rapportant aux déclarants et un fichier de données sur le revenu rajustées de 1984 de la Division SOI, nous avons déterminé que 4 ou 5 pour cent au plus des déclarations produites pour l'année d'imposition 1986 ne seront pas produites pour l'année d'imposition 1987 du seul fait des changements apportés aux exigences relatives à la production des déclarations (Valdisera 1987). En outre, ces non-déclarants éventuels ne représentent que 2 ou 3 pour cent de l'ensemble des personnes (déclarants et personnes à charge) qui composent l'univers des déclarants de 1986. Par ailleurs, nous prévoyons que les estimations de non-déclarants varieront considérablement selon la région géographique et le groupe démographique.

Nous savons par expérience que de nombreux non-déclarants éventuels continueront de produire une déclaration par choix, par habitude, parce qu'ils ne seront pas au courant des nouvelles exigences de déclaration ou auront négligé d'aviser leur employeur de cesser de faire des retenues, et nous avons tenu compte de ces facteurs dans nos projections. Nous poursuivons notre étude de cette question afin d'améliorer nos estimations du nombre de non-déclarants éventuels. À mesure que ces estimations augmenteront, nous tenterons de trouver une autre source de renseignements sur les migrations des défavorisés.

3.3 Autres facteurs

Outre l'ajout ou la perte de données, plusieurs facteurs influent sur la qualité et l'actualité des renseignements déclarés et traités. Par exemple:

- Comme dans le cas de n'importe quel remaniement important d'un système de traitement, les changements d'envergure à apporter aux formules d'impôt ne peuvent pas toujours se faire dans les délais prescrits. Les nouvelles versions des programmes peuvent ne pas être terminées à temps ou avoir des défauts.

- L'envoi des fichiers au Census Bureau peut être retardé par le codage manuel et l'introduction par clavier des arriérés de travail attribuables aux nouvelles procédures.

- Les renseignements recueillis sur l'ensemble des formules d'impôt ne sont pas tous codés. Il est essentiel de savoir quelles données ne sont que partiellement codées avant de pouvoir utiliser les fichiers fiscaux.

- L'importance de la vérification du codage et de l'édition des données varie selon l'utilisation adminitrative de ces données. La qualité des données et leur utilité à des fins statistiques varieront donc aussi.

- De façon générale, il est possible que les contribuables produisent leur déclaration en retard, ce qui donnerait lieu à un volume important de formules qui ne pourraient pas être traitées assez rapidement avec les ressources disponibles.

Certains indices nous portent à croire que beaucoup de contribuables produiront leur déclaration pour 1987 en retard parce qu'ils auront mal rempli ou rempli tardivement la formule W4, Employee's Withholding Allowance Certificate (certificat de retenue d'indemnité de l'employé). Les formules W4, qui doivent être produites avant les déclarations, ont tendance à présenter des retenues insuffisantes des gains à cause des rajustements apportés aux tables des taux d'impôt en janvier 1987. Les contribuables devaient remplir les nouvelles formules W4 pour expliquer les changements relatifs aux déductions permises. Toutefois, comme cette exigence a créé une certaine confusion et

que la formule était assez complexe, l'IRS a établi une version révisée, prolongé le délai accordé pour remplir la formule et, dans certains cas, laissé tomber les pénalités prévues. Selon l'importance du retard dans l'envoi des W4, bon nombre de contribuables pourraient avoir à payer plus d'impôt. De nombreux déclarants qui remplissent habituellement leur déclaration en janvier ou en février pourraient repousser cette tâche au mois d'avril.

Comme il importe que les données fiscales nous parviennent à temps pour la réalisation de notre programme d'estimations démographiques et la tenue de nos recensements économiques, nous devrons surveiller de près la charge de travail afin de déceler les points de congestion possibles. Nous avons demandé à l'IRS de nous fournir des copies de leurs rapports hebdomadaires sur les rentrées et le traitement. En comparant ces rapports avec ceux de l'année précédente, nous devrions pouvoir cerner les problèmes et prendre les mesures qui s'imposent.

Il pourrait aussi y avoir des problèmes concernant la qualité des données reçues ou traitées. Par exemple, la confusion créée par la réforme fiscale peut donner lieu à des problèmes de déclaration, du moins au début. Par ailleurs, on pourrait constater une amélioration du taux de déclaration par suite des nouvelles vérifications faites pour s'assurer que les déclarants se conforment aux nouvelles exigences. Nous espérons pouvoir relever, du moins dans le cas des données des déclarations d'impôt des entreprises, tout changement et problème éventuel à l'aide de notre programme de contrôle qualitatif qui permet de comparer les habitudes réelles de déclaration avec les normes de déclaration établies.

D'autres changements peuvent aussi influer plus ou moins sur les données: 1) les sociétés de personnes et les "S" corporations ("small" corporations/les petites sociétés)* doivent maintenant déclarer leurs gains et leurs frais suivant l'année civile et non l'exercice financier, ce qui signifie une meilleure uniformité des déclarations de ce genre d'entreprises; et 2) les contribuables peuvent choisir de dissoudre leur société ou de la transformer en petite société afin de tirer profit des taux plus bas pour les particuliers (taux maximal de 28 pour cent pour les particuliers contre un taux maximal de 34 pour cent pour les sociétés). Dans ce dernier cas, les contribuables qui décident de transformer leur société en petite société peuvent demander de nouveaux EIN (Employeur Identification Numbers/numéros d'identification de l'employeur), bien qu'ils ne soient pas tenus de le faire. De telles initiatives pourraient fausser les données sur les nouvelles entreprises ou celles qui n'existent plus et créer un décalage dans les listes des entreprises qui sont effectivement nouvelles et qui sont requises pour l'échantillonnage. Même si les entreprises ne demandent pas de nouveaux EIN, ces décalages, s'ils sont importants, entraîneront une augmentation considérable du nombre de déclarations portant seulement sur une partie de l'année, ce qui compliquerait le traitement des données au Census Bureau.

Toutefois, la décision de transformer une société en petite société donne lieu à des complications pour les raisons suivantes: 1) le taux marginal d'impôt, de 33 pour cent, prévu pour les particuliers dont le revenu imposable se situe entre $71,900 et $149,250; 2) les limites imposées aux petites sociétés en ce qui concerne les emprunts aux régimes de pension; 3) les restrictions imposées aux petites sociétés au sujet de l'émission d'actions; et 4) l'exigence selon laquelle les petites sociétés doivent produire leur déclaration suivant l'année civile. "Le statut de petite société demeure peut-être le meilleur choix pour certains types d'entreprises. Mais ce ne sera pas une solution facile" (Quinn 1987).

---

* Une S-corporation ou petite société est une forme juridique d'organisation créée pour les "petites" entreprises qui permet à celles-ci de profiter de l'allégement fiscal dont bénéficient les entreprises non constituées en société par actions tout en ayant les avantages qu'offre la constitution en société par actions. Pour avoir le statut de petite société, une entreprise doit compter moins de 35 actionnaires et ces derniers doivent tous opter pour cette forme juridique.

En 1985, il y avait 735,000 petites sociétés et auparavant, leur nombre augmentait à un rythme d'environ 30,000 par année. Pour l'année d'imposition 1986, nous prévoyons qu'il y aura 950,000 petites sociétés alors qu'on en avait prévu seulement 770,000. D'après le nombre de demandes de conversion en petite société qui avaient été reçues en juillet 1987, l'IRS prévoit qu'il y aura environ 1.1 million de petites sociétés pour l'année d'imposition 1987. Nous avons l'intention de suivre la situation de près et de tenir compte de toute augmentation substantielle du nombre de petites sociétés à mesure que nous traitons les données des recensements économiques.

Comme nous l'avons mentionné précédemment, un autre problème vient de l'attitude des contribuables (qui sont aussi des répondants des enquêtes du Census Bureau) devant les limites et les possibilités qui résultent de la réforme fiscale. Par exemple, le fait de reporter le revenu et d'anticiper les déductions et les gains en capital influera sur les comparaisons faites d'une année à une autre de ces articles tels qu'ils sont déclarés au Census Bureau. En outre, les abris fiscaux, l'amortissement de biens immobiliers et les dispositions concernant l'épargne-retraite ont une incidence sur le volume des placements pour ces catégories. En tenant compte des changements dans l'attitude des contribuables, il nous est possible de prévoir dans quelle mesure ces changements peuvent influer sur nos données.

4. CE QUE L'AVENIR NOUS RÉSERVE

Le Census Bureau s'intéresse à la réforme fiscale parce qu'il désire accorder suffisamment d'attention aux données qu'il ne recueille pas mais qui sont néanmoins un élément essentiel de ses opérations. Il serait trop facile de supposer que les organismes administratifs continueront de nous fournir des données complètes et sûres qui répondent à nos besoins. Nous pourrions oublier que les objectifs des organismes statistiques sont différents de ceux des utilisateurs de statistiques. Quelle que soit l'importance que nous attribuons à nos activités, nous représentons un aspect relativement insignifiant des opérations d'un organisme administratif. Par conséquent, il faut faire connaître nos besoins, nous tenir au courant des faits nouveaux et parer à toute éventualité. À cette fin, il faudrait proposer des changements et veiller à ce que les organismes administratifs s'engagent à nous fournir des données actuelles. C'est un tel équilibre que nous désirons atteindre relativement à toutes nos utilisations des dossiers administratifs pour les besoins de nos programmes statistiques.

Compte tenu de la réforme fiscale de 1986 et des données fiscales en général, le Census Bureau est en train d'élaborer un système qui permettrait à tous les statisticiens d'évaluer les principaux systèmes de dossiers administratifs de l'administration fédérale américaine afin de déterminer si les données de ces systèmes sont applicables à leurs programmes statistiques. Ce système, que l'on désigne sous le nom de ARIS (Administrative Records Information System/système d'information sur les dossiers administratifs), a été établi à partir des résultats d'une étude préliminaire entreprise par le Subcommittee on Statistical Uses of Administrative Records (sous-comité des utilisations statistiques des dossiers administratifs) qui relève du U.S. Federal Committee on Statistical Methodology (comité fédéral américain de la méthodologie statistique). Il vise à recueillir les renseignements détaillés dont les utilisateurs de statistiques ont besoin pour déterminer si un des principaux systèmes fédéraux de dossiers administratifs peut convenir à leurs besoins, s'ils peuvent avoir accès aux données de ces systèmes et si ces données sont de qualité suffisante.

Nous avons commencé par déterminer les principaux systèmes fédéraux qui feraient l'objet de notre étude. À cette fin, nous nous sommes fondés sur les résultats du Project Link-Link* (projet couplage-couplage) pour relever les systèmes qui avaient déjà fait l'objet d'études de couplage et nous avons pris note des principaux systèmes présentés dans

le Statistical Policy Working Paper #6 (voir réf.) (document de travail n° 6 sur la politique statistique) ainsi que d'autres systèmes traités dans le document intitulé **Federal Information Sources and Systems** (voir réf.). Nous avons ensuite établi une liste d'environ 55 systèmes administratifs exploités par vingt organismes gouvernementaux. Des questionnaires ont été envoyés à chacun de ces organismes afin de recueillir les renseignements suivants:

- la structure matérielle du fichier et sa taille;
- la population visée (par ex. les particuliers, les ménages, les entreprises);
- le genre de données;
- les méthodes de collecte des données;
- la fréquence de la collecte, de la mise à jour et de la correction des données;
- les possibilités d'accès au fichier;
- les études de la qualité des données;
- la documentation disponible;
- les personnes avec qui il faut communiquer pour obtenir plus de renseignements.

Les organismes visés se sont montrés très coopératifs et nous sommes très satisfaits du taux de réponse. Il va sans dire que de nombreuses questions ont été posées concernant l'utilisation des données. Nous avons parlé au téléphone avec les représentants des organismes et nous les avons rencontrés afin de leur expliquer l'objet de notre étude et de coordonner la tâche de déclaration à l'intérieur de chaque organisme. Dans certains cas, nous avons constaté que le système observé n'était pas réellement un fichier administratif ou ne constituait qu'une sous-série d'importance secondaire d'un système plus important. À l'occasion, l'organisme expliquait que les fichiers et les données étaient de nature confidentielle et protégés par la **Privacy Act** ou par d'autres dispositions légales et pouvaient être consultés uniquement par le personnel de l'organisme. De façon générale, nous avons pu convaincre les organismes que les personnes qui utiliseraient la base de données seraient bien informées des restrictions d'accès et qu'il était important que leur système soit inclus dans notre étude afin d'assurer l'intégralité de la couverture.

Jusqu'à présent, nous avons reçu 47 questionnaires remplis dont les renseignements ont été introduits dans une base interactive de données relationnelles (voir le tableau 2 qui donne la liste actuelle des fichiers). Cette base de données a été constituée de façon à pouvoir être utilisée directement sur un ordinateur individuel compatible avec le matériel IBM qui utilise un système d'exploitation à disques. Nous pourrons fournir cette base de données sur disques souples, ou bien l'utilisateur pourra composer le numéro d'un panneau d'affichage électronique du Census Bureau et transférer le fichier directement. Pour avoir accès au panneau d'affichage, l'utilisateur a besoin d'un modem, d'un logiciel de transmission et du PC/MS-DOS, version 2.0 ou postérieure.

---

* Le projet LINK-LINK (projet couplage-couplage) consiste en une collecte de renseignements relatifs aux études du couplage des dossiers administratifs. Ces renseignements portaient sur 26 études et ont été recueillis au début de 1985 par l'Administrative Records Subcommittee du Federal Committee on Statistical Methodology. Chaque étude comprend des renseignements sur l'objet du couplage, les méthodes utilisées, les fichiers qui ont fait l'objet du couplage, les exigences juridiques dont il faut tenir compte pour le couplage, les méthodes de diffusion des données et les personnes avec lesquelles il faut communiquer pour avoir plus de renseignements. Les données sont contenues dans une base interactive et sont disponibles sur disques souples ou peuvent être obtenues en direct en passant par un panneau d'affichage électronique du Census Bureau.

Une des caractéristiques les plus intéressantes de ce système est qu'il permet de garder les données à jour. Nous avons conçu un tableau d'affichage qui servira aux organismes visés à corriger et à mettre à jour facilement les données que nous introduirons ensuite dans la base de données. En outre, nous enverrons chaque année des imprimés des renseignements contenus dans la base de données aux organismes à des fins de vérification. Nous espérons ainsi que l'ARIS demeurera un système actuel et utile.

Comme nous cherchons de plus en plus à réduire le fardeau de réponse et les coûts, nous devons considérer les dossiers administratifs comme une source possible d'information qui compléterait ou remplacerait les données recueillies au moyen de recensements et d'enquêtes. Nous croyons que l'ARIS nous permettra d'évaluer cette possibilité. De plus, ce système nous fournira des renseignements importants sur les changements apportés aux fichiers que nous utilisons actuellement. Enfin, il aidera à mesurer et à améliorer la couverture de nos enquêtes et de nos recensements, à améliorer les techniques d'estimation ainsi qu'à évaluer et à compléter les données d'enquête.

Tableau 2

Systèmes de dossiers administratifs contenus dans l'ARIS

1. **THE NATIONAL DEATH INDEX/indice national des décès**
   National Center for Health Statistics
2. **THE INDIVIDUAL INCOME TAX EXTRACT FILE/fichier de données extraites sur l'impôt sur le revenu des particuliers**
   Internal Revenue Service
3. **THE W2/W2P WAGE AND TAX STATEMENT EXTRACT FILE/fichier des données extraites des formules W2 et W2P - Salaires et relevé d'impôt**
   Internal Revenue Service
4. **STATISTICS OF INCOME - INDIVIDUAL INCOME TAX RETURNS/statistiques du revenu - déclaration d'impôt sur le revenu des particuliers**
   Internal Revenue Service
5. **COMPENSATION AND PENSION MASTER RECORD FILE/fichier principal d'enregistrements sur les indemnités et les pensions**
   Veterans Administration
6. **THE LIST SAMPLING FRAME/base de sondage sur liste**
   National Agricultural Statistical Service
7. **THE BUSINESS MASTER FILE/fichier principal des entreprises**
   Internal Revenue Service
8. **THE EMPLOYMENT/PAYROLL EXTRACT FILE (FORMS 941/943)/fichier de données extraites sur l'emploi et la paye (formules 941-943)**
   Internal Revenue Service
9. **STATISTICS OF INCOME DIVISION PARTNERSHIP SAMPLE FILE/fichier-échantillon des sociétés de personnes de la Division des statistiques du revenu**
   Internal Revenue Service
10. **THE CHARACTERISTICS OF FOOD STAMP HOUSEHOLDS FILE/fichier des caractéristiques des ménages recevant des bons de nourriture**
    Food and Nutrition Service
11. **THE RETURN PEACE CORPS FILE/fichier du personnel de retour du Peace Corps**
    Peace Corps

12. THE INDIAN HEALTH SERVICE HEALTH CARE STATISTICS SYSTEMS/systèmes statistiques sur les services et soins de santé pour les Indiens
Indian Health Service

13. THE MASTER ESTABLISHMENT LIST/liste principale des établissements
U.S. Small Business Administration

14. THE U.S. ESTABLISHMENT AND ENTERPRISE MICRODATA FILE/fichier de microdonnées sur les établissements et les entreprises aux É.-U.
U.S. Small Business Administration

15. THE VA CHAPTER 106 EDUCATION MASTER FILE/fichier principal de l'enseignement - chapitre 106 de la VA
Veterans Administration

16. THE MASTER PROVIDER OF SERVICES FILE/fichier principal des organismes qui fournissent les services
Health Care Financing Administration

17. THE CENTRAL PERSONNEL DATA FILE/fichier de données sur le personnel central
Office of Personnel Management

18. THE RESIDENTIAL ENERGY CONSUMPTION SURVEY/enquête sur la consommation énergétique des particuliers
Energy Information Administration

19. THE 1980 DECENNIAL CENSUS 100% FILES*/fichiers intégraux du recensement décennal de 1980
Bureau of the Census

20. THE 1980 DECENNIAL CENSUS SAMPLE FILES*/fichiers-échantillon du recensement décennal de 1980
Bureau of the Census

21. THE INDIVIDUAL MASTER FILE/fichier principal des particuliers
Internal Revenue Service

22. THE PAYER MASTER FILE/fichier principal des payeurs
Internal Revenue Service

23. THE EMPLOYMENT AND WAGES SYSTEM (ES 202)/système d'emploi et de salaires (ES 202)
Bureau of Labor Statistics

24. THE SUPPLEMENTAL SECURITY RECORD/fichiers des enregistrements sur les pensions supplémentaires
Social Security Administration

25. THE UNEMPLOYMENT INSURANCE NAME AND ADDRESS FILE/fichier des noms et adresses des prestataires d'assurance-chômage
Bureau of Labor Statistics

26. THE SMALL BUSINESS ADMINISTRATION LOAN ACCOUNTING FILE/fichier de comptabilité des emprunts
Small Business Administration

27. THE INFORMATION RETURNS PROGRAM FILE/fichier du programme des déclarations de renseignements
Internal Revenue Service

28. THE HEALTH INSURANCE MASTER ENTITLEMENT'S FILE/fichier principal de l'admissibilité au régime d'assurance-maladie
Health Care Financing Administration

29. THE GUARANTEED AND INSURED LOAN SYSTEM/système de prêts garantis et assurés
Veterans Administration

30. THE DEFICIENCY, DISASTER, AND DIVERSION PAYMENTS SYSTEM/système de paiements d'appoint, de paiements en cas de désastres et de paiements de diversification
Agricultural Stabilization and Conservation Service

31. OFFICE OF GENERAL SALES MANAGER SYSTEM/système du bureau du directeur général des ventes
USDA: Agricultural Stabilization and Conservation Service

32. THE STATISTICS OF INCOME CORPORATE SAMPLE FILE/fichier-échantillon des statistiques du revenu des sociétés
Internal Revenue Service

33. THE IRS ESTATE TAX RETURNS FILE/fichier des déclarations d'impôt sur les successions du IRS
Internal Revenue Service

34. RAILROAD EMPLOYER'S CREDITABLE COMPENSATION RECORDING SYSTEM/système d'enregistrement des indemnités pouvant être créditées des sociétés ferroviaires
U.S. Railroad Retirement Board

35. RAILROAD RETIREMENT, DISABILITY, AND SURVIVOR BENEFIT PAYMENT SYSTEM/système de paiement des prestations de retraite, des indemnités d'invalidité et des prestations de survivant des sociétés ferroviaires
U.S. Railroad Retirement Board

36. THE GENERAL REFUGEE FILE/fichier général des réfugiés
Health and Human Services

37. THE STANDARD STATISTICAL ESTABLISHMENT LIST*/liste statistique type des établissements
Bureau of the Census

38. SUMMARY EARNINGS FILE/fichier sommaire des gains
Social Security Administration

39. THE MASTER BENEFICIARY RECORD/fichier principal des enregistrements sur les bénéficiaires
Social Security Administration

40. THE NUMIDENT FILE/fichier d'identification numérique
Social Security Administration

41. THE EMPLOYER REPORT RECORD/fichier des enregistrements sur les rapports des employeurs
Social Security Administration

42. THE SINGLE-UNIT CODE FILE/fichier à codes numériques uniques
Social Security Administration

43. THE MULTI-UNIT CODE FILE/fichier à codes numériques multiples
Social Security Administration

44. THE CHAPTER 30 EDUCATION MASTER FILE/fichier principal de l'enseignement - chapitre 30
Veterans Administration

45. THE CHAPTER 31 TARGET MASTER RECORD/fichier principal prototype - chapitre 31
Veterans Administration

46. THE CHAPTER 32 EDUCATION MASTER FILE/fichier principal de l'enseignement - chapitre 32
Veterans Administration

47. THE CHAPTER 34 AND 35 EDUCATION MASTER FILE/fichier principal de l'enseignement - chapitres 34 et 35
Veterans Administration

BIBLIOGRAPHIE

Comptroller General, Federal Information Sources and Systems 1984, Washington, U.S. Government Printing Office, 1984.

Jabine, T.B. (1987). "Statistical Uses of Administrative Records in the United States: Some Recent Developments"; document présenté à la réunion annuelle de la Société statistique du Canada, Québec.

Jonas, J., et Hanczaryk, P. (1987). "Traitement automatisé de l'assurance de la qualité des fichiers de dossiers administratifs". Communication présentée lors du Symposium sur les utilisations statistiques des données administratives.

Pechman, J.A. (1987). Tax Reform: Theory and Practise. *Economic Perspectives*, 11-28.

Quinn, J.B. (1987). "The Pluses and Minuses of Becoming an "S" Corporation", Washington Post, section des affaires.

Sater, D. (1987). Document non-publié, U.S. Bureau of the Census.

"Statistical Policy Working Paper #6", Report on Statistical Uses of Administrative Records, U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, décembre 1980.

Valdisera, V. (1987). Document non-publié, U.S. Bureau of the Census.

Wakefield, J.C. (1987). *The Tax Reform Act of 1986*. Survey of Current Business, 21.

SESSION VI: COMMUNICATIONS OFFERTES

Président: M.P. Singh, Statistique Canada

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

L'UTILISATION DE DOSSIERS ADMINISTRATIFS POUR L'ENQUÊTE SUR LE REVENU ET LA PARTICIPATION AUX PROGRAMMES (SIPP)

CHESTER BOWIE et DANIEL KASPRZYK[1]

La SIPP (Survey of Income and Program Participation) est une nouvelle enquête du Census Bureau qui fournit des renseignements sur les caractéristiques sociales, démographiques et économiques des particuliers et des familles aux États-Unis. Au nombre des données recueillies dans le cadre de la SIPP, on compte les revenus selon la source et le montant, la participation aux programmes, l'activité sur le marché du travail et les éléments d'actif selon le genre et le montant. Depuis le début, la SIPP vise à créer un système d'information intégrant des données administratives et des données d'enquête recueillies auprès des ménages. Même si cet objectif n'est pas encore atteint, plusieurs projets ont été mis sur pied à cette fin. Ce document examine donc les divers projets SIPP axés sur l'utilisation éventuelle de données de source administrative: 1) enrichississement des données d'enquête par des données complémentaires sur les revenus antérieurs et les prestations d'aide sociale reçues; 2) étude de vérification des enregistrements visant à évaluer les réponses aux questions relatives à certains types de revenu; 3) travaux de recherche sur l'estimation à partir de données de dossiers administratifs.

INTRODUCTION

La SIPP (Survey of Income and Program Participation/enquête sur les revenus et la participation aux programmes) a pour objet de fournir une information complète et détaillée au sujet des ressources économiques de la population américaine et de l'incidence des programmes fiscaux et de transfert de l'administration publique sur leur situation financière. Les données tirées de l'enquête devraient offrir aux responsables des politiques fédérales une source précieuse de renseignements pour déterminer l'efficacité des programmes fiscaux et de transfert, estimer les coûts futurs des programmes et le nombre des bénéficiaires éventuels et évaluer les répercussions de toute modification envisagée.

La SIPP fournit des données périodiques et annuelles détaillées sur le revenu et la participation à des programmes de transfert publics et privés des ménages américains. Des efforts considérables sont également consacrés à la mesure de divers types de ressources économiques autres que les revenus monétaires courants; il s'agit principalement des données SIPP sur les éléments d'actif, les dettes et les ressources non financières comme l'aide au logement déterminée après enquête sur les revenus,

[1] Chester E. Bowie, Chef, Income Surveys Branch, Demographic Surveys Division, Bureau of the Census, Washington, D.C. 20233, Daniel Kasprzyk, Chef, SIPP Research and Coordination Staff, Office of the Director, Bureau of the Census, Washington, D.C. 20233.

l'assurance-maladie au titre de régimes publics et privés, les fonds de retraite et autres avantages sociaux liés à l'emploi.

La SIPP a été créée quand on s'est rendu compte que la principale source d'information sur la distribution du revenu des particuliers et des ménages aux États-Unis — l'enquête sur le revenu menée au mois de mars en tant que complément à la CPS (Current Population Survey) — comportait des lacunes à ce point importantes qu'on n'avait d'autre choix que de remanier en profondeur les méthodes de collecte de données et le questionnaire lui-même. Une des principales faiblesse de la CPS était l'absence de couplages possibles entre les données tirées de l'enquête et les données des dossiers administratifs, à des fins statistiques. Tenant compte de cette limite de la CPS et de l'utilité analytique certaine du couplage des données d'enquête et des données des dossiers administratifs, les concepteurs de l'enquête SIPP ont finalement fait savoir que les travaux devaient être orientés vers l'établissement d'un système où les données de source administrative pourraient être jumelées aux données obtenues auprès des ménages, au moyen du numéro de sécurité sociale. Dans l'énoncé des objectifs de la SIPP, Lininger (1980) déclarait que les dossiers administratifs serviraient:

1. à accroître l'efficacité de l'échantillonnage pour certaines sous-populations (p. ex. les bénéficiaires du programme des prestations de sécurité de la vieillesse, de survivant et d'assurance-invalidité (OASDI) et les bénéficiaires du revenu supplémentaire de sécurité sociale (SSI);
2. à établir des comparaisons avec les données d'enquête pour la validation d'éléments communs aux deux sources; et
3. à compléter les données d'enquête dans le cas de renseignements difficiles à obtenir par sondage (p. ex. au sujet des revenus antérieurs et des prestations reçues à ce jour).

Ces objectifs ont été considérés d'abord dans le cadre du programme de mise au point de l'enquête SIPP et ensuite dans la mise en oeuvre de l'enquête elle-même.

Ce document décrit l'évolution constante de l'objectif de la SIPP concernant l'utilisation des dossiers administratifs à des fins statistiques. La section I est une revue des travaux de recherche et développement sur l'utilisation des données administratives entrepris dans le cadre du programme ayant donné naissance à la SIPP; la section II décrit le plan de sondage et le contenu de la SIPP ainsi que le programme mis sur pied pour assurer l'exactitude des SSN (Social Security Numbers/numéros de sécurité sociale) des répondants et faciliter ainsi les rapprochements entre les renseignements fournis dans l'enquête et les données de sources administratives; la section III décrit cinq domaines d'application ou projets pour lesquels on a déjà entrepris certains travaux axés sur l'utilisation conjointe des données d'enquête et des données administratives à des fins statistiques; enfin, la section IV présente quelques exemples de couplage des données SIPP et des données de sources administratives.

1. UTILISATION DES DOSSIERS ADMINISTRATIFS DANS L'ISDP

L'ISDP (Income Survey Development Program/programme de mise au point de l'enquête sur le revenu), approuvé en 1975, avait pour objectif l'élaboration de méthodes et d'un plan de sondage pouvant contribuer à résoudre les problèmes de non-réponse et d'erreurs de classification associés à la Current Population Survey (Yeas et Lininger, 1981). L'ISDP a également permis d'améliorer les méthodes de collecte des SSN. L'orientation donnée à ce programme s'inspirait largement des travaux de recherche et développement de Scheuren et de ses collaborateurs (Scheuren et coll., 1975) concernant le "1973 Exact Match File" (fichier des concordances) (voir Kills et Scheuren, 1978).

Kasprzyk (1983) et Griffith (1980) passent en revue l'utilisation qui a été faite des dossiers administratifs dans le cadre de l'ISDP. Nous allons en présenter les grandes lignes ici afin de mieux faire ressortir le contexte dans lequel s'incrivent les projets SIPP actuels et futurs axés sur l'utilisation des dossiers administratifs à des fins statistiques.

Dans l'ISDP, on jugeait essentiel de pouvoir obtenir un SSN (numéro de sécurité sociale) valide pour chaque personne faisant partie de l'échantillon. En insistant ainsi sur l'importance de la collecte des SSN et grâce aussi au système de validation et de correction des SSN qui a été mis au point, on a observé que 95.5% des unités échantillonnées avaient déclaré ou s'étaient vu attribuer un SSN approprié (Kasprzyk, 1983). C'est le même système qui a servi de prototype au système SIPP décrit dans la section qui suit.

L'ISDP a consisté en quatre essais sur le terrain effectués dans le but d'étudier un certain nombre de notions de base, de méthodes et de questionnaires. Pour chacun de ces essais, on a décidé d'utiliser des dossiers administratifs comme base de sondage. Même si un tel procédé avait surtout pour objet d'accroître l'efficacité de l'échantillonnage dans certaines sous-populations au moyen d'estimateurs pour bases de sondage multiples, le résultat le plus frappant de ces études de faisabilité est qu'elles auront permis aux planificateurs de l'enquête de mieux comprendre les difficultés administratives, méthodologiques et opérationnelles qui sont liées à l'utilisation de sources administratives à des fins d'échantillonnage.

Au cours des années de mise en oeuvre de l'ISDP, les sources de dossiers administratifs suivantes ont été utilisées: 1) le fichier principal de l'AFDC (Aid to Families with Dependent Children/aide aux familles à faible revenu avec enfants à charge) tenu par le ministère du Bien-être social de l'État du Texas[2]; le SSR (Supplemental Security Record/fichier de base sur le progamme du revenu supplémentaire de sécurité sociale)[3]; le MBR (Master Beneficiary Record/fichier maître des bénéficiaires)[4]; le fichier du BEOG (Basic Educational Opportunity Grant/fichier sur les personnes ayant présenté une demande en vertu du programme de subventions à l'éducation de base)[5]; le fichier VAPCF

---

[2] Le fichier principal de l'AFDC (Aid to Families with Dependent Children/aide aux familles à faible revenu avec enfants à charge) est un système de dossiers administratifs que tient chaque État et dans lequel on trouve des données sur les montants des prestations, des données chronologiques sur les versements faits à ce jour, des données sur les caractéristiques démographiques des bénéficiaires et autres renseignements nécessaires à l'administration de ce programme.

[3] Le SSR (Supplemental Security Record) est le fichier de base national contenant des données sur le programme SSI (Supplemental Security Revenue/revenu supplémentaire de sécurité sociale), dont des données sur les montants de prestations, sur les versements à ce jour et des données démographiques.

[4] Le MBR (Master Beneficiary Record/fichier maître des bénéficiaires) est un fichier de base national sur le programme des prestations de sécurité de la vieillesse, de survivant et d'assurance-invalidité (prestations de classe II); on y trouve des données chronologiques et courantes sur les demandeurs de prestations de classe II, sur les bénéficiaires passés et actuels de prestations en espèces, sur les demandeurs jugés non admissibles et sur les demandes refusées.

[5] Le fichier du BEO (Basic Educational Opportunity/programme de subventions à l'éducation de base) est un fichier administratif que tient le Department of Education. On y trouve des données sur toutes les personnes ayant présenté une demande au cours d'une année scolaire donnée, y compris sur les personnes non admissibles, sur les personnes admissibles n'ayant pas utilisé leur subvention et sur les personnes admissibles ayant utilisé leur subvention. Le programme BEOG est maintenant appelé le programme des subventions Pell.

(Veterans Administration Pension and Compensation file/fichier des indemnités et pensions des anciens combattants)[6]; l'IMF (Individual Master File/fichier de base des particuliers) du Internal Revenue Service[7]; et les fichiers administratifs des États sur l'assurance-chômage et l'indemnisation des accidents du travail.

Dans le cadre de l'ISDP, les dossiers administratifs n'ont pas seulement servi à l'échantillonnage, mais ont aussi été utilisés pour clarifier des cas d'absence de déclaration ou de déclaration erronée de prestations versées au titre de divers programmes, les données d'enquête étant comparées aux données administratives. Vaughan (1978) et Goudreau, Oberheu et Vaughan (1981, 1984) rendent compte des résultats d'études ISDP qui ont entraîné la modification des questionnaires d'enquête pour réduire les erreurs de classification des sources de revenu.

Enfin, bien qu'on visait aussi au départ à créer une base de des données enrichie de données administratives pour les renseignements difficiles à recueillir dans une enquête auprès des ménages, cet objectif n'a pu être atteint. On n'a jamais pu donner suite comme prévu à un projet d'appariement avec les données du fichier SER (Summary Earnings Record/état récapitulatif des gains assurés)[8], en raison d'autres priorités.

2. PLAN DE SONDAGE ET CONTENU

Cette section fournit une vue d'ensemble du plan de sondage et du contenu de l'enquête SIPP et elle est suivie d'une description du programme de collecte et de validation des numéros de sécurité sociale qui a été établi dans le cadre de cette enquête.

1. Caractéristiques du plan de sondage de la SIPP

Au moment du choix du plan de sondage de la SIPP, on avait pour principaux objectifs d'améliorer la collecte des données sur le revenu et des autres données relatives aux programmes sociaux et de faciliter l'analyse détaillée des changements dans le temps. Le plan de sondage devait aussi permettre de recueillir un volume élevé de données, tout en étant suffisamment souple pour que certains renseignements puissent être obtenus plus fréquemment que d'autres. Pour atteindre ces objectifs, on a donc conçu un plan de sondage selon lequel les mêmes personnes sont interviewées plus d'une fois (enquête par

---

[6] Le VAPCF (Veteran's Administration Pension and Compensation File/fichier des indemnités et pensions des anciens combattants) est un fichier maître national contenant des dossiers sur les prestations versées sous forme d'allocations pour personnes à charge, de pensions d'invalidité, de pensions au décès et d'allocations d'enterrement.

[7] L'IMF (Individual Master File/fichier de base des particuliers) du Internal Revenue Service est un fichier national contenant certaines données relatives au revenu et à l'impôt qui sont tirées de toutes les déclarations d'impôt sur le revenu des particuliers, notamment sur les salaires, les revenus sous forme de dividendes et d'intérêts, les impôts payés et les exemptions.

[8] Le SER (Summary Earnings Record/état cumulatif des gains) est un fichier dans lequel on trouve des données sur le montant total des gains assurés d'une personne depuis son entrée en vie active (jusqu'à concurrence du montant maximum pour chaque employeur) ainsi que sur le nombre de trimestres dans lesquels cette personne a versé des cotisations au titre de la sécurité sociale. Ces données servent à déterminer si une personne a droit aux prestations et à calculer le montant des prestations. Les personnes sont identifiées par leur numéro de sécurité sociale dans ce fichier.

panel). Une fois tous les quatre mois, tous les membres âgés de 15 ans et plus des ménages sélectionnés pour faire partie d'un panel donné sont interrogés au sujet de leurs revenus et autres questions, et ce, sue une période d'environ deux ans et demi. S'ils déménagent, les répondants sont interviewés à leur nouvelle adresse, de même qu'on interview toute personne qui décide de former un ménage avec un répondant ou avec lequel le répondant lui-même emménage. De cette façon, on peut constituer un dossier chronologique très détaillé sur chaque membre d'un ménage et sur chaque ménage d'un panel. Ce genre de plan de sondage impose un fardeau moins lourd aux répondants étant donné qu'ils n'ont qu'à remonter à quelques mois pour se rappeler la plupart des renseignements demandés et qu'il est également possible de limiter le nombre des questions posées au cours d'une interview.

Pour accroître la précision des estimations du changement, en particulier d'une année à une autre, un nouveau panel est créé chaque année au lieu d'attendre à la fin de la période allouée pour chaque panel. Par conséquent, deux et parfois trois panels peuvent faire en même temps l'objet d'une enquête sur le terrain. Grâce à un tel chevauchement des panels, il est possible de produire des estimations recoupées à partir des résultats obtenus d'un échantillon beaucoup plus grand puisque sa taille double ou triple lorsque deux ou trois panels se chauvauchent.

Le premier panel SIPP, dit panel de 1984 bien que les opérations sur le terrain aient débuté en octobre 1983, comptait environ 20,000 ménages au départ. Le deuxième panel, soit le panel de 1985, était constitué d'environ 14,000 ménages, pour lesquels les interviews ont commencé en février 1985. On s'attend à créer en février de chaque année un nouveau panel d'environ 12,300 ménages. Au cours de chaque "vague" (voir paragraphe suivant) d'interview d'un panel, la taille d'un échantillon varie en fonction des pertes (membres qui quittent un ménage) et des gains (personnes retracées à leur nouvelle adresse).

Pour les principaux éléments d'information recueillis dans le cadre de l'enquête, la période de référence correspond aux quatre mois précédant l'interview; p. ex. en février, la période de références va des mois d'octobre à janvier précédents. Lorsqu'un ménage est interviewé à nouveau en juin, la période de référence s'étend de février jusqu'à la fin de mai. Au lieu d'avoir à assumer une lourde charge de travail d'interview et de traitement à tous les quatre mois, on a jugé plus pratique de répartir le travail chaque mois. A cette fin, les ménages d'un **panel** donné sont divisés en quatre sous-échantillons ayant sensiblement la même taille. Ces sous-échantillons s'appellent des **groupes de renouvellement**, et un groupe de renouvellement, ou le quart de l'échantillon, est interviewé chaque mois. Il faut donc quatre mois consécutifs pour que tout l'échantillon ait fait l'objet de l'interview. Cette période d'interview de quatre mois est ce qu'on appelle une "**vague**".

2. Contenu de l'enquête SIPP

Chaque interview ne devrait pas prendre plus de 30 minutes du temps d'un répondant et comprend trois grands groupes de questions. Les deux premiers groupes de questions restent sensiblement les mêmes pour chaque vague et pour chaque panel. Le troisième groupe de questions porte sur des sujets qui changent au cours de chaque vague d'interview des ménages d'un panel. De cette façon, il est possible de varier quelque peu le contenu d'un panel à l'autre, bien que nombre de sujets soient répétés à tous les panels. Chaque groupe de renouvellement d'une vague répond à la même série de questions bien que, comme nous l'avons vu, la période de référence soit différente.

Les questions du premier groupe sont des éléments d'information recueillis sur une fiche de contrôle. Cette fiche de contrôle est un document distinctif du questionnaire et on s'en sert à plusieurs fins importantes. La fiche de contrôle est utilisée pour dresser la liste de chaque membre d'un même ménage situé à une adresse donnée et pour enregistrer

les caractéristiques socio-démographiques de base (âge, race, sexe et ainsi de suite) de chaque personne dénombrée au moment de la première interview. La fiche permet aussi de recueillir certains renseignements relatifs à l'unité de logement ou au ménage, notamment sur le nombre de pièces, le mode d'occupation du logement et autres renseignements du genre. La fiche est réutilisée lors des interviews subséquentes pour noter tout changement dans les caractéristiques tels que l'âge, le niveau de scolarité et l'état matrimonial ainsi que les dates auxquelles les personnes ont intégré ou quitté le ménage. Enfin, au cours de chaque interview, on transcrit sur la fiche de contrôle l'information obtenue au sujet de chaque source de revenu de même que le nom de chaque emploi ou entreprise, de façon à pouvoir utiliser ces renseignements à des fins de mise à jour à l'interview suivante.

Le deuxième grand groupe de questions constitue la partie principale du questionnaire, laquelle est répartie en cinq sections. Toutes les questions de cette série sont posées lors de la première interview et font ensuite l'objet d'une mise à jour lors de chaque interview subséquente.

La première section de la partie principale recueille des données de base sur l'activité au cours des quatre mois de la période de référence. C'est également dans cette première section que les répondants fournissent la plupart des renseignements sur les revenus de diverses sources qu'ils ont touchés au cours de la période de référence. Cela comprend les revenus de sources publiques comme les allocations d'aide aux familles à faible revenu avec enfants à charge, le revenu supplémentaire de sécurité sociale, les prestations d'aide générale et les indemnités pour accident du travail. On demande aussi aux répondants d'indiquer non seulement les prestations de sécurité sociale de la vieillesse, mais aussi tout autre revenu de retraite comme les pensions versées par les compagnies de chemin de fer, les entreprises, les syndicats et la Fonction publique. C'est également dans cette section que sont déclarées diverses sources de revenu comme les pensions alimentaires ou les pensions pour l'entretien des enfants, les intérêts sur placements, les revenus pour garde d'un enfant placé en famille d'accueil et les subventions d'aide à l'éducation. Cette section comporte également des questions sur les principales sources d'avantages non financiers tels que les bons alimentaires, le programme WIC (Women, Infants, and Children Nutrition Program/programme axé sur l'alimentation des femmes, des nouveaux-nés et des enfants), Medicaid, Medicare, et les prestations d'assurance-maladie.

La deuxième section de la partie principale du questionnaire SIPP vise à recueillir des renseignements associés à la rémunération. Cette section comprend des données sur la branche activité, la profession ainsi que le salaire horaire du répondant, jusqu'à concurrence de deux emplois.

La troisième section de la partie principale met l'accent sur le revenu provenant d'un travail autonome et sur la collecte de renseignements précis sur le genre d'entreprise--société de capitaux, entreprise individuelle ou encore société de personnes--et sur les pertes et profits enregistrés. Les répondants peuvent là aussi déclarer jusqu'à deux emplois indépendants.

La quatrième section est celle qui est consacrée à la déclaration des montants de toutes sources. En effet, c'est dans cette section que le répondant inscrit le montant mensuel des sources de revenus précisées dans la première section et se rapportant aux quatre mois de la période de référence. Les montants provenant d'un maximum de six sources peuvent être indiqués dans cette section.

La cinquième et dernière section du questionnaire central porte sur les gains provenant de divers éléments d'actif comme les comptes d'épargne, les obligations, les actions et les biens en location et autres possessions du genre. L'information s'applique à la période de référence de quatre mois et les sommes peuvent être déclarées sur une base individuelle ou conjointe, si le répondant est codétenteur.

Le troisième grand groupe de questions consiste en divers suppléments ou modules thématiques joints au questionnaire principal au cours des vagues qui suivent la première interview. Le recours à des modules thématiques permet l'étude d'un large éventail de sujets. La diversité des données ainsi recueillies contribue à accroître l'utilité et les possibilités d'application de la base de données SIPP. Un module peut être utilisé pour les vagues 2 à 8 d'un panel (et même 9 dans le cas du panel de 1984) puisqu'on consacre généralement moins de temps à la mise à jour de l'information de base après l'interview initiale. De plus, selon le temps dont on dispose et la longueur des modules, on peut prévoir plus d'un module au cours d'une même vague. Les modules thématiques portent sur des sujets qui n'ont pas à être examinés tous les quatre mois et peuvent se rapporter à une période de référence différente de celle des questions principales. Certains modules ne sont utilisés qu'au cours d'une seule vague d'un panel, tandis que d'autres seront répétés sur plus d'une vague. Les modules offrent des perspectives plus vastes d'analyse en recueillant de l'information sur une diversité de sujets non traités dans la partie principale du questionnaire. Les données des modules thématiques peuvent faire l'objet d'une analyse distincte ou encore être associées aux éléments d'information de la fiche de contrôle ou aux réponses fournies aux questions principales. Souvent, les ménages de panels différents doivent répondre aux questions d'un même module et les données sont traitées conjointement pour accroître la fiabilité des estimations.

3. Collecte et validation des numéros de sécurité sociale dans la SIPP

Depuis le début, il a été prévu que le système de données SIPP utiliserait à la fois des données des dossiers administratifs et des données d'enquêtes auprès des ménages. Une telle façon de procéder permet de réduire le fardeau de réponse puisqu'on a recours à d'autres sources d'information pour les données difficiles à obtenir par sondage. Les réponses fournies dans le cadre des interviews peuvent être complétées à l'aide des renseignements contenus dans les fichiers des organismes responsables des programmes, comme les dossiers sur les gains assurés et les prestations que tient la SSA (Social Security Administration/Administration de la sécurité sociale). Ce genre de couplage de données peut notamment servir à l'analyse de l'incidence à long terme de diverses formules en matière de prestations de sécurité sociale.

Pour que de tels liens puissent être établis avec exactitude, il faut obtenir le numéro de sécurité sociale (SSN) de toutes les personnes faisant partie de l'échantillon. On insiste auprès des interviewers sur la très grande importance de veiller à obtenir un SSN pour chacun des membres des ménages participant à la SIPP. Ces numéros sont inscrits sur la fiche de contrôle puis vérifiés et corrigés au besoin afin que l'on puisse établir de manière certaine le plus grand nombre de liens avec les autres systèmes de dossiers.

Le processus de vérification et de correction a été élaboré en fonction des résultats des travaux d'essai de l'ISDP (Kasprzyk, 1983). À la fin de chaque interview mensuelle de la première vague d'un panel SIPP, le Census Bureau crée pour la SSA un fichier spécial de données. Ce fichier contient un petit nombre de variables-clés (SSN, nom, date de naissance, sexe) pour toutes les personnes comprises dans l'échantillon initial et ayant un SSN (enfants y compris), lesquelles sont présentées sous une forme appropriée à une validation machine. On traite à part et avec une méthode manuelle les cas des personnes ayant déclaré ne pas posséder de SSN ou ne pas être en mesure de fournir ce numéro.

Les personnes qui ont refusé de fournir un SSN sont exclues de la recherche. La SSA identifie, par validation automatisée, les numéros erronés et procède ensuite à une vérification manuelle de ces cas et des cas des personnes n'ayant pas fourni de SSN. Ce travail est achevé à l'interview de la quatrième vague, au moment où un suivi sur le terrain est réalisé pour obtenir les SSN manquants (à condition qu'il ne s'agisse pas de "refus") et pour résoudre les cas de non-concordance du SSN ou des données démographiques constatés lors de l'appariement automatisé ou de la vérification manuelle.

Les numéros de sécurité sociale des personnes entrées dans l'échantillon après la vague 1 (parce qu'elles ont commencé à vivre avec une personne faisant partie de l'échantillon initial) sont validés au début du panel suivant. Par exemple, l'information recueillie au cours des vagues 2 à 5 au sujet des nouveaux membres (personnes hors échantillon) du panel de 1984 a été conservée et a fait l'objet d'une validation automatisée en même temps que les données de la vague 1 du panel de 1985. C'est ainsi également que les renseignements sur les personnes hors échantillon obtenus au cours des vagues 6, 7 et 8 du panel de 1984 et des vagues 2, 3 et 4 du panel de 1985 ont été soumis pour validation automatisée avec les données de la vague 1 du panel de 1986.

Voici un résumé des résultats de la validation des SSN réalisée pour l'échantillon de la vague 1 du panel de 1984:

53,588	nombre total de personnes comprises dans l'échantillon à la vague 1
-1,674	personnes ayant refusé de fournir un SSN et qui ont été exclues du processus de validation
51,914	personnes admissibles à la validation du SSN
-42,128	personnes ayant déclaré un SSN utilisable et admissibles à la validation automatisée
9,786	personnes n'ayant pas fourni un SSN et admissibles à la recherche manuelle (dans la plupart des cas des enfants)
44,172	SSN validés (85% des personnes admissibles)
-7,742	SSN non validés (dans la plupart des cas des enfants n'ayant pas de SSN)
51,914	personnes admissibles à la validation du SSN

En se fondant sur ces résultats, Sater (1986) a conclu que pour pour les personnes qui ont un SSN, le taux d'acquisition se situe entre 93 et 97%.

2. COUPLAGE DES DONNÉES DE LA SIPP ET DES DOSSIERS ADMINISTRATIFS

Cette section décrit brièvement cinq domaines d'application ou projets pour lesquels on a déjà entrepris certains travaux d'utilisation conjointe des données d'enquête et des données de dossiers administratifs: 1) le projet de couplage des données SIPP/SSA; 2) l'étude de faisabilité de la collecte auprès des employeurs de données sur le montant des contributions; 3) l'étude de vérification des enregistrements SIPP; 4) l'utilisation des dossiers administratifs pour les estimations SIPP; et 5) la fusion des données économiques et des données démographiques SIPP.

1. Projet de couplage des données SIPP/SSA

L'intérêt que représente pour la SSA un ensemble intégré de données administratives et de données d'enquêtes-ménages est très proche des applications prévues au moment de la création de la SIPP. En effet, la SSA pourrait utiliser un tel ensemble de données aux fins suivantes:

1. L'estimation des coûts futurs des programmes — La SSA doit établir des prévisions des coûts des principaux programmes dont elle est responsable, notamment l'OASDI (Old Age, Survivor, and Disability Program/programme des prestations de

sécurité de la vieillesse, de survivant et d'assurance-invalidité), le SSI (Supplemental Security Income/revenu supplémentaire de sécurité sociale) et l'AFDC (Aid to Families with Dependent Children/aide aux familles à faible revenu avec enfants à charge). Pour accroître la précision des méthodes de projection des coûts, les données recueillies auprès des ménages des panels SIPP peuvent être appariées avec les données du SSA se rapportant à un certain nombre d'années. Ceci devrait permettre une analyse des entrées et sorties en plus des estimations de la prévalence de la participation aux programmes de la SSA à un moment donné dans le temps. Les relations ainsi établies entre la participation aux programmes et les caractéristiques des bénéficiaires peuvent alors aider la SSA à mieux prévoir à l'avance l'évolution ainsi que les coûts futurs des programmes.

2. Évaluation des répercussions possibles de changements d'orientation des programmes — Un couplage SIPP-SSA permettrait de réunir des données sur la famille, les revenus et les prestations de sécurité sociale. Grâce à une telle combinaison de renseignements, la SSA serait en mesure d'estimer les coûts directs de changements d'orientation reliés à ces trois facteurs et aussi d'évaluer les effets que ces changements pourraient avoir sur le bien-être économique des bénéficiaires.

3. Description des caractéristiques particulières des bénéficiaires (c.-à-d. non directement liées aux programmes) — À la demande de membres du Congrès et d'autres personnes, la SSA est souvent appelée à fournir des renseignements au sujet des bénéficiaires de ses programmes qui ne sont pas saisis de manière systématique dans les systèmes de dossiers administratifs. Par le passé, pour pouvoir fournir une telle information, la SSA a mené des enquêtes très sporadiques et souvent même uniques. Comme la conjoncture n'est actuellement pas favorable à une nouvelle série d'enquêtes spéciales du genre, un système permanent de couplage des données SIPP/SSA fournirait des données relativement à jour à intervalles réguliers.

4. Vérification des théories du domaine des sciences sociales pouvant être appliquées aux programmes de sécurité sociale — Le plan de sondage de la SIPP offre en soi des possibilités intéressantes d'analyses longitudinales. Les données recueillies au moyen des questions de base et des modules thématiques sont suffisamment nombreuses et variées pour les besoins de la vérification de certaines théories socio-économiques associées à la participation aux programmes. Par conséquent, ces données devraient contribuer de manière significative à l'avancement de la recherche fondamentale dont tout programme social doit faire l'objet pour pouvoir continuer à évoluer.

En gros, le projet repose sur un couplage maximum entre les données SIPP et les données de la SSA. Les données recueillies au cours de chacune des vagues de chaque panel SIPP, y compris les données du questionnaire central et des modules thématiques, seront appariées aux données extraites des principaux fichiers de base de la SSA: le MBR (Master Beneficiary Record/fichier maître des bénéficiaires), qui contient des données chronologiques sur l'admissibilité et les prestations versées au titre du programme OASDI; le SSR (Supplemental Security Record), qui contient aussi des données chronologiques sur l'admissibilité et les prestations versées au titre du programme SSI (revenu supplémentaire de sécurité sociale); et le SER (Summary Earnings Recors/état récapitulatif des gains), qui contient des données chronologiques sur les gains assurés de chaque travailleur. Les registres de la SSA seront mis à jour périodiquement de façon à pouvoir introduire dans les fichiers établis pour chaque panel des données couvrant une plus longue période. Il est possible également que nous procédions plus tard à un couplage avec les données d'un nouveau fichier administratif sur l'assurance-invalidité que la SSA est en train de

constituer. Évidemment, un accord devra être conclu entre la SSA et le Bureau of the Census pour tout projet de couplage donné.

Les données ainsi jumelées seront conservées dans l'ordinateur du Census Bureau. L'accès au fichier sera limité au personnel du Census Bureau et aux employés de la SSA spécialement désignés à cette fin par le Census Bureau. La SSA pourra publier seulement des statistiques sommaires de façon que les données ne puissent révéler l'indentité d'un ménage, d'une famille ou d'un particulier.

Les principales tâches associées au projet de couplage sont les suivantes:

1. La vérification et la recherche des SSN (numéros de sécurité sociale) — Cette tâche s'inscrit dans les activités courantes de la SIPP et a déjà été décrite. Il faut cependant souligner que c'est le personnel de la SSA qui a été chargé des opérations de validation de la très grande majorité des SSN des membres du panel SIPP de 1984.
2. L'obtention des dossiers administratifs de la SSA — comme nous l'avons déjà mentionné, ce projet nécessite le rapprochement des données SIPP et des données du MBR, du SSR et du SER. Des décisions devront être prises au sujet du genre de données extraites de ces fichiers qui pourraient être utilisées dans l'appariement.
3. La fusion des données administratives et des données d'enquête SIPP — Le rapprochement des données doit se faire dans une perspective de continuité. En effet, nous prévoyons qu'un certain nombre d'opérations de traitement des données se répéteront pour chaque panel SIPP.
4. La pondération, l'imputation et l'estimation des erreurs d'échantillonnage — Nous devrons étudier et élaborer des méthodes de pondération et d'imputation pour tenir compte des enregistrements SIPP non appariés. Nous aurons besoin de coefficients de pondération aux fins d'analyse tant transversale que longitudinale. La SSA devra aussi être en mesure d'estimer les erreurs d'échantillonnage.
5. Constitution d'une documentation relative aux fichiers faisant l'objet d'un rapprochement — Une telle documentation devrait comprendre une description des bandes et des renseignements sur leur utilisation, les questionnaires SIPP, une description des dossiers administratifs de la SSA et des méthodes d'échantillonnage et d'imputation ainsi que toute autre information utile pour l'estimation ou l'analyse.

2. Étude de la faisabilité de la collecte auprès des employeurs de données sur les montants des contributions

Les contributions des employeurs aux régimes d'assurance-maladie, de retraite et d'assurance-vie ont occupé une place prépondérante dans les récents débats à l'échelle nationale sur des questions telles que les soins de santé, les personnes âgées et la réforme fiscale, auxquels ont pris part les membres du congrès, d'autres décisionnaires et de nombreux chercheurs. La SIPP permet de savoir si une personne participe à un régime d'assurance-maladie et si l'employeur y contribue, mais le répondant n'a à préciser ni le montant de ses cotisations ni celui des contributions de l'employeur. Pour ce qui est de l'assurance-vie, les seuls renseignements obtenus portent sur les garanties, le capital assuré et si la police est ou non offerte par l'intermédiaire de l'employeur. Aucune information n'est recueillie sur le montant des primes de l'employé et des contributions de l'employeur.

Pour l'étude de faisabilité en question, il s'agissait d'obtenir une autorisation écrite du répondant au moment de l'interview, d'entrer en communication avec son employeur et de lui demander de remplir un questionnaire abrégé afin de pouvoir obtenir des données sur les contributions tant de l'employeur que de l'employé à un régime d'assurance-maladie, à

un régime de retraite ou à une police d'assurance-vie. L'information ainsi obtenue pouvait compléter les données SIPP.

Pour réaliser l'étude, on a choisi la moitié de l'échantillon des ménages d'un groupe de renouvellement. Le test a eu lieu en août 1987 (groupe de renouvellement 4) au cours de l'interview de la vague 8 du panel de 1985. Il s'agissait de la dernière interview menée auprès de ces ménages.

Le test portait uniquement sur les personnes occupées, âgées de 18 ans et plus, ayant répondu au questionnaire SIPP de la vague 8. Sur les 1,352 personnes admissibles, 569 (42%) ont signé la formule d'autorisation, 446 (33%) ont refusé de signer et 337 enquêtés-substituts ou répondants interviewés par téléphone (25%) n'ont pas retourné la formule d'autorisation qui leur a été laissée ou postée. Aucun suivi n'a été fait dans les cas de refus ou de non-retour des formules.

Sur les 569 questionnaires qui ont été postés à un employeur, 548 (96%) ont été remplis et retournés. Une évaluation plus détaillée des données recueillies dans le cadre de cette étude sera entreprise l'an prochain, et l'on analysera la possibilité d'étendre l'étude à l'ensemble de l'échantillon.

3. Étude de vérification des enregistrements SIPP

Une autre avenue explorée dans le cadre de la recherche relative au système de dossiers administratifs est la réalisation d'études de validation des éléments communs aux deux sources, c.-à-d. aux données d'enquête et aux données administratives. L'étude de vérification des enregistrements SIPP a pour objet d'évaluer les divers aspects liés à la qualité des réponses fournies dans le cadre de l'enquête au moyen d'une comparaison, cas par cas, des données SIPP et de l'information contenue dans les dossiers administratifs. Une fois achevée, l'étude devrait permettre de mieux juger de la qualité des diverses données SIPP. De plus on pourra produire des estimations précises des erreurs de réponse et autres erreurs, afin que l'on puisse corriger les données d'enquête et même modifier les méthodes de collecte pour améliorer la qualité de ces données.

Moore et Marquis (1987) présentent une vue d'ensemble de l'étude et de l'état d'avancement des travaux. En gros, l'étude vise à évaluer les aspects suivants:

1. la qualité des données déclarées par les répondants pour indiquer qu'ils sont bénéficiaires de divers programmes de transfert administrés par l'État ou le gouvernement fédéral;
2. la qualité des données recueillies au sujet du montant des prestations versées en vertu de ces programmes;
3. les corrélations démographiques pouvant être établies pour mesurer la qualité des données;
4. l'importance des erreurs de classification;
5. les effets (non expérimentaux) sur la qualité des données déclarées qui sont imputables aux répondants eux-mêmes et aux enquêtés-substituts;
6. les effets dus au renouvellement des bénéficiaires entre les vagues d'interview (le problème d'un "point de jonction" (Burkhead et Coder, 1985; Moore et Kasprzyk, 1984)).

Pour l'évaluation de ces diverses questions, on se servira de l'information des fichiers administratifs concernant les bénéficiaires de neuf programmes de transfert gouvernementaux dans quatre États: la Floride, New York, la Pensylvanie et le Wisconsin. Quatre des neuf programmes à l'étude sont administrés par l'État (l'aide aux familles à faible revenu avec enfants à charge, les bons alimentaires, l'assurance-chômage et l'indemnisation des accidents du travail) et cinq sont administrés par le gouvernement fédéral (pensions de retraite des employés de la Fonction publique, subventions Pell,

prestations de sécurité de la vieillesse, de survivant et d'assurance-invalidité (OASDI), revenu supplémentaire de sécurité sociale et pensions et indemnisations des anciens combattants. Le projet a déjà permis de réunir une importante documentation sur l'acquisition des systèmes de dossiers administratifs, sur les particularités de chaque système et sur les méthodes générales de couplage utilisées au Census Bureau. Quelques résultats, qui restent cependant préliminaires, sont présentés dans Moore et Marquis (1987).

4. Utilisation des dossiers administratifs pour les estimations SIPP

Les effets connus de la réduction de la taille de l'échantillon sur la variance des estimations et sur notre capacité de mesurer les variations du nombre des paramètre statistiques sont devenus une source de préoccupation majeure. C'est ce qui nous a amenés à essayer de trouver de nouvelles façons de réduire cette variance. Une des solutions envisagées est l'utilisation de données de sources administratives pour une stratification a posteriori. Les méthodes d'estimation transversales utilisées présentement reposent sur un ajustement du second degré visant à accroître la précision des estimations. A cette fin, on procède à un ajustement par quotient des estimations pour les mois de collecte et les mois de référence en fonction des estimations de population. Cependant, le Census Bureau a accès à certains fichiers de l'IRS (Internal Revenue Service) et de la SSA (Social Security Administration) qui peuvent être utilisés pour la production de distributions détaillées selon l'âge, la race et le sexe du revenu brut ajusté. Nous venons tout juste de commencer à étudier de quelle façon ces données administratives pourraient servir à une stratification a posteriori visant à améliorer les estimations du revenu moyen et médian des particuliers et des ménages ainsi que les estimations des déciles de la distribution du revenu des particuliers et des ménages. Nous devons surtout nous demander quel degré de réduction de la variance des estimations on peut obtenir avec une telle méthode. Ce sont les deux questions que nous étudierons plus en détail au cours de l'année à venir.

La première phase de cette recherche (Huggins, 1987) permettra d'évaluer à quel point on peut réduire la variation des estimations SIPP en utilisant les données de l'IRS comme variables auxiliaires dans les méthodes d'estimation. La procédure à l'étude a été recommandée par Herriot (1985) et Scheuren (1983). Dans l'étude SIPP, la méthode d'estimation repose sur un rajustement par quotient des estimations SIPP qui est fait au second degré dans les cases définies selon l'âge + la race + le sexe + le "revenu", celui-ci étant le revenu brut ajusté, tel que déclaré à l'Internal Revenue Service.

Des totaux de contrôle sont établis à partir d'un échantillon de 1% des données du fichier 1984 de l'IRS en fonction des caractéristiques d'âge, de race et de sexe que l'on trouve dans le fichier SER; les données du revenu brut ajusté tirées du fichier intégral de l'IRS sont appariées aux données SIPP. Les données SIPP sont ensuite repondérées en fonction des totaux de contrôle du fichier 1984 de l'IRS. A cette fin, un facteur f_j qui est le rapport entre le total de contrôle IRS dans la case$_j$ et l'estimation SIPP du nombre de personnes après rapprochement avec les données de l'IRS de 1984 sur le revenu pour la case$_j$, est appliqué aux personnes incluses dans la case$_j$ selon les données de l'IRS. On utilisera ces nouveaux poids et les poids non fondés sur ces méthodes pour calculer les estimations de certaines caractéristiques SIPP et de leurs variances.

5. Fusion des données économiques et des données démographiques SIPP

Au cours des deux premières années de mise en oeuvre du programme SIPP, on a réalisé un assez grand nombre d'études sur la possibilité d'enrichir la base de données SIPP avec les données sur les établissements et les entreprises tirées du recensement économique et les données d'autres fichiers tenus par le Bureau of the Census (Haber, Ryscavage, Sater, Valdisera, 1984). Haber (1985) décrit les possibilités analytiques du

couplage des données économiques et des données démographiques se rapportant aux personnes participant à la SIPP. Selon Haber, un tel couplage offrirait de nouvelles possibilités d'analyse dans les domaines suivants : lien entre les investissements et les salaires; étude comparative de la mobilité des travailleurs les moins bien et les mieux rémunérés; étude des répercussions du passage d'une économie productrice de biens à une économie productrice de services; et analyse des effets des syndicats sur le marché du travail. On a aussi entrepris un projet pilote ayant pour objet la recherche de méthodes de couplage des répondants SIPP (qui indiquent le nom de leur employeur) avec les données sur les employeurs recueillies dans le cadre du recensement économique. On effectue également la vérification de ces méthodes afin de cerner les problèmes et de proposer des solutions et le couplage proprement dit d'un échantillon de données. Sater (1985) décrit le projet ainsi que les problèmes rencontrés. Malheureusement, en raison des coûts, de priorités plus grandes et du manque de personnel, ce projet n'a jamais été achevé.

3. COUPLAGES POSSIBLES AVEC D'AUTRES ENSEMBLES DE DONNÉES ADMINISTRATIVES

La SIPP est une enquête permanente relativement nouvelle dont les données servent à dresser un tableau socio-économique détaillé des ménages américains. Comme nous l'avons déjà souligné, dans cette enquête, on insiste aussi beaucoup sur l'importance de bien déclarer le numéro de sécurité sociale (SSN). Ces deux éléments réunis sont ce qui fait la force de la base de données SIPP. Dans les années à venir, l'excellente variable d'appariement (le SSN) que fournit la SIPP pourrait servir au couplage des données de l'enquête et des données du fichier principal de l'assurance-maladie (Medicare) de la Health Care Financing Administration en vue de l'étude de la relation entre l'utilisation des services hospitaliers, l'état de santé, l'emploi et le revenu. De la même façon, le SNN pourrait servir de lien entre les répondants décédés et l'indice national des décès. Dans ce dernier cas, de nombreux panels SIPP seraient nécessaires, de façon à ce qu'on puisse disposer d'un échantillon de taille suffisamment grande pour les besoins de l'analyse. Quoi qu'il en soit, ces possibilités existent bel et bien. De fait, tout couplage avec un système de dossiers administratifs pour lequel le SSN est utilisé comme principal identificateur est possible. Les obstacles les plus importants sont toutefois le coût de tels projets et la difficulté que pose le partage entre tous les chercheurs des données administratives et des données d'enquête ainsi couplées.

BIBLIOGRAPHIE

Burkead D., et Coder J. (1985). "Gross Changes in Income Recipiency from the Survey of Income and Program Participation", *Proceedings of the Social Statistics Section, American Statistical Association*, 351-356.

Goudreau, K., Oberheu, H., et Vaughan, D. (1981). "An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Program", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 377-382.

Goudreau, K., Oberheu, H., et Vaughan, D. "An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Programm", *Journal of Business and Economic Statistics*, 179-186.

Griffith, J., et Kasprzyk, D. (1980). "The Use of Administrative Records in the Survey of Income and Program Participation", étude présentée dans *Report on Statistical Uses of Administrative Records; Statistical Policy Working Paper 6*, U.S. Government Printing Office, Washington, D.C. 20402.

Haber, S., Ryscavage, P., Sater, D., et Valdisera, V. (1984). "Matching Economic Data to the Survey of Income and Program Participation: A Pilot Study", *Proceedings of the Social Statistics Section, American Statistical Association*, 529-533.

Haber, S. (1985). "Applications of a Matched File Linking the Bureau of the Census Survey of Income and Program Participation and Economic Data ", *SIPP Working Paper Series No. 8502*, U.S Bureau of the Census, Washington, D.C.

Herriot, R. (1983). "The Use of Administrative Records in Social and Demographic Statistics", document présenté dans le cadre de la réunion de l'*Institut* international de statistique , Madrid, Espagne.

Huggins, V. (1987). *Research Plans. Memorandum for the Record, April 12, 1987*, Statistical Methods Division, U.S. Bureau of the Census.

Kasprzyk, D. (1983). "Social Security Number Reporting, the Use of Administrative Records, and the Multiple Frame Design in the Income Survey Development Program, dans *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program (ISDP)*, M. David (éditeur), 123-141, New York: Social Science Research Council.

Kilss, B., et Scheuren, F. (1978). "The 1973 CPS-IRS-SSA Exact Match Study", *Social Security Bulletin*, vol. 41, No. 10, 14-22.

Lininger, C. (1980). "The Goals and Objectives of the Survey of Income and Program Participation", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 480-485.

Moore, J., et Kasprzyk, D. (1984). "Month-to-Month Recipiency Turnover in the ISDP", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 726-731.

Moore, J., et Marquis, K. (1987). "Utilisation des données des dossiers administratifs pour l'évaluation de la qualité des estimations d'enquêtes, communication présentée dans le cadre du Symposium international sur les utilisations statistiques des données administratives, 23-25 novembre 1987, Ottawa, Canada.

Sater, D.K. (1985). "Enhancing Data from the Survey of Income and Program Participation with Data from Economic Census and Surveys", *SIPP WORKING Paper Series No. 8505*, U.S. Bureau of the Census, Washington, D.C.

Sater, D.K. (1986). *SSN Responses Rates and Results of SSN Validation/Improvement Operation*, rapport présenté à Roger Herriot, le 11 mars 1986, Population Division, U.S. Bureau of the Census.

Scheuren, F., Herriot, R., Vogel, L. Vaughan, D., Kilss, B., Tyler, B. Cobleigh, C., et Alvey, W. (1975). "Report No 4: Exact Match Research using the March 1973 Current Population Survey--Initial States", *Studies from Interagency Data Linkages*, U.S. Department of Health, Education, and Welfare, Social Security Administration, Office of Research and Statistics, Department of Health, Education and Welfare, publication No. SSA 76-11750.

Scheuren, F. (1983). "Design and Estimation for Large Federal Surveys Using Administrative Records", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 377-381.

Vaughan, D. (1978). "Errors in Reporting Supplemental Security Income Recipiency in a Pilot Household Survey", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 288-293.

Yeas, M., et Lininger, C. (1981). "The Income Survey Development Program: Design Features and Initial Findings", *Social Security Bulletin*, Vol. 44, No. 11, 13-19.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

MODÉLISATION DE SÉRIES CHRONOLOGIQUES POUR L'ÉTABLISSEMENT D'ESTIMATIONS RÉGIONALES

G.H. CHOUDHRY et L.A. HUNTER[1]

RÉSUMÉ

La plupart des organismes statistiques gouvernementaux comme Statistique Canada administrent des enquêtes permanentes et offrent ainsi des données chronologiques sur les petites régions. Comme le degré de fiabilité des données régionales n'est pas satisfaisant en règle générale, on peut envisager la modélisation de séries chronologiques afin d'obtenir des estimations lissées, qui auront une plus grande fiabilité. L'information supplémentaire utilisée dans le modèle de séries chronologiques est fondée sur les données locales, qui comprennent des estimations d'enquête, et les données tirées du dernier recensement et de dossiers administratifs. En nous servant des données de l'enquête sur la population active du Canada (EPA), nous avons fait une modélisation de séries chronologiques avec structure d'erreur autorégressive afin d'obtenir des estimations lissées du chômage pour les divisions de recensement.

1. INTRODUCTION

Depuis quelques années, on insiste de plus en plus pour obtenir des données régionales plus récentes et plus fiables à cause de leur rôle dans l'élaboration des programmes, la répartition des fonds publics et la planification régionale. Statistique Canada a réagi aux demandes des utilisateurs en mettant sur pieds un programme de développement des données régionales. Brackstone (1986) s'intéresse aux questions entourant le développement et la production de données régionales.

Les estimations établies par sondage ne suffisent pas à satisfaire toutes les demandes de données régionales puisque la taille d'échantillon requise dans les enquêtes par sondage pour obtenir des estimations régionales fiables suppose des coûts et un fardeau de réponse trop élevés. Il est donc important d'élaborer des méthodes d'estimation qui combinent à l'aide de modèles statistiques des données provenant d'enquêtes par sondage, de dossiers administatifs et du dernier recensement. Parmi ces méthodes, on note les estimateurs synthétiques (voir, par exemple, Gonzalez et Hoza (1978) et Ghangurde et Singh (1978)) et les estimateurs dépendants de la taille de l'échantillon, définis par Drew, Singh et Choudhry (1982), et Hidiroglou et Särndal (1985). Toutefois, aucune de ces méthodes n'utilise l'information découlant de la corrélation entre les données chronologiques pour produire des estimations régionales plus fiables. Dans cet article, nous utilisons des

[1] G.H. Choudhry et L.A. Hunter, Division des méthodes d'enquêtes sociales, Statistique Canada, 4ième étage, Section C-1, Édifice Jean Talon, Ottawa, Ontario. K1A 0T6.

données chronologiques transversales groupées dans un modèle de régression avec structure d'erreur autorégressive. En nous servant de données de l'enquête sur la population active du Canada (EPA), nous calculons des estimations lissées du niveau de chômage pour des divisions de recensement. L'information supplémentaire utilisée dans le modèle repose sur les données locales, qui comprennent des estimations d'enquête, les données du dernier recensement ainsi que des données administratives de l'assurance-chômage. Comme prévu, la modélisation de séries chronologiques produit des estimations plus fiables que celles obtenues par des méthodes qui n'exploitent pas la dimension temporelle.

L'article est divisé de la façon suivante. Nous décrivons le modèle de base à la section 2 et la méthode d'estimation de ses paramètres à la section 3. Ensuite, dans la section 4 nous évaluons le modèle pour ce qui a trait à l'estimation du niveau de chômage dans les divisions de recensement. Dans la section 5, nous considérons diverses façons de mettre à jour le modèle et enfin dans la section 6, nous tirons quelques conclusions.

2. MODÈLE DE BASE

La statistique utilise habituellement des données transversales qui décrivent de nombreuses unités distinctes (par exemple, des petites régions) à une période donnée tandis que l'économétrie utilise en règle générale des données chronologiques qui décrivent une seule unité. Puisque la plupart des grandes enquêtes sont des enquêtes continues, il existe également une base de données chronologiques transversales pour les petites régions. On peut agréger ces données pour obtenir des estimations lissées pour les petites régions. L'agrégation consiste à combiner des données transversales et des données chronologiques sur les variables, afin d'estimer les coefficients du modèle de régression. Bien que l'on puisse estimer un modèle de régression pour chaque région ou secteur, l'agrégation de données de plusieurs régions produit une estimation plus efficace du modèle si on peut définir un modèle commun pour toutes les petites régions avec, peut-être, des ordonnées à l'origine différentes. Dielman (1983) s'est intéressé à ce genre de modèle pour données transversales et chronologiques combinées. Dans le présent exposé, nous ajustons un modèle de régression à des données chronologiques transversales afin d'accroître le degré de fiabilité des estimations du niveau de chômage pour les divisions de recensement. Cronkite (1984) a utilisé des modèles de régression différents pour obtenir le niveau d'emploi et le niveau de chômage dans les États mais a élaboré un seul modèle pour l'estimation des mêmes variables au niveau des régions administratives en groupant de façon transversale des données chronologiques de manière à former un agrégat d'observations pour l'estimation. Binder et Dick (1987) ont appliqué le même genre de modélisation pour l'établissement d'estimations régionales dans le cadre de l'enquête sur les voyages des Canadiens.

Nous allons maintenant définir les éléments du modèle. Pour la petite région "a" à la période t, définissons ${}_aY_t$ comme l'observation pour la variable dépendante et ${}_aX_{kt}$ ($k=0, 1, \ldots, K$) comme les observations pour les variables explicatives. On compte K+1 variables de ce genre et k=0 correspond à la constante. Nous définissons aussi les variables auxiliaires ${}_aZ_{a'}$, qui sont égales à 1 si a est égal à a' et à 0 dans le cas contraire. Ainsi, le modèle de base est défini

$$ {}_aY_t = \sum_{k=0}^{K} \beta_k ({}_aX_{kt}) + \sum_{a'=2}^{A} \gamma_{a'} ({}_aZ_{a'}) + {}_aU_t, $$

$$ a = 1, 2, \ldots A $$
$$ t = 1, 2, \ldots T \qquad (2.1) $$

Les β et γ sont les coefficients de régression et ${}_aU_t$ représente les erreurs stochastiques. La structure d'erreur hypothétique pour le modèle (2.1) est

$$ {}_aU_t = \rho\ {}_aU_{t-1} + {}_a\varepsilon_t $$

où ρ est le coefficient d'autocorrélation, qui est censé être le même pour toutes les petites régions et toutes les périodes. De plus, nous supposons que les ${}_a\varepsilon$ sont indépendants et identiquement distribués avec une moyenne nulle et une variance ${}_a\sigma^2$ pour a=1, 2, ... A.

L'observation ${}_aY_t$ est le logarithme naturel de l'estimation (selon l'EPA) de la proportion de personnes en chômage par rapport à l'ensemble de la population. La conversion en logarithme accroît la symétrie et l'homoscédasticité des erreurs et atténue l'effet des valeurs extrêmes. Les observations ${}_aX_{kt}$ sont les (K+1) variables explicatives, y compris la constante.

Les K variables explicatives dépendent de la région, de la période ou des deux à la fois; autrement dit, les données administratives de l'assurance-chômage et les estimations de l'EPA dépendent à la fois de la région et de la période, les variables de recensement dépendent uniquement de la région et les variables auxiliaires "mois" et "année" dépendent évidemment de la période. Les variables auxiliaires ${}_aZ_{a'}$ sont définies pour tenir compte d'ordonnées à l'origine différentes pour les petites régions. Le terme d'erreur ${}_aU_t$ comprend l'erreur d'échantillonnage et reflète également la structure corrélée des estimations dans le temps à l'intérieur des petites régions. Toutes les variables, à l'exception des variables auxiliaires, sont exprimées comme des proportions ou des taux ou des fonctions de ceux-ci.

3. ESTIMATION

En premier lieu, le choix des variables s'est fait au moyen d'une analyse de régression échelonnée, sauf en ce qui concerne l'ordonnée à l'origine, le taux d'activité selon l'EPA et le logarithme de la proportion de population qui reçoit des prestations d'assurance-chômage, qui sont toutes des variables qui ont toujours été incluses dans le modèle. Le taux d'activité selon l'EPA a été inclus dans le modèle parce qu'il est un très bon prédicteur des fluctuations saisonnières du chômage. Pour obtenir des estimateurs efficaces des paramètres du modèle, nous avons appliqué la transformation suivante, proposée par Cochran et Orcutt (1949), pour le modèle (2.1).

$$ ({}_aY_t - \rho\ {}_aY_{t-1}) = \sum_{k=0}^{K} \beta_k\ ({}_aX_{kt} - \rho\ {}_aX_{k,t-1}) $$

$$ + \sum_{a'=2}^{A} \gamma^{*}_{a'}\ ({}_aZ_{a'}) + {}_a\varepsilon_t \qquad (3.1) $$

où $\gamma^{*}_{a'} = (1 - \rho)\ \gamma_{a'}$.

Nous avons pu définir la transformation ci-dessus pour $t \geq 1$ puisque nous diposions des données pour t=0. Si cela n'avait pas été le cas, nous aurions obtenu la tranformation pour t=1 en multipliant les deux membres de l'équation (2.1) par $(1 - \rho^2)^{\frac{1}{2}}$. Nous ne l'avons pas fait toutefois car cela aurait compliqué les calculs. Le modèle transformé (3.1) a été réécrit de la façon suivante

$$ {}_{a}y_t = \rho\ {}_{a}y_{t-1} + \sum_{K=0}^{K} \beta_k ({}_{a}X_t - \rho\ {}_{a}X_{t-1}) $$

$$ + \sum_{a'=2}^{A} \gamma^{*}_{a'}\ ({}_{a}Z_{a'}) + {}_{a}\varepsilon_t \qquad (3.2) $$

et nous avons calculé les estimateurs par les moindres carrés des paramètres du modèle à l'aide de la méthode de Gauss-Newton modifiée (Hartley, 1961). Nous désignerons ces estimateurs comme les estimateurs par les moindres carrés non pondérés. Nous nous sommes servis des résidus estimés pour calculer les estimateurs pour ${}_{a}\hat{\sigma}^2$ de ${}_{a}\sigma^2$ pour a = 1, 2, ..., A. Ensuite, nous avons déterminé les paramètres du modèle à l'aide des moindres carrés pondérés, les poids étant égaux à ${}_{a}\hat{\sigma}^{-2}$. Nous avons utilisé une fois de plus la méthode Gauss-Newton modifiée de Hartley pour estimer les paramètres du modèle. Nous désignerons les estimateurs correspondants comme les estimateurs par les moindres carrés pondérés. Ceux-ci ont permis d'estimer les résidus pour le modèle transformé (3.2) ainsi que les variances ${}_{a}\sigma^2$. À l'aide du test de Student, nous avons testé, pour chacune des petites régions, l'hypothèse selon laquelle la moyenne des résidus est nulle. Si, pour une petite région donnée, la variable t était significative à une seuil de 5%, nous introduisions la variable auxiliaire pour cette région dans le modèle et nous calculions les estimateurs par les moindres carrés pondérés en refaisant les mêmes opérations. Cette méthode d'estimation est très comparable à la méthode des moindres carrés généralisés d'Aitken.

Suivant Goldberger (1962), le meilleur estimateur linéaire non biaisé de ${}_{a}y_t$ est défini

$$ {}_{a}\tilde{y}_t = \sum_{k=0}^{K} \hat{\beta}_k\ {}_{a}X_{kt} + \sum_{a'=2}^{A} \hat{\gamma}_{a'}\ ({}_{a}Z_{a'}) $$

$$ + \hat{\rho}\ \{{}_{a}y_{t-1} - \sum_{k=0}^{K} \hat{\beta}_k\ {}_{a}X_{k,\,t-1} - \sum_{a'=2}^{A} \hat{\gamma}_{a'}({}_{a}Z_{a'})\}, \qquad (3.3) $$

où $\hat{\beta}$'s, $\hat{\gamma}$'s, et $\hat{\rho}$ sont les estimateurs par les moindres carrés des paramètres correspondants du modèle. Il convient de souligner que

$$ \hat{\gamma}_{a'} = \frac{\hat{\gamma}^{*}_{a'}}{1 - \hat{\rho}}\ . $$

L'estimateur défini par (3.3) est la somme de la composante systématique et de la composante chronologique. On constate une amélioration de l'estimateur traditionnel obtenu par une composante systématique grâce à l'addition de la composante chronologique. Il convient aussi de souligner que la modélisation de séries chronologiques ne sert pas à établir des prévisions au-delà des limites du modèle mais à calculer des estimations lissées ayant une plus grande fiabilité. Les estimateurs de la proportion de personnes en chômage selon le modèle sont calculés comme suit

$$ {}_a\tilde{p}_t = \text{Exp}\ ({}_a\tilde{y}_t) $$

et leur variance est définie par

$$ v({}_a\tilde{p}_t) = \text{Exp}\ (2\ {}_a\tilde{y}_t)\ v({}_a\tilde{y}_t) $$

pour a = 1, 2, ... A et t = 1, 2, ... T, où $v({}_a\tilde{y}_t)$ est la variance estimée de ${}_a\tilde{y}_t$. Le coefficient de variation (CV) de ${}_a\tilde{p}_t$ peut être calculé à l'aide de la formule suivante:

$$ CV({}_a\tilde{p}_t) = [\ v({}_a\tilde{y}_t)\ ]^{\frac{1}{2}}. $$

Ainsi, pour obtenir le CV de l'estimation de la proportion de personnes en chômage selon le modèle, il suffit d'extraire la racine carrée de la variance estimée de l'estimation de la variable dépendante du modèle, qui est le logarithme de la proportion de personnes en chômage. Le calcul de $v({}_a\tilde{y}_t)$ est décrit en annexe.

On obtient des estimations lissées du niveau de chômage en multipliant l'estimation de la proportion de personnes en chômage par les estimations démographiques postcensitaires pour les petites régions tirées de sources externes (Verma et coll., 1983).

Il est aussi possible de calculer la variance des estimations lissées du niveau de chômage moyen (trimestriel, annuel, etc.) en se servant de la matrice des variances-covariances des estimateurs ${}_a\tilde{y}_t$, t=1, 2, ...T, définie également en annexe.

4. ÉVALUATION DU MODÈLE

À cause des différences de conditions du marché du travail, nous avons estimé des modèles différents pour les cinq régions de recensement suivantes: i) Maritimes, ii) Québec, iii) Ontario, iv) Prairies et v) Colombie-Britannique. La sixième région, qui comprend le Yukon et les Territoires du Nord-Ouest, n'a pu être incluse dans notre analyse parce qu'elle n'est pas couverte par l'EPA. Un modèle de régression a été défini pour chacune des cinq régions à l'aide de trente-six mois de données chronologiques (janvier 1983 - décembre 1985). Les petites régions étaient en l'occurrence des divisions de recensement ou des groupes de divisions de recensement. Le nombre de petites régions pour lesquelles des données ont été groupées à des fins d'estimation dans chacune des régions de recensement figure dans le tableau 4.1.

La valeur R^2 qui est une mesure de la validité de l'ajustement du modèle, a été calculée pour les régressions pondérées et non-pondérées pour chacune des régions. Les données pertinentes figurent également dans le tableau 4.1. Nous constatons que la validité de l'ajustement est largement supérieure dans le cas de la régression pondérée à cause de la forte variation des variances estimées pour les petites régions. De plus, les valeurs élevées de R^2 pour la régression pondérée indiquent que le modèle proposé permet d'obtenir un ajustement précis des données.

Tableau 4.1

Valeurs de R^2 pour les régressions pondérée et non pondérée et nombre de petites régions par région de recensement

Région de recensement	Nombre de petites régions	Valeur de R^2 pour	
		Régression non-pondérée	Régression pondérée
1	43	76.1	97.2
2	42	69.7	96.6
3	37	70.2	97.6
4	44	69.3	98.3
5	21	64.4	97.7

Au moyen de la régression pondérée, nous avons calculé les estimations du niveau de chômage selon le modèle et avons évalué la cohérence de ces estimations par rapport aux estimations de domaines stratifiés à posteriori. Nous avons obtenu ce dernier genre d'estimations en corrigeant l'estimation de la population en chômage dans une petite région (selon l'EPA) par le rapport entre l'estimation postcensitaire de la population de cette région et l'estimation de cette même population selon l'EPA. L'estimateur pour les domaines stratifiés à posteriori est non biaisé sauf s'il s'agit d'un estimateur par quotient et encore là, le biais est négligeable pour de grands échantillons. À partir des estimations mensuelles de domaines stratifiés à posteriori, nous avons établi des estimations trimestrielles et annuelles de même que des estimations moyennes pour trois ans pour chacune des petites régions et les avons comparées aux estimations correspondantes du modèle. L'écart relatif absolu (ERA) entre les deux estimations mensuelles pour la région "a" a été défini de la façon suivante:

$$a(\text{ERA}) = \frac{100}{T} \sum_t \left| \frac{a\hat{u}_t - a\tilde{u}_t}{a\tilde{u}_t} \right|,$$

où $a\hat{u}_t$ et $a\tilde{u}_t$ représentent respectivement les estimations de domaines stratifiés à posteriori et les estimations du modèle pour la petite région a et la période t et T est le nombre de périodes. L'ERA moyen d'une région de recensement a ensuite été défini

$$\overline{\text{ERA}} = \frac{1}{A} \sum_a a(\text{ERA}).$$

Nous avons défini de la même façon des ERA moyens pour les estimations trimestrielles et annuelles et les estimations moyennes pour trois ans. Les ERA moyens pour les estimations mensuelles, trimestrielles et annuelles de même que les estimations moyennes pour trois ans sont reproduits dans le tableau 4.2 pour chacune des régions. Nous remarquons que dans chaque cas, l'ERA moyen décroît de façon monotone à mesure que s'accroît la période de calcul de la moyenne, ce qui dénote l'absence de biais dans les estimations fondées sur le modèle. En outre, comme nous l'avons indiqué dans la section

précédente, le test de Student pour les résidus tirés du modèle final a été non significatif à un seuil de 5% pour toutes les petites régions. Il n'y a donc aucune preuve que les estimations du modèle soient entachées d'une erreur systématique quelconque.

Tableau 4.2

ERA moyen entre les estimations de domaines stratifiés à posteriori et les estimations du modèle, par région estimations mensuelles, trimestrielles et annuelles et estimations moyennes pour trois ans

Régions de recensement	ERA moyen pour			
	estimations mensuelles	estimations trimestrielles	estimations annuelles	estimations moyennes pour trois ans
1	15.3	8.6	4.5	2.5
2	18.4	10.7	6.5	3.4
3	15.9	9.3	4.9	2.7
4	21.6	12.2	7.3	4.2
5	13.9	8.5	4.8	1.5

On réussit à accroître sensiblement le degré de fiabilité des estimations fondées sur le modèle en i) combinant des données transversales par groupe de petites régions pour l'estimation et en ii) utilisant l'information qui découle de la corrélation entre les termes de la série chronologique. Drew, Singh et Choudhry (1982) ont défini un estimateur dépendant de l'échantillon, qui est une combinaison linéaire de l'estimateur pour domaines stratifiés à posteriori et de l'estimateur synthétique; dans ce dernier cas, le poids de la composante synthétique dépend de l'échantillon. Si l'échantillon pour la petite région est de taille "suffisante", le poids de la composante synthétique est nul. À mesure que la taille de l'échantillon diminue, la composante synthétique prend de plus en plus d'importance. Le poids de cette composante devient égal à 1 lorsqu'il n'y a pas d'échantillon dans la petite région. Nous avons comparé les CV moyens des estimations moyennes pour trois ans fondées sur le modèle avec ceux des estimations dépendantes de l'échantillon. Les résultats de cette comparaison figurent dans le tableau 4.3. Nous constatons que l'utilisation d'un modèle en série chronologique réduit de plus de moitié les CV estimés des estimations dépendantes de l'échantillon pour les cinq régions. Le tableau 4.3 donne aussi le coefficient d'autocorrélation estimé pour chacune des régions avec l'écart type (e.t.) correspondant entre parenthèses. Les degrés d'autocorrélation élevés expliquent en partie la forte augmentation du degré de fiabilité des estimations du modèle lorsqu'on utilise des séries chronologiques.

Tableau 4.3

CV moyens pour les estimations moyennes pour trois ans fondées sur le modèle et les estimations dépendantes de l'échantillon et autocorrélation estimée ($\hat{\rho}$) par région

Région	$\hat{\rho}$(E.T.)	CV estimé pour		Rapport
		estimations fondées sur le modèle	estimations dépendantes de l'échantillon	
1	0.60(0.02)	3.0	7.2	0.4
2	0.56(0.02)	3.4	7.3	0.5
3	0.53(0.02)	2.8	6.6	0.4
4	0.49(0.02)	3.6	8.2	0.4
5	0.53(0.03)	2.4	5.8	0.4

5. MISE À JOUR DU MODÈLE

Si la méthode proposée dans ce document doit être utilisée régulièrement pour produire des séries chronologiques d'estimations lissées, il est essentiel de déterminer si le modèle demeure efficace avec le temps et s'il y a lieu de définir une formule de mise à jour du modèle. Si l'on juge la mise à jour nécessaire, il faut aussi déterminer à quelle fréquence elle se fera.

Pour répondre à ces questions, nous avons ajusté les "meilleurs" modèles pour trois périodes différentes (1981-1983, 1982-1984, 1983-1985) aux données de la région 5 (Colombie-Britannique). Nous avons ensuite appliqué les variables du modèle pour 1981-1983 et leurs coefficients estimés aux données de 1982-1984 et de 1983-1985 pour évaluer le degré de fiabilité des estimations lorsqu'on utilise le même modèle sans le mettre à jour. Il est possible de mettre à jour le modèle sans le redéfinir entièrement en conservant les variables originales mais en réestimant les coefficients de régression. Pour tester la validité de cette méthode, nous avons appliqué les variables définies dans le modèle de 1981-1983 aux données de 1982-1984 et de 1983-1985 et avons estimé de nouveaux coefficients de régression pour chacune de ces variables. Nous avons calculé l'ERA moyen pour les estimations selon chaque modèle et avons comparé ces valeurs entre elles afin de déterminer les conséquences des diverses options (sans mise à jour, avec mise à jour des coefficients, redéfinition du modèle) en choisissant à nouveau des variables. Les résultats pertinents figurent dans le tableau 5.1.

Tableau 5.1

Effets de la mise à jour du modèle sur les données de C.-B.

Modèle	ERA (moyenne sur trois ans)
Modèle ajusté pour 1981-1983	2.9
Mise à jour du modèle pour 1982-1984	
(1) Mêmes variables et mêmes coefficients que pour 1981-1983	5.1
(2) Mêmes variables que pour 1981-1983; coefficients réestimés	3.2
(3) Redéfinition de variables	2.7
Mise à jour du modèle pour 1983-1985	
(1) Mêmes variables et mêmes coefficients que pour 1981-1983	8.2
(2) Mêmes variables que pour 1981-1983; coefficients réestimés	2.7
(3) Redéfinition de variables	1.5

En l'absence de mise à jour, le modèle se dégrade très rapidement. Après un an, l'ERA a presque doublé et après deux ans, il a presque triplé par rapport à 1981-1983. Toutefois, si l'on se sert des mêmes variables mais que l'on met à jour les coefficients de régression à chaque année, la fiabilité du modèle s'en trouve largement accrue. L'ERA pour les données de 1983-1985 est comparable à l'ERA pour les données de 1981-1983. Si nous comparons l'ERA obtenu par la mise à jour du modèle avec celui obtenu par la redéfinition du modèle ("meilleur" modèle), nous constatons que même après deux années,

le premier n'est que légèrement supérieur au second. Ces résultats montrent que même s'il est nécessaire de mettre à jour le modèle si nous souhaitons l'utiliser longtemps, il n'est pas essentiel de le redéfinir entièrement à mesure que de nouvelles données paraissent.

On se préoccupe aussi des conséquences que peut avoir l'utilisation de données de recensement périmées dans le modèle. Les données de recensement relatives à l'activité ne sont recueillies que dans les recensements décennaux et ne sont publiées, en général, que deux ans plus tard. Autrement dit, les données de recensement utilisées dans le modèle peuvent remonter jusqu'à douze ans. Pour analyser l'effet que cela peut avoir sur les estimations, nous avons ajusté le "meilleur" modèle pour la C.-B. pour la période 1981-1983 à des données du recensement de 1971, puis à des données du recensement de 1981 et avons comparé les ERA obtenus dans chaque cas.

L'ERA moyen pour trois ans était de 3.3% pour les données du recensement de 1971 et de 2.9% pour celles de 1981. Ces chiffres semblent indiquer que l'ancienneté des données du recensement influe très peu sur la fiabilité du modèle. Les variables de recensement utilisées dans le modèle sont les proportions des catégories de population active dans chaque branche d'activité et ces proportions sont demeurées à peu près inchangées avec les années.

Même si l'analyse que nous venons de faire ne portait que sur des données de la C.-B., les résultats auraient vraisemblablement été similaires pour les autres régions.

6. CONCLUSIONS

En combinant des données régionales, la modélisation de séries chronologiques produit une estimation plus efficace que l'estimateur dépendant de l'échantillon parce qu'elle exploite la structure de corrélation présente dans la série chronologique. En outre, il n'y a rien qui laisse croire que les estimations du modèle sont entachées d'un biais systématique.

Des tests portant sur une région de recensement ont montré qu'il n'est pas nécessaire de redéfinir entièrement le modèle lorsque paraissent de nouvelles données. En effet, les variables qui ont été choisies pour le modèle à une période quelconque peuvent servir pendant un certain temps; il suffit de mettre à jour les coefficients correspondants. Des tests devraient être faits pour d'autres régions afin de vérifier ces résultats mais tout indique qu'ils aboutiraient aux mêmes conclusions.

Les variances ont été estimées suivant l'hypothèse que la variance demeure fixe dans le temps pour chacune des petites régions (c.-à-d., $_a\sigma^2_t = {}_a\sigma^2$ pour $a = 1, 2, \ldots, A$ et $t = 1, 2, \ldots, T$). Sur le plan de la recherche, on pourrait vérifier cette hypothèse à l'aide de méthodes fondées sur la méthode du "Jackknife" (Wu, 1986).

ANNEXE
Matrice des variances-covariances des estimations du modèle

Le modèle en série chronologique (2.1) peut être exprimé sous la forme matricielle:

$$\underset{\sim}{Y}(1,T) = \underset{\sim}{X}(1,T)\ \underset{\sim}{B} + \underset{\sim}{U}(1,T) \qquad \text{(A1)}$$

où $\underset{\sim}{Y}(1,T)$ est le vecteur formé de n $(=A \cdot T)$ observations de la variable dépendante pour les T périodes $t = 1, 2, \ldots, T$ et les A petites régions $a = 1, 2, \ldots, A$ et $\underset{\sim}{X}(1,T)$ est la matrice correspondante des observations des variables explicatives, y compris l'ordonnée à l'origine et les variables auxiliaires. $\underset{\sim}{B}$ est le vecteur des paramètres inconnus du modèle, sauf en ce qui concerne le paramètre d'autocorrélation. $\underset{\sim}{U}(1,T)$ est le vecteur des erreurs stochastiques qui dépendent du coefficicient d'autocorrélation ρ.

On supprime l'autocorrélation par transformation linéaire en se servant des observations à la période $t > 0$ pour chacune des petites régions. La transformation linéaire est définie par la matrice de transition R_1, de dimension $T \times (T+1)$ elle-même définie $\underset{\sim}{R}_1$, $R_1(t,t) = -\rho$, $R_1(t,t+1) = 1$ pour $t = 1, 2, \ldots, T$, et zéro dans les autres cas.

Définissons maintenant $\underset{\sim}{R} = \underset{\sim}{I}_A \otimes \underset{\sim}{R}_1$, où $\underset{\sim}{I}_A$ est matrice d'identité de dimension A (nombre de petites régions) et $\otimes$ est le produit de Kronecker. Nous pouvons alors écrire sous la forme suivante le modèle transformé (4.3)

$$\underset{\sim}{R}\ \underset{\sim}{Y}(0,T) = \underset{\sim}{R}\ \underset{\sim}{X}(0,T)\ \underset{\sim}{B} + \underset{\sim}{E} \qquad \text{(A2)}$$

où $\underset{\sim}{E} = \underset{\sim}{R}\underset{\sim}{U}(0,T)$, $\underset{\sim}{Y}(0,T)$, $\underset{\sim}{X}(0,T)$, et $\underset{\sim}{U}(0,T)$ étant définis de la même façon que $\underset{\sim}{Y}(1,T)$, $\underset{\sim}{X}(1,T)$, et $\underset{\sim}{U}(1,T)$. La matrice des variances-covariances des erreurs du modèle transformé (A2) est définies par $\underset{\sim}{\Sigma} = \underset{\sim}{\sigma} \otimes \underset{\sim}{I}_T$ où $\underset{\sim}{\sigma}$ est la matrice diagonale de dimension A avec a-l'élément diagonal a étant égal à $_a\sigma^2$, et $\underset{\sim}{I}_T$ est la matrice unité de dimension T.

Définissons maintenant $\underset{\sim}{X}^* = \underset{\sim}{\Sigma}^{-\frac{1}{2}}\ \underset{\sim}{R}\ \underset{\sim}{X}(0,T)$ et $\underset{\sim}{Y}^* = \underset{\sim}{\Sigma}^{-\frac{1}{2}}\ \underset{\sim}{R}\ \underset{\sim}{Y}(0,T)$. Nous avons alors la relation suivante $\hat{\underset{\sim}{B}} = (X^{*\prime}\ X^*)^{-1}\ (X^{*\prime}\ Y^*)$. Signalons que la matrice de transition $\underset{\sim}{R}$ est évaluée à $\rho = \hat{\rho}$.

Définissons $\tilde{\underset{\sim}{Y}}(1,T)$ comme le vecteur des estimations du modèle définies en (3.3); nous pouvons alors exprimer $\tilde{\underset{\sim}{Y}}(1,T)$ de la façon suivante:

$$\begin{aligned}\tilde{\underset{\sim}{Y}}(1,T) &= \underset{\sim}{R}\ \underset{\sim}{X}(0,T)\ \hat{\underset{\sim}{B}} + \underset{\sim}{Y}(1,T) - \underset{\sim}{R}\ \underset{\sim}{Y}(0,T) \\ &= \underset{\sim}{R}\ \underset{\sim}{X}(0,T)\ \hat{\underset{\sim}{B}} + \rho\ \underset{\sim}{Y}(0,T-1)\end{aligned}$$

où $\underset{\sim}{Y}(0,T-1)$ est défini comme $\underset{\sim}{Y}(1,T)$.

Or,

$\text{var}\ (\hat{\underset{\sim}{B}}) = (\underset{\sim}{X}^{*'}\ \underset{\sim}{X}^{*})^{-1}$, et

$$\text{var}(\underset{\sim}{Y}^{(0,T-1)}) = \text{var}(\underset{\sim}{U}^{(0,T-1}) = \frac{1}{1-\rho^2}\ \text{var}(\underset{\sim}{E}) = \frac{1}{1-\rho^2}\ \underset{\sim}{\Sigma}\ .$$

et la matrice des variances-covariances de $\underset{\sim}{Y}(1,T)$ peut être exprimée selon une formule simple

$$\text{var}(\tilde{\underset{\sim}{Y}}(1,T)) = \underset{\sim}{\Sigma}^{1/2}\ \underset{\sim}{X}^{*}\ (\underset{\sim}{X}^{*'}\ \underset{\sim}{X}^{*})^{-1}\ \underset{\sim}{X}^{*'}\ \underset{\sim}{\Sigma}^{1/2} + \frac{\rho^2}{1-\rho^2}\ \underset{\sim}{\Sigma}\ .$$

Soit $_{a}\tilde{\underset{\sim}{Y}}(1,T)$ l'estimation du modèle pour la a-ième petite région, alors

$$\text{var}(_{a}\tilde{\underset{\sim}{Y}}(1,T)) = \{_{a}\underset{\sim}{X}^{*}\ (\underset{\sim}{X}^{*'}\ \underset{\sim}{X}^{*})^{-1}\ _{a}\underset{\sim}{X}^{*'} + \frac{\rho^2}{1-\rho^2}\ \underset{\sim}{I}_{T}\}\ _{a}\sigma^2\ ,$$

où $_{a}\underset{\sim}{X}^{*}$ est la sous-matrice de $\underset{\sim}{X}^{*}$ qui correspond à la a-ième petite rétion et $\underset{\sim}{I}_{T}$ est la matrice d'identité de dimension T. On estime la variance de $_{a}\tilde{\underset{\sim}{Y}}(1,T)$ en remplaçant ρ et $_{a}\sigma^2$ dans l'équation ci-dessus par les estimateurs correspondants.

REMERCIEMENTS

Les auteurs tiennent à exprimer leur reconnaissance à des collègues de Statistique Canada pour les nombreux commentaires utiles que ceux-ci leur ont fourni.

BIBLIOGRAPHIE

Binder, D.A., et Dick, J.P. (1987). "Estimation and Modelling in Repeated Surveys," document de travail interne, Division des méthodes d'enquêtes sociales, Statistique Canada.

Brackstone, G.J. (1986). "Small Area Data: Policy Issues and Technical Challenges," dans *Small Area Statistics, An International Symposium* (colligé par R. Platek, J.N.K. Rao, C.E. Sarndal, et M.P. Singh). John Wiley and Sons, 3-20.

Cochran, W.G., et Orcutt, G.H. (1949). "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 44, 32-61.

Cronkite, F.R. (1984). "A Proposed New System for Developing State and Area Employment and Unemployment Estimates: An Overview," rapport technique interne, Bureau of the Labour Statistics, Washington, D.C.

Dileman, T.E. (1983). "Pooled Cross-Sectional and Time Series Data: A Survey of Current Statistical Methodology," *The American Statistician*, 37, 111-122.

Drew, J.D., Singh, M.P., et Choudhry, G.H. (1982). "Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active au Canada", *Techniques d'enquête*, volume 8, 19-52.

Ghangurde, P.D., et Singh, M.P. (1978). "Evaluation of Efficiency of Synthetic Estimates," *Proceedings of the American Statistical Association, Social Statistics Section*, 52-61.

Goldberger, A.S. (1962). "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, 57, 369-375.

Gonzalez, M.E., et Hoza, C. (1978). "Small Area Estimation with Application to Unemployment and Housing Estimates," *Journal of the American Statistical Association*, 73, 7-15.

Hartley, H.O. (1961) "The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares," *Technometrics*, 3, 269-280.

Hidiroglou, M.A., et Sarndal, C.E. (1985). "Étude empirique de quelques estimateurs de régression pour petits domaines" *Techniques d'enquête*, vol. 11, 73-85.

Verma, R.B.P., Basavarajappa, K.G., et Bender, R. (1983). "Estimation par régression de la population à l'échelon infraprovincial au Canada", *Techniques d'enquête*, vol. 9, no. 2, .

Wu, C.F.J. (1986). "Jackknife, Bootstrap and other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14, 1261-1294.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

UNE SITUATION INVERSÉE: L'IMPUTATION DE VALEURS À DES ÉLÉMENTS D'INFORMATION MANQUANTS LORSQUE LES DONNEURS SONT RARES

JOHN L. CZAJKA[1]

RÉSUMÉ

On a généralement recours à la méthode "hot deck" ou à d'autres méthodes d'imputation de valeurs à des éléments d'information manquants dans les cas où il y a un rapport élevé entre les donneurs potentiels et les enregistrements pour lesquels une valeur doit être imputée. Le U.S. Internal Revenue Service a toutefois eu recours dans le cadre de son programme de la statistique du revenu des sociétés à l'imputation conjuguée avec échantillonnage à deux degrés pour réduire les coûts de collecte des données supplémentaires requises pour contrôler les déclarations d'impôt sur le revenu des sociétés. Dans ce cas précis, les enregistrements pour lesquels une valeur doit être imputée sont jusqu'à neuf fois plus nombreux que les donneurs potentiels. Le présent exposé porte sur les problèmes posés par le manque de donneurs et examine plusieurs modifications apportées cette année à la méthodologie pour améliorer les estimations des statistiques des sociétés désagrégées en sous-groupes d'activités économiques. La méthodologie peut être appliquée à d'autres cas où les données administratives doivent subir un contrôle élaboré avant de servir à des fins statistiques.

1. INTRODUCTION

La conversion des données d'enregistrements administratifs en une forme utile à des fins statistiques peut être une entreprise considérable et coûteuse. Cette observation vaut certainement pour le travail de la Statistics of Income (SOI) Division du U.S. Internal Revenue Service (IRS), qui produit des fichiers de données analytiques en extrayant de l'information des déclarations d'impôt sur le revenu produites chaque année par les particuliers et les sociétés. Les enregistrements échantillonnés par la SOI Division font l'objet d'un contrôle élaboré, qui peut comporter la collecte et le traitement de données supplémentaires pouvant être tirées seulement des déclarations elles-mêmes.

Pour réduire un des éléments des coûts de production des fichiers de données dans son programme de la statistique du revenu des sociétés, l'IRS a utilisé une méthode combinant l'échantillonnage à deux degrés et l'imputation au lieu de contrôler tous les enregistrements. Ces procédures ont été appliquées à un ensemble d'éléments de revenu décomposables pour lesquels un contribuable indiquant un montant différent de zéro doit joindre au document principal une annexe détaillant les sources du montant. Les préposés

[1] John L. Czajka, chercheur principal, Mathematica Policy Research, Inc., 600 Maryland Ave., S.W., pièce 550, Washington, D.C., 20024, É.U.

au contrôle de la SOI Division qui vérifient les annexes constatent souvent qu'une partie de ce qui est déclaré, par exemple "Autres revenus", peut être reclassée sous une rubrique plus précise (par exemple "Revenus bruts" ou "Autres dividendes") qui donnerait plus d'information aux analystes. Ils contrôlent chaque zone de données, enlevant une partie de ce qui a été déclaré comme "Autres revenus" et l'ajoutant à une ou plusieurs autres zones. Au lieu de vérifier toutes les annexes jointes aux déclarations échantillonnées, l'IRS a procédé de la façon suivante pour un sous-ensemble d'annexes:

1. révision des annexes de toutes les déclarations de "très grandes entreprises",
2. révision des annexes d'un certain nombre de déclarations d'entreprises plus petites en l'occurence les annexes pour lesquelles la probabilité de modification résultant d'un contrôle est élevée par rapport aux autres déclarations des entreprises du même groupe d'activités économiques,
3. révision des annexes d'un sous-échantillon aléatoire des autres déclarations (20% de celles des entreprises financières et 10% de celles des entreprises non financières),
4. imputation des résultats des contrôles aux déclarations qui restent en utilisant comme donneurs les unités du sous-échantillon aléatoire.

Ces procédures ont d'abord été appliquées à sept annexes se rapportant aux années d'imposition 1981 et 1982. La méthode employée dans ces deux applications a été décrite par Hinkins (1983, 1984), Czajka (1986, 1987) et Hinkins et Scheuren (1986). Après une interruption de deux ans, un plan d'imputation modifié a été appliqué à trois annexes se rapportant à l'année d'imposition 1985 (Czajka, 1987).

L'approche générale est applicable à d'autres programmes de données d'enregistrements administratifs. En ces temps de resserrement des budgets gouvernementaux, beaucoup d'organismes qui produisent des fichiers de données administratives à des fins analytiques pourraient gagner à appliquer des techniques analogues à leurs propres programmes de données. Le présent exposé porte sur un problème technique inhérent à l'utilisation d'une méthode d'imputation pour compenser le manque de données causé par l'échantillonnage à deux degrés: autrement dit, pour compenser le fait qu'on a un petit nombre de donneurs par rapport au nombre d'enregistrements pour lesquels on doit imputer une valeur.

2. UTILISATION D'UNE MÉTHODE D'ÉCHANTILLONNAGE À DEUX DEGRÉS CONJUGUÉE AVEC UNE MÉTHODE D'IMPUTATION POUR RÉDUIRE LE CONTRÔLE

Comme on peut le voir dans Cochran (1977), l'échantillonnage à deux degrés consiste notamment à tirer un échantillon préliminaire pour obtenir de l'information sur les caractéristiques nécessaires pour stratifier l'échantillon principal. Dans l'application de l'IRS, l'échantillon préliminaire est aussi l'échantillon primaire, qui est sous-échantillonné pour obtenir les renseignements supplémentaires de certains types de déclaration. L'information supplémentaire est ensuite utilisée pour contrôler les zones enregistrées de toutes les déclarations de l'échantillon primaire. L'échantillonnage à deux degrés réduit le volume des données à recueillir pour contrôler ces zones. En imputant des valeurs aux éléments d'information manquants dans les déclarations qui n'ont pas été choisies pour faire partie du sous-échantillon, on peut produire des micro-enregistrements complets pour toutes les déclarations de l'échantillon. Toutefois, cette nouvelle façon de procéder par rapport à l'utilisation habituelle de l'échantillonnage à deux degrés produit aussi des conditions inhabituelles à l'utilisation de l'imputation.

2.1 Échantillonnage à deux degrés pour obtenir l'information nécessaire au contrôle

On peut illustrer le rôle de l'échantillonnage à deux degrés dans le programme de la statistique du revenu des sociétés de la SOI Division en montrant comment la révision de l'annexe "Autres revenus" peut influer sur le montant final inscrit à l'élément "Autres revenus" dans le fichier analytique. On peut obtenir la valeur finale de l'élément "Autres revenus" d'une entreprise de l'échantillon à l'aide de l'équation suivante:

$$Y_i = B_i - C_i$$

où Y_i représente le montant final, B_i, le montant initial, et C_i, le changement. Sans tenir compte de la pondération, on peut donc exprimer l'estimateur de l'élément "Autres revenus" d'un agrégat d'entreprises de l'échantillon comme suit:

$$\Sigma Y_i = \Sigma B_i - \Sigma C_i$$

L'estimation agrégée correspond au montant total initial auquel on soustrait le montant total enlevé. Par conséquent, si on ne fait pas de contrôle, ΣC_i représente le biais agrégé.

Étant donné qu'on a procédé à un échantillonnage à deux degrés, on n'observe C_i que pour un cinquième des entreprises financières sous-échantillonnées et un dixième des entreprises non financières. Pour estimer l'élément "Autres revenus" (ou n'importe quel des autres éléments d'information touchés par la révision de l'annexe "Autres revenus") d'une population d'entreprises sous-échantillonnées, on pourrait multiplier chaque changement observé par cinq, s'il s'agit d'une déclaration appartenant à une entreprise financière, et par dix, s'il s'agit d'une déclaration appartenant à une entreprise non financière. La repondération est une méthode d'estimation couramment utilisée quand on a recours à l'échantillonnage à deux degrés. Toutefois, le traitement de la non-réponse par repondération n'est pas la meilleure méthode à utiliser quand on construit un fichier de microdonnées qui pourrait servir à d'autres fins qu'à produire des totalisations agrégées. En outre, dans ce dernier cas, on publie des estimations agrégées pour plus de mille sous-populations. Pour beaucoup de sous-populations, la repondération aboutirait à des estimations des montants finals qui, quoique sans biais, auraient une très grande variance. Par contre, la méthode d'imputation de la valeur C_i non observée qu'on propose produit des micro-enregistrements complets qui augmentent de beaucoup l'utilité du fichier de données. De plus, en procédant par imputation, on peut plus facilement trouver une solution au problème de la variance des estimations qui risque d'être grande dans des sous-populations très petites.

Les résultats des imputations de valeurs aux éléments "Autres revenus", "Autres déductions" et "Coûts des biens vendus" sont résumés dans le tableau 1. Au total, pour ces trois éléments d'information, l'utilisation d'une méthode d'échantillonnage à deux degrés et d'une méthode d'imputation a remplacé le contrôle de 57,136 annexes. Des changements ont été imputés à 25.9 pour cent des éléments d'information.

Tableau 1

Résumé des Imputations Faites en 1985 aux Éléments "Autres Revenus", "Autres Déductions" et "Coût des Biens Vendus"

Élément	nombre d'enregistrements retenus pour imputation	Enregistrements pour lesquels le montant initial a été changé par imputation	Enregistrements pour lesquels le montant initial n'a pas été changé par imputation	Proportion des enregistrements pour lesquels le montant initial a été changé par imputation
Autres revenus	18,657	4,149	14,508	22.2%
Autres déductions	23,719	7,662	16,057	32.3%
Coût des biens vendus	14,760	2,982	11,778	20.2%
Total	57,136	14,793	42,343	25.9%

NOTE: Tous les chiffres sont des valeurs non pondérées.

2.2 Imputation de valeurs aux éléments non contrôlés: le problème du manque de donneurs

Les méthodes d'imputation fondées sur la substitition de valeurs d'enregistrements complets pour les enregistrements incomplets sont très intéressantes parce qu'elles peuvent être automatisées en grande partie et parce qu'elles permettent d'imputer à plusieurs éléments d'information sur un même enregistrement des valeurs qui sont compatibles entre elles. La nécessité d'automatiser le processus et d'imputer des valeurs à plusieurs éléments d'information par enregistrement est dans l'un et l'autre cas caractéristique de l'application de la SOI Division et explique pourquoi on a utilisé une méthode "hot deck" modifiée pour imputer des valeurs aux enregistrements des sociétés en 1981 et 1982 (Hinkins, 1984).

En règle générale, on a recours à des méthodes de substitution quand le nombre d'enregistrements où manquent des données sur un élément d'information ou sur un groupe d'éléments d'information est faible par rapport au nombre d'enregistrements où les données sont complètes et où la taille de l'échantillon est très grande (de dix mille unités ou plus). Le grand nombre de donneurs potentiels permet de trouver des donneurs qui concordent assez bien avec les enregistrements où il manque des données et d'imputer des valeurs aux caractéristiques correspondant aux données manquantes. Toutefois, lorsque les données manquent en raison de l'échantillonnage à deux degrés, le rapport entre le nombre d'enregistrements où les données sont complètes et le nombre d'enregistrements où les données sont incomplètes est inversé. Dans l'application de l'IRS dont nous parlons ici, il y a quatre ou neuf fois plus d'enregistrements pour lesquels on doit imputer des valeurs que de donneurs potentiels, d'où la "situation inversée" dont fait mention le titre du présent exposé. Malgré la taille de l'échantillon des sociétés, le nombre de donneurs potentiels pour l'un ou l'autre des trois éléments d'information s'élève à deux à trois mille seulement.

La rareté des donneurs pose plusieurs problèmes quand on veut utiliser des méthodes d'imputation fondées sur la substitution. Outre la variance introduite par l'échantillonnage à deux degrés, trois problèmes se posent:

1. les chances qu'on fasse correspondre à chaque enregistrement un donneur qui concorde assez bien sont limitées par le petit nombre de donneurs;
2. la fréquence d'utilisation des donneurs risque d'être très inégale, ce qui augmente encore plus l'effet de l'imputation sur la variance des estimations;
3. le risque d'imputer des résultats peu plausibles est sensiblement accru.

Décrivons maintenant chacun de ces problèmes.

L'utilisation d'une des méthodes de substitution demande d'avoir beaucoup de donneurs potentiels pour permettre au processus d'appariement entre donneurs et receveurs (c'est-à-dire les enregistrements pour lesquels on doit imputer une valeur) d'incorporer plusieurs variables auxiliaires aux éléments d'information manquants. Quand on a peu de donneurs, on peut apparier les enregistrements en fonction d'un petit nombre seulement de caractéristiques regroupées en très peu de catégories. Pour illustrer le problème, le degré de concordance entre les éléments de chaque paire d'enregistrements, receveur et donneur, dans l'exemple de l'IRS est limité par le fait que chacun des deux à trois mille donneurs doit être apparié à environ 4 à 9 donneurs en moyenne. À un micro-niveau, l'erreur quadratique moyenne comportera donc un biais important ainsi qu'une grande variance. L'erreur attribuable à l'imputation risque de demeurer grande même à des niveaux d'agrégation très élevés.

Dans les applications habituelles des méthodes de substitution, c'est-à-dire quand le nombre de donneurs potentiels est beaucoup plus élevé que le nombre de receveurs, la plupart des donneurs servent une fois ou ne servent pas du tout; il arrive rarement qu'un donneur soit utilisé plus d'une fois. Toutefois, quand le rapport est inversé, la fréquence d'utilisation des donneurs risque de varier beaucoup d'un donneur à l'autre. Des écarts importants dans la fréquence d'utilisation des donneurs risquent fort d'augmenter la variance des estimations produites par imputation. Les procédures d'imputation utilisées par l'IRS en 1981 et 1982 incluaient un élément visant à minimiser la variation dans le degré d'utilisation des donneurs à l'intérieur des cases d'ajustement. Les donneurs étaient tirés séquentiellement similairement à un tirage effectué à partir d'un jeu de cartes mêlé une fois. Dans une même case d'ajustement, toutefois, tout donneur était utilisé au maximum une fois de plus que tout autre donneur. Afin d'avoir un effet sur la variation de la fréquence d'utilisation des donneurs, cette méthode exige que chaque case d'ajustement contienne plusieurs donneurs. Quand le nombre de donneurs est très faible, il faut restreindre de beaucoup le nombre de cases d'ajustement.

Un des problèmes que pose un rapport faible entre les donneurs et les enregistrements receveurs est faible, est le risque d'imputer des valeurs peu plausibles ou aberrantes. Les travaux de l'IRS fournissent des exemples pertinents. Dans le cas de l'application de l'IRS, les imputations peu plausibles incluent des changements qu'un préposé au contrôle est peu susceptible de faire. Il est peu probable que les préposés au contrôle enlèvent de très petites proportions de petits montants initiaux ou enlèvent de grandes proportions de grands montants initiaux. Cependant, il n'y a pas de règle absolue qui puisse être appliquée à toutes les procédures d'imputation ni qui puisse être incluse dans les tests de compatibilité auxquels les enregistrements receveurs doivent être soumis. Pour faire ses imputations en 1981 et 1982, l'IRS s'est fié au début à son plan de conception initial des cases d'ajustement pour s'assurer d'obtenir des valeurs d'imputation plausibles. Il s'est toutefois produit des cas en 1982 où l'on a apparié des donneurs pour lesquels on avait enlevé une petite partie d'un montant élevé de la catégorie "Autres revenus" et des enregistrements pour lesquels le montant initial de la catégorie "Autres revenus" était faible. Comme la valeur monétaire imputée est exprimée en variation relative plutôt qu'absolue, cette méthode d'imputation a produit des changements de valeurs intimes de l'ordre de moins d'un dollar dans plusieurs cas. Le faible rapport entre le nombre de

donneurs et le nombre receveurs a joué un rôle dans ce résultat en limitant le degré de ressemblance qui pouvait être obtenu entre chaque receveur et ses donneurs potentiels.

2.3 Les imputations de valeurs aux sociétés en 1985: une vue d'ensemble

En révisant les procédures d'imputation devant servir dans le programme de la statistique des sociétés de 1985, nous avons apporté plusieurs modifications pour résoudre un certain nombre de problèmes posés par la rareté des donneurs. Ces nouveaux éléments sont résumés dans les paragraphes qui suivent.

L'élément clé de l'approche révisée consiste à distinguer deux opérations, soit premièrement l'imputation ou non d'un changement (c'est-à-dire l'imputation d'un changement différent de zéro ou d'un changement nul) et, deuxièmement, l'imputation de la valeur conditionnelle du changement. Cette distinction des deux composantes de l'imputation par suite d'un contrôle est à la fois possible et efficace parce que le contrôle des donneurs révèle que dans bien des cas il n'y a pas de changement à apporter aux montants initiaux. (Il faut se rappeler que deux groupes de déclarations n'étaient pas soumis au sous-échantillonnage: les déclarations des grandes entreprises et les déclarations des entreprises pour lesquelles il était fort probable que le contrôle produirait un changement). Le fait de distinguer ces deux composantes de l'imputation nous a permis de régler en partie le cas du manque de donneurs de deux façons: (1) en imputant ou pas un changement à partir d'une matrice de probabilités et (2) en imputant la valeur conditionnelle du changement dans des cases d'ajustement définies de façon plus appropriée.

Lorsqu'il manque de donneurs, la possibilité d'imputer ou non un changement à partir d'une matrice de probabilités présente l'avantage d'éliminer la nécessité de compter sur un appariement explicite des enregistrements et des donneurs. Ainsi, on peut lisser les probabilités de changement de manière à réduire la variabilité due de l'échantillonnage. Cela permet de relâcher la contrainte de taille minimum acceptable des cases et donc d'inclure plus de variable auxiliaire ou d'augmenter le nombre de catégories pour une ou plusieurs dimensions de la matrice, ce qui aura pour effet de réduire le biais dû à l'imputation.

L'imputation de changements différents de zéro oblige tout de même à procéder à l'appariement d'enregistrements receveurs et donneurs. À cette étape, toutefois, on utilise seulement les donneurs avec changement (c'est-à-dire les donneurs auxquels on a apporté des changements), ce qui permet de définir des cases d'ajustement répondant aux critères de taille minimum. De cette façon, on peut contrôler la variance due à l'échantillonnage des changements imputés. En outre, dans la mesure où les variables auxiliaires de la **valeur** d'un changement différent de zéro diffèrent des covariables de la **probabilité** du changement en question, on peut diminuer le biais d'imputation en déterminant un ensemble de covariables différent de celui qu'on a utilisé pour définir la matrice de probabilités.

Pour diminuer le risque d'imputer des valeurs peu plausibles, on a apparié les donneurs et les receveurs figurant dans une même case d'ajustement en utilisant le logarithme du montant initial inscrit à l'élément concerné (c'est-à-dire à l'élément "Autres revenus", "Autres déductions" ou "Coût des biens vendus"). Cette solution directe au problème observé dans les imputations précédentes a un effet secondaire non intentionnel -- à savoir qu'elle entraîne une grande variation dans la fréquence d'utilisation des donneurs à l'intérieur d'une même case d'ajustement. Nous reparlerons de ce résultat un peu plus loin dans le texte.

Les modifications que nous venons de résumer ici sont maintenant expliquées plus en détail dans des parties distinctes ci-dessous et illustrées pour l'élément "Autres revenus".

3. IMPUTATION D'UN CHANGEMENT NUL OU D'UN CHANGEMENT DIFFÉRENT DE ZÉRO

Le choix binaire entre l'imputation à l'élément d'information "Autres revenus" d'un changement nul ou d'un changement différent de zéro a été effectué à partir d'une matrice de probabilités à trois dimensions représentant la classification des déclarations selon l'industrie, la taille des entreprises et le ratio Autres revenus/Revenu total. Beaucoup de cases étant très petites, on a lissé les probabilités calculées à partir de l'élément "Autres revenus" chez les donneurs pour diminuer l'effet de la variabilité due à l'échantillonnage. Les résultats ont ensuite été imputés au hasard à l'intérieur des cases de la matrice en fonction des probabilités lissées correspondantes. La matrice et l'algorithme de lissage sont décrits plus bas.

3.1 Dimensions de la matrice de probabilités

Pour déterminer quelles seraient les dimensions de la matrice de probabilités, on a accru le niveau de désagrégation industrielle et raffiné la classification des entreprises en fonction de leur taille par rapport à ce qui avait été utilisé en 1981 et 1982. On a aussi ajouté une nouvelle variable auxiliaire.

La classification des activités économiques se composait de 23 sous-groupes d'activités économiques obtenus à partir de sept groupes. (Voir Appendice.) Dans ses imputations précédentes, l'IRS ne reconnaissait que dix classes d'activités économiques. La désagrégation en un plus grand nombre de classes a été rendue possible par la décision d'imputer un changement différent de zéro ou un changement nul à partir de probabilités lissées.

Trois classes d'entreprises selon leur taille ont été définies pour les entreprises de chacun des 23 sous-groupes d'activités économiques, en fonction de la répartition des déclarations dans chaque sous-groupe selon les montants d'actifs et de revenu net indiqués dans les déclarations. En 1981 et 1982, l'IRS avait appliqué à toutes les déclarations une méthode de classement uniforme selon la taille. Cette révision était justifiée par le fait qu'on a observé des différences importantes dans la répartition des entreprises selon la taille.

Cinq autres catégories ont été définies en fonction du ratio Autres revenus/Revenu total de chaque groupe d'activités économiques. Ce ratio d'éléments de revenu est le critère utilisé par l'IRS pour déterminer si les déclarations seront classées dans la catégorie où les annexes "Autres revenus" devront être contrôlées plutôt que dans la catégorie où les annexes seront soumises à un sous-échantillonnage. Ce critère n'était cependant pas utilisé auparavant dans le processus d'imputation. Nous avons décidé que ce "ratio de sélection" serait une dimension de la matrice de probabilités puisque celle-ci semblait hautement liée avec la probabilité que l'élément Autres revenus ait été changé sur un enregistrement donneur en 1982.

Le tableau 2 présente les résultats des calculs de la probabilité moyenne qu'un changement ait été apporté à l'élément "Autres revenus" chez les donneurs classés selon chacune des trois dimensions. Les probabilités moyennes de changement selon la classe de taille et le ratio de sélection sont présentées séparément pour les entreprises financières et les entreprises non financières. Les probabilités varient beaucoup selon n'importe laquelle des trois dimensions. Pour ce qui est des catégories d'entreprises selon la taille et le ratio de sélection, il convient toutefois de noter que les probabilités de changement varient beaucoup moins pour les entreprises financières que pour les entreprises non financières.

Tableau 2

Probabilité qu'un changement soit apporté à l'élément "Autres Revenus" chez les enregistrements donneurs classés selon l'industrie, la taille et le ratio de sélection

Sous-groupe d'activités économiques			Taille de la déclaration			Ratio de sélection		
Classe	Prob. de change-ments	Nombre de donneurs	classe	Prob. de change-ments	Nombre de donneurs	classe	Prob. de change-ments	Nombre de donneurs
			Entreprises non financières					
1	.211	90	1	.087	391	1	.047	492
2	.194	72	2	.163	787	2	.121	381
3	.259	216	3	.195	780	3	.190	410
4	.239	109				4	.207	372
5	.421	152				5	.297	303
6	.158	57						
7	.068	59						
8	.176	119						
9	.145	76						
10	.103	194						
11	.083	108						
12	.077	117						
13	.133	75						
14	.088	137						
15	.095	21						
16	.080	200						
17	.052	77						
18	.101	79						
			Entreprises fiancières					
19	.958	289	1	.565	147	1	.552	181
20	.724	123	2	.638	168	2	.549	133
21	.260	50	3	.601	296	3	.596	151
22	.161	93				4	.714	119
23	.244	156				5	.661	127

3.2 Lissage de la matrice de probabilités

On a lissé les valeurs des cases de la matrice de probabilités 23x3x5 pour diminuer la variabilité due à l'échantillonnage. L'algorithme de lissage utilisait les proportions relatives observées d'enregistrements donneurs avec changement c'est-à-dire où un changement a été apporté et d'enregistrements donneurs sans changement c'est-à-dire où aucun changement n'a été apporté, par classe à l'intérieur de chacune des trois dimensions, pour calculer le nombre d'enregistrements avec changement prévu c'est-à-dire pour prédire le nombre d'enregistrements où un changement serait apporté et le nombre d'enregistrements sans changement prévu c'est-à-dire pour prédire le nombre d'enregistrements où aucun changement ne serait apporté **par case.** À partir de ces prédictions, on a pu calculer des probabilités de changement en divisant le nombre

d'enregistrements avec changement prévu par la somme des enregistrements avec changement prévu et des enregistrements sans changement prévu. La probabilité lissée dans chaque case a ensuite été calculée comme une somme pondérée de la probabilité observée et de la probabilité prédite, les poids étant une fonction du nombre de donneurs échantillonnés dans la case.

En termes explicites, l'algorithme de lissage peut être décrit comme suit. Supposons que N_{1ijk} représente le nombre d'enregistrements de la case (i,j,k) avec changement et que N_{0ijk} représente le nombre d'enregistrements sans changement. La proportion relative d'enregistrements de la classe de taille i=1 avec changement est:

$$SIZ_{11} = \sum_{jk} N_{11jk} / \sum_{ijk} N_{1ijk}$$

De même, la proportion relative d'enregistrements **sans changement** est:

$$SIZ_{01} = \sum_{jk} N_{01jk} / \sum_{ijk} N_{0ijk}$$

Parallèlement, on calcule les proportions relatives d'enregistrements avec changement par rapport au nombre total d'enregistrements des autres classes de taille et des enregistrements classés selon le ratio de sélection (SEL_{1j}) et la classification des activités économiques (IND_{1k}). De même, on calcule les proportions relatives d'enregistrement sans changement par rapport au nombre total d'enregistrements classés selon la taille, le ratio de sélection et la classification des activités économiques.

Soit N_1 le nombre total d'enregistrements avec changement. On peut exprimer le nombre d'enregistrements de la case (i,j,k) avec changement prévu de la façon suivante:

$$PRED_{1ijk} = N_1 * SIZ_{1i} * SEL_{1j} * IND_{1k}$$

et le nombre d'enregistrements sans changement prévu de la façon suivante:

$$PRED_{0ijk} = N_0 * SIZ_{0i} * SEL_{0j} * IND_{0k}$$

(À noter que la sommation de $PRED_{1ijk}$ sur i, j et k est N_1 et que, de même, la sommation de $PRED_{0ijk}$ sur i, j et k est N_0.) On calcule ensuite la probabilité prédite de changement dans chaque case comme suit:

$$PROB_{ijk} = PRED_{1ijk} \quad (PRED_{1ijk} + PRED_{0ijk})$$

Ces probabilités prédites sont utilisées conjointement avec les probabilités observées pour produire les probabilités lissées.

Quand la taille de la case égale ou dépasse une constante précisée à l'avance dans le programme d'imputation (on a utilisé la taille de 50 en 1985), on ne substitue pas de valeur lissée à la probabilité observée de changement. Par contre, quand la taille de la case est inférieure à la constante, on fait une somme pondérée des probabilités observées et prédites, où les poids sont une fonction de la racine carrée de la taille de la case. Plus précisément, le poids attribué à la probabilité observée est la racine carrée du rapport entre la taille de la case et la constante. Le poids attribué à la probabilité prédite est la différence entre un et ce rapport. Si par exemple la taille de la case est 25, la probabilité

lissée correspondrait à 70.7 pour cent de la probabilité observée plus 29.3 pour cent de la probabilité prédite. Si la taille de la case n'était que de 10, la probabilité lissée correspondrait à 44.7 pour cent de la probabilité observée plus 55.3 pour cent de la probabilité prédite. La diminution du poids attribué à la probabilité observée à mesure que la taille de la case diminue est inversement proportionnelle à l'accroissement de l'erreur-type de la probabilité observée.

Tableau 3

Lissage des probabilités de changement à l'élément "Autres Revenus": industries du commerce de gros

Classe de taille	Classe selon le ratio de sélection 1	2	3	4	5	Classe selon le ratio de sélection 1	2	3	4	5
	Sous-groupe d'activités économiques 1									
	Enregistrements avec changement/total					**Probabilités observées**				
1	0/4	0/4	0/2	0/4	0/0	.000	.000.	.000	.000	.000
2	1/11	2/9	0/4	3/9	3/4	.091	.222	.000	.333	.750
3	1/11	1/10	4/8	1/6	3/4	.091	.100	.500	.167	.750
	Probabilités prédites					**Probabilités lissées**				
1	.029	.082	.097	.167	.294	.021	.059	.078	.120	.294
2	.070	.183	.213	.336	.512	.080	.200	.153	.335	.580
3	.082	.210	.243	.375	.554	.086	.161	.346	.303	.610
	Sous-groupe d'activités économiques 2									
	Enregistrements avec changement/total					**Probabilités observées**				
1	0/8	0/3	0/4	1/2	0/1	.000	.000	.000	.500	.000
2	0/7	2/6	0/4	2/6	2/5	.000	.333	.000	.333	.400
3	1/2	2/8	1/7	1/6	2/3	.500	.250	.143	.167	.667
	Probabilités prédites					**Probabilités lissées**				
1	.026	.074	.088	.154	.274	.016	.056	.063	.223	.235
2	.064	.168	.196	.314	.487	.040	.226	.141	.320	.459
3	.075	.193	.224	.351	.529	.160	.216	.194	.287	.562
	Sous-groupe d'activités économiques 3									
	Enregistrements avec changement/total					**Probabilités observées**				
1	1/10	2/11	2/11	3/12	0/1	.100	.182	.182	.250	.000
2	2/25	4/16	7/25	8/14	3/11	.080	.250	.280	.571	.273
5	1/16	3/16	5/20	9/21	6/7	.062	.188	.250	.429	.857
	Probabilités prédites					**Probabilités lissées**				
1	.038	.104	.123	.208	.353	.066	.141	.151	.229	.303
2	.090	.227	.262	.398	.579	.083	.240	.275	.490	.435
3	.105	.258	.295	.439	.619	.081	.218	.267	.432	.708

Le tableau 3 présente les probabilités observées, prédites et lissées calculées pour les donneurs de la catégorie "Autres Revenus" des industries du commerce de gros. On donne également la fréquence initiale du totals des enregistrements et des enregistrements modifiés. L'éparpillement des effectifs dans les tableaux de fréquence, est caractéristique de la plupart des catégories d'industries importantes.

Les avantages du lissage sont évidents. Le très petit nombre de donneurs dans la plupart des cases fait qu'on a une grande variation due à l'échantillonnage des probabilités observées de changement. En lissant les probabilités observées de changement, on diminue de beaucoup l'effet de la composante du bruit dans la variation entre les cases, mais on continue quand même d'avoir une forte relation entre les probabilités de changement et les trois dimensions de la matrice.

Il est évident qu'on pourrait appliquer d'autres algorithmes de lissage, peut-être même avec plus de succès. L'algorithme décrit ici équivaut, quoique pas exactement, à ajuster un modèle log-linéaire composé des interactions à deux sens entre le résultat changement/aucun changement et chacune des trois covariables. Le lissage fondé sur l'ajustement d'un modèle log-linéaire sur les fréquences dans chaque case est une autre façon évidente de procéder qui mérite d'être considérée. Le fait qu'il soit souhaitable de faire des ajustements pour éliminer les sommes nulles dans certaines industries a été le principal élément qui nous a dissuadés d'explorer davantage cette approche quand nous avons élaboré la méthode d'imputation que nous voulions utiliser en 1985.

4. DÉFINITION DES CASES D'AJUSTEMENT POUR L'IMPUTATION DE VALEURS DIFFÉRENTES DE ZÉRO

Des recherches effectuées sur les statistiques des sociétés de 1982 ne nous ont pas permis de trouver de variables auxiliaires fortement liées aux valeurs conditionnelles de changement, ce qui contraste beaucoup avec ce que nous avons observé pour les probabilités de changement. Par conséquent, pour imputer des valeurs de changement nous avons pris les mêmes dimensions pour les cases d'ajustement que pour les matrices de probabilités, sauf que dans le cas des éléments "Autres déductions" et "Coût des biens vendus", nous n'avons pas repris le ratio de sélection.

Le lissage n'est pas applicable dans les cas où les donneurs à l'intérieur d'une case d'ajustement doivent être appariés individuellement aux enregistrements receveurs (c'est-à-dire aux enregistrements auxquels on doit imputer une valeur). Quand les cases ne contiennent pas assez de donneurs, il faut les combiner pour former des cases de taille acceptable. En utilisant un ensemble précis de règles, on peut facilement automatiser un algorithme qui servira à grouper les cases d'ajustement de manière qu'elles aient au moins une taille limite minimum satisfaisante.

Pour imputer des valeurs de changement à l'élément "Autres revenues", on a procédé de la façon suivante pour grouper les cases d'ajustement de manière à obtenir au moins dix donneurs (avec changements différent de zéro) dans chaque case. On a d'abord groupé les cases d'ajustement selon les classes définies par le ratio de sélection, en combinant les cases adjacentes les unes après les autres jusqu'à ce que le nombre de donneurs ait atteint ou dépassé dix. Si cela ne suffisait pas, on les a groupées selon la taille. Si cela ne suffisait pas encore, on a groupé les cases selon la classification des activités économiques, mais seulement à l'intérieur du même groupe d'entreprises.

Pour l'élément "Autres revenus", cette stratégie de groupement a produit 45 cases d'ajustement sur un total possible de 345 cases. Dans la catégorie des entreprises non financières, il a souvent fallu aller jusqu'à la classification des activités économiques pour grouper les cases de manière à atteindre le seuil limite (parce qu'en général il y a moins de 20 donneurs parmi les entreprises non financières de chacun des sous-groupes de la

classification des activités économiques). Par contre, on n'a pas eu à faire beaucoup de groupements dans les entreprises du secteur des banques.

5. APPARIEMENT EN FONCTION DU MONTANT INITIAL

Dans chaque case d'ajustement, comme on l'a dit précédemment, on a choisi comme donneur pour chaque imputation le plus proche voisin de l'enregistrement receveur en fonction du montant initial. Un des dangers toutefois de cette stratégie était qu'un même donneur puisse servir dans un trop grand nombre d'appariements. Pour réduire ce risque, on a prévu qu'au cas où le plus proche donneur aurait déjà été utilisé 20 fois ou plus, on prendrait le plus proche donneur suivant. Toutefois, quand le plus proche voisin suivant était situé à une distance de l'enregistrement receveur supérieure à une certaine distance précisée au préalable, ou avait lui-même déjà servi 20 fois ou plus, on revenait au choix initial -- c'est-à-dire au plus proche voisin.

Malgré la possibilité de choisir un donneur secondaire, le couplage des enregistrements receveurs avec le plus proche voisin a abouti à une utilisation très inégale des donneurs à l'intérieur d'une même case d'ajustement. Cela est attribuable au fait que, dans une même case d'ajustement, la distribution des montants initiaux à l'intérieur du petit nombre de donneurs est différente de la distribution à l'intérieur des enregistrements receveurs. Les donneurs très proches d'autres donneurs ont été utilisés moins souvent que les donneurs plus éloignés de leurs plus proches voisins. Dans certaines cases d'ajustement, la différence était assez importante.

Le tableau 4 présente le logarithme du montant "Autres revenus" et les valeurs obtenues pour la fréquence d'utilisation des donneurs et la proportion d'"Autres revenus" enlevée de chaque donneur dans la première case d'ajustement, qui correspond au premier sous-groupe d'activités économiques (voir appendice). Les donneurs sont énumérés les uns après les autres selon le montant initial de l'élément "Autres revenus". La fréquence d'utilisation des 19 donneurs varie selon un rapport d'un à 20, la moyenne se situant à 8.3 utilisations par donneur environ. On peut voir que la fréquence d'utilisation est inversement proportionnelle à la distance séparant chaque enregistrement donneur des deux enregistrements adjacents, qui figure dans la deuxième colonne. Le premier donneur et le dernier donneur sont également utilisés plus souvent parce qu'ils sont les plus proches voisins de tous les enregistrements receveurs situés aux extrémités de la distribution de l'élément "Autres revenus".

La proportion d'"Autres revenus" enlevée de chaque donneur varie beaucoup autour d'une moyenne de 0.519 (ou de 0.608 si l'on pondère par la fréquence d'utilisation). La proportion d'"Autres revenus" enlevée de chaque donneur varie beaucoup entre donneurs adjacents -- par exemple, seulement 0.116 a été enlevée du donner numéro 18 tandis que 0.954 a été enlevée du dernier donneur. Par conséquent, la variation de la fréquence d'utilisation explique en grande partie la variance due à l'imputation. Dans cette case d'ajustement, le couplage avec le plus proche voisin selon le montant initial d'"Autres revenus" empêche, comme on le voulait, qu'on impute des changements anormalement petits à des donneurs indiquant de petits montants initiaux. Toutefois, on impute aussi en même temps des changements proportionnellement élevés aux donneurs indiquant les plus gros montants initiaux, contrairement à ce que l'on espère.

De toute évidence, cette approche doit être révisée. L'appariement en fonction d'un classement par ordre de grandeur (rang percentile) à l'intérieur d'une case d'ajustement égaliserait l'utilisation des donneurs à l'intérieur des cases, mais une utilisation uniforme n'est pas le résultat le plus favorable à rechercher quand la distribution des montants initiaux parmi les donneurs diffère beaucoup de la distribution parmi les enregistrements receveurs. Un appariement en fonction **d'intervalles** de montants initiaux devrait produire les meilleurs résultats. Cette façon de procéder permet de faire varier le degré

d'utilisation des donneurs en fonction de la forme générale de la courbe de distribution des montants initiaux chez les enregistrements pour lesquels on veut imputer une valeur (enregistrements receveurs), mais ne lie pas assez l'utilisation des donneurs aux fluctuations dues à l'échantillonnage.

Quelle que soit la méthode utilisée pour apparier les enregistrements à l'intérieur d'une case d'ajustement, le fait demeure que l'imputation des valeurs de changement donnerait des résultats bien meilleurs si l'on identifiait de meilleures variables auxiliaires. Comme le démontre le tableau 4, la proportion d'"Autres revenus" enlevée de chaque donneur d'une même case d'ajustement varie beaucoup d'un donneur à l'autre, et les valeurs de changement imputées deviennent d'autant plus sensibles à la distribution des donneurs à l'intérieur de chaque case d'ajustement.

Tableau 4

Donneurs utilisés pour imputer des montants en valeur absolue de changement à l'élément "Autres Revenus" des enregistrements des entreprises du premier sous-groupe d'activités économiques

Numéro du donneur	Logarithme du montant "Autres revenus"	Distance entre les donneurs	Fréquence d'utilisation	Proportion du montant "Autres revenus" enlevée
1	7.89	0.03	19	1.000
2	7.92	1.23	6	0.438
3	9.15	1.16	20	0.035
4	10.31	0.39	20	0.800
5	10.70	0.08	9	0.626
6	10.78	0.19	6	0.230
7	10.97	0.37	16	1.000
8	11.34	0.07	5	0.729
9	11.41	0.17	3	0.122
10	11.58	0.01	3	0.952
11	11.59	0.09	1	0.987
12	11.68	0.12	1	0.264
13	11.80	0.27	7	0.756
14	12.07	0.08	7	0.370
15	12.15	0.09	2	0.006
16	12.24	0.43	6	0.081
17	12.67	0.39	6	0.387
18	13.06	0.12	5	0.116
19	13.18		16	0.954
Moyenne			8.3	0.519

6. CONCLUSION

Dans le contexte budgétaire actuel, il ne fait aucune doute qu'on a besoin de méthodes qui réduisent les coûts de production de bases de données analytiques à partir de dossiers administratifs. Pour diminuer le coût d'une des composantes de son programme de statistiques sur les sociétés, l'IRS a utilisé une méthode d'échantillonnage à deux degrés et imputé des valeurs aux enregistrements non contrôlés. Les microdonnées produites suivant cette technique peuvent être totalisées de nombreuses de façons pour des sous-populations très désagrégées. Le calendrier de production et la nécessité de faire plusieurs imputations par enregistrement militent en faveur d'une méthode d'imputation fondée sur la substitution de valeurs de donneurs contrôlés à des enregistrements non contrôlés. Toutefois, si l'on procède à un sous-échantillonnage en vue de réduire les coûts de contrôle, il faut se rappeler qu'on peut aboutir à un nombre de donneurs représentant une fraction seulement de la taille totale de l'échantillon. Même avec des bases de données de presque 100,000 enregistrements, il est possible qu'il n'y ait qu'un petit nombre de donneurs potentiels par rapport au nombre qu'il faudrait pour utiliser le mieux possible les méthodes traditionnelles d'imputation par substitution.

Le problème du manque de donneurs a été étudié par l'IRS dans le cadre de la révision des procédures d'imputation devant servir en 1985 à la production des statistiques du revenu des sociétés. Un élément clé était la distinction de deux opérations, à savoir s'il faut imputer un changement différent de zéro ou un changement nul et, deuxièmement, l'imputation d'une valeur conditionnelle du changement. La première opération pouvait incorporer des éléments d'imputation obtenus à partir de modèles, tout en continuant de se prêter à l'automatisation. Les résultats de l'imputation de valeurs conditionnelles confirment la difficulté de faire des imputations quand les donneurs sont rares. Il faut poursuivre les travaux dans ce domaine et il ne fait aucun doute qu'on réussirait à améliorer les résultats si l'on pouvait identifier de meilleures variables auxiliaires aux valeurs conditionnelles des changements à apporter aux enregistrements contrôlés.

REMERCIEMENTS

Cette recherche a été effectuée par la société Mathematica Policy Research, Inc. (MPR) par contrat avec la Statistics of Income Division du U.S. Internal Revenue Service. Je tiens à remercier tout le personnel de la SOI Division pour son appui et en particulier Fritz Scheuren et Susan Hinkins pour m'avoir permis d'utiliser leurs travaux précédents et m'avoir fait de précieux commentaires en cours de route. Je voudrais aussi remercier David Edson et Cavan Capps de MPR pour avoir écrit le logiciel d'imputation et produit les tableaux présentés dans le document ainsi que Donald Rubin et Roderick Little pour leur contribution au développement des méthodes d'imputation de 1985.

BIBLIOGRAPHIE

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

Czajka, J.L. (1986). "Imputation of Selected Items in corporate tax data: improving upon the earlier hot deck", *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Czajka, J.L. (1987). "Predicting edit outcomes: the strategic use of imputation in estimating corporate income statistics", *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Hinkins, S. (1983). "Matrix sampling and the related imputation of corporate income tax returns", *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Hinkins, S. (1984). "Matrix sampling and the effects of using hot deck imputation", *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Hinkins, S., et Scheuren, F. (1986). "L'imputation par la méthode "hot deck" appliquée à un plan d'échantillonnage à deux degrés", *Techniques d'enquête*, volume 12, décembre 1986, 189-204.

Appendice
Classification des activités économiques

Commerce de gros

1. Produits alimentaires et produits connexes; véhicules à moteur et automobiles; ameublement et meubles de maison; sports et loisirs; médicaments et produits pharmaceutiques divers; vêtements, tissu à la verge et articles divers; boissons alcooliques.
2. Machines, matériel et fournitures.
3. Bois et matériaux de construction; métaux et minéraux; appareils électriques; quincaillerie, plomberie, matériel de chauffage et fournitures; autres biens durables; papier et produits du papier; produits agricoles bruts; produits chimiques, produits du pétrole et produits connexes; biens non durables divers.

Commerce de détail

4. Matériaux de construction, quincailleries et articles de jardinage; magasins de marchandises diverses et magasins d'alimentation.
5. Détaillants en véhicules automobiles et stations-service.
6. Magasins de vêtements et d'accessoires vestimentaires; magasins de meubles et d'articles d'ameublement.
7. Restaurants et débits de boissons.
8. Magasins de vente au détail divers; magasins de commerce de gros et de commerce de détail non classés.

Industries manufacturières

9. Produits alimentaires et produits de même nature; fabricants de produits du tabac; produits manufacturés divers; entreprises non classées.
10. Produits textiles; vêtements et autres produits textiles; produits en bois.
11. Meubles et articles d'ameublement; industrie du papier et activités annexes; première transformation des métaux; fabrication de machines, sauf électriques; fabrication de véhicules automobiles et de matériel de transport; instruments et produits apparentés.
12. Produits métalliques fabriqués; matériel électrique et matériel électronique.

Services

13. Hôtels et autres installations d'hébergement; services personnels; ateliers de réparation d'automobiles et services de réparation divers; divertissements et loisirs.
14. Services fournis aux entreprises; services médicaux; bureaux d'architectes et bureaux d'ingénieurs; bureaux de comptables, de vérificateurs et services de tenue de livres; services divers.
15. Publicité; services juridiques, éducatifs et sociaux; associations.

Autres industries non financières

16. Agriculture, exploitation forestière et pêche; mines; entrepreneurs généraux en bâtiment, constructeurs spéculateurs et entrepreneurs en travaux publics.
17. Entrepreneurs spécialisés.
18. Transports et services publics.

Banques, crédit et finances

19. Banques
20. Organismes de crédit; courtiers en valeurs mobilières et courtiers en marchandises.

Autres industries financières

21. Assurances; agents d'assurances et courtiers d'assurances.
22. Exploitants immobiliers et agents de location d'immeubles.
23. Autres affaires immobilières; sociétés de portefeuille et sociétés d'investissement.

SESSION VII: COMMUNICATIONS OFFERTES

Président: Daniel Kazprzyk, U.S. Bureau of the Census

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

RAPPORT ENTRE LES CARACTÉRISTIQUES DE MEURTRES, L'ISSUE DES PROCÈS ET LA PEINE CAPITALE: Canada, 1961-1983

JANE F. GENTLEMAN et PAUL B. REED[1]

RÉSUMÉ

Statistique Canada recueille des données sur tous les homicides (meurtres, homicides involontaires et infanticides) commis au Canada et portés à l'attention des services policiers, y compris de l'information sur les circonstances de l'homicide, les caractéristiques des personnes impliquées et les activités du système judiciaire déclenchées par cet homicide. Cet ensemble détaillé de données est formé principalement à l'aide de dossiers administratifs. Dans cet article, nous exposons quelques résultats d'une étude portant sur les meurtres qui ont été commis entre 1961 et 1983. Nous ajustons des modèles de régression logistique afin d'analyser 1) l'effet des caractéristiques de meurtres et de l'abolition de la peine de mort au Canada en 1976 sur l'issue des procès et 2) les caractéristiques de meurtres qui ont un rapport avec l'abolition de la peine de mort.

1. INTRODUCTION

La série de données sur laquelle repose cette étude renferme des renseignements très complets sur tous les homicides (meurtres, homicides involontaires et infanticides) qui ont été commis au Canada et rapportés à la police de 1961 à 1983; cette série de données renferme notamment de l'information sur la nature et les circonstances de l'homicide, les caractéristiques des personnes impliquées et les liens qui existent entre ces personnes de même que la nature des accusations portées et l'issue du procès. Ce fichier statistique détaillé a été produit essentiellement à partir de dossiers administratifs. Les données ont été recueillies par Statistique Canada dans le cadre du Programme des homicides du Centre canadien de la statistique juridique (CCSJ). Ce programme sert à mesurer un phénomène dont les manifestations **ne peuvent être constatées que** par des dossiers administratifs. Les homicides constituent une population d'événements relativement restreinte et supposent un comportement tellement répréhensible que les autorités publiques n'hésitent pas à consacrer beaucoup de ressources à leur élucidation. La conséquence de ces deux caractéristiques - la faible population d'événements et l'importance exceptionnelle qui leur est accordée - est que le Programme des homicides de Statistique Canada permet de dénombrer les homicides connus avec un très faible taux d'erreur de couverture. Cela démontre la capacité des données administratives de produire des séries de micro-données longitudinales se rapportant à des unités géographiques et temporelles extrêmement précises.

[1] Jane F. Gentleman et Paul B. Reed, Direction des études analytiques, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6

Les données sont organisées par "affaire", c'est-à-dire selon chaque événement où un homicide est commis, peu importe le nombre de suspects ou de victimes. Elles décrivent toutes les victimes et tous les suspects pour chaque affaire. Le fichier de données qui a été utilisé pour la présente étude renferme en tout 9 642 affaires d'homicide, y compris des données pour 10 627 victimes et 9 954 suspects.

À chaque mois, les services de police de tout le pays font parvenir au CCSJ quelque 1 500 copies d'un rapport intitulé "Formule C - Statistique de la criminalité" (un spécimen de ce rapport est reproduit dans Statistique Canada, 1986a). Cette formule (ou la bande magnétique correspondante) renferme des données agrégées sur des affaires telles que les homicides, les infractions d'ordre sexuel, les voies de fait, les vols qualifiés, la prostitution, le trafic de drogue, etc. Pour chaque affaire d'homicide rapportée sur la Formule C, le service de police concerné fournit au Programme des homicides du CCSJ un "Rapport sur les homicides" qui contient des renseignements détaillés sur l'affaire. Un spécimen de ce rapport est reproduit dans Statistique Canada, 1986b. Les codeurs de Statistique Canada utilisent ces données et d'autres sources d'information, comme les coupures de journaux et les appels téléphoniques faits à des services de police, pour remplir un "rapport sur la victime et l'infraction" pour chaque victime ainsi qu'un "rapport sur le suspect" pour chaque suspect. On reproduit ensuite les données de ces rapports sur une bande magnétique qui contient des données chronologiques sur les homicides et qui a servi, de fait, à la réalisation de la présente étude.

Statistique Canada envisage actuellement de modifier le Rapport sur les homicides et quelques-unes des définitions de codes de données (voir Nabata, 1986, et Gentleman et Dixon, 1985, 1986 et 1987). Il est nécessaire de renouveler périodiquement le système de collecte des données pour diverses raisons. Il y a bien sûr l'évolution de la nature et du degré de diffusion des données de même que l'évolution des besoins et des méthodes d'analyse. Lorsqu'il recueille des données sur les homicides, le personnel du CCSJ doit concilier des exigences parfois contradictoires, c'est-à-dire limiter le coût de production des données, trouver un juste équilibre entre le fardeau de réponse des autorités policières et les besoins des utilisateurs de données, conserver les définitions de variables tout en renouvelant les concepts et en résolvant les difficultés, et garantir la protection des renseignements personnels tout en fournissant des séries de données détaillées au public. Lorsqu'on analyse de près les grandes séries de données, on s'aperçoit qu'elles présentent toutes certaines faiblesses. Grâce à une analyse "interne" des données de Statistique Canada, les auteurs ont pu proposer certains moyens pour résoudre ces faiblesses, leur analyse étant d'autant plus valable qu'ils avaient accès aux fichiers de micro-données et pouvaient à loisir consulter les producteurs de données, qui étaient tout près.

Toutes les affaires d'homicide sont suivies (au sens statistique) par le CCSJ. Les fichiers sont mis à jour selon les besoins en y incorporant les données de la cour sur l'instance des affaires de même que des données fournies par le Service correctionnel du Canada. Le fichier ayant servi à la présente étude a été mis à jour en 1986. Les homicides survenus après 1983 n'ont pas été étudiés car on a jugé que dans un bon nombre de cas, les dossiers judiciaires avaient peu de chances d'être complets. Les suspects qui n'avaient pas encore subi au moins un procès ont été soustraits à l'analyse. Cela comprend, entre autres, les personnes qui attendent de subir leur procès ou celles qui sont décédées avant leur condamnation ou leur acquittement. Dans le cas d'une personne qui avait déjà subi plusieurs procès, c'est le dernier procès qui a servi de source de renseignements pour la déclaration de culpabilité et le prononcé de la sentence.

La base de données sur les homicides ne contient pas de micro-données concernant les homicides involontaires et les infanticides qui se sont produits avant 1974 car à l'origine, la série chronologique ne portait que sur les meurtres (l'homicide involontaire et l'infanticide n'étant pas définis comme des meurtres). En conséquence, pour ne pas grossir artificiellement le taux d'homicide pour la période 1974-1983, nous avons exclu de notre étude toutes les personnes qui ont été reconnues coupables d'homicide involontaire ou

d'infanticide. De même, nous n'avons pas tenu compte des personnes mineures qui ont été jugées au tribunal de la jeunesse. (En revanche, nous avons tenu compte de tous les suspects - adolescents ou adultes - qui ont subi un procès pour meurtre devant les tribunaux ordinaires.)

Le tableau 1 énumère les variables qui ont servi à cette étude. Parmi celles-ci, notons deux variables chronologiques : la date de règlement (c'est-à-dire la date où le suspect a été acquitté ou condamné) et la date de l'infraction. Deux variables indicatrices ont été définies pour l'existence de la peine capitale, l'une d'elles indiquant si la peine capitale était en vigueur à la date de règlement et l'autre indiquant si elle était en vigueur à la date de l'infraction. Deux variables fractionnaires ont été utilisées : le taux mensuel moyen de meurtres par rapport à l'année précédente à la date de règlement et le taux mensuel moyen de meurtres par rapport à l'année précédente à la date de l'infraction.

Pour chaque variable discrète - c'est-à-dire pour toutes les variables sauf les variables chronologiques et les variables fractionnaires - nous avons défini des variables auxiliaires distinctes correspondant à chaque résultat, celles-ci devant être utilisées comme variables indépendantes dans les régressions logistiques (un résultat - la "catégorie de référence" - étant omis pour chaque variable afin d'éviter que la matrice descriptive ne soit singulière).

Tableau 1

Variables décrivant une affaire d'homicide

VARIABLES CONCERNANT LA VICTIME

- Sexe
- Groupe d'âge
- État matrimonial
- Origine raciale

VARIABLES CONCERNANT LE SUSPECT

- Sexe
- Groupe d'âge
- État matrimonial
- Origine raciale
- Niveau d'instruction
- Condamné ou acquitté
- Sévérité de la sentence
- Date de règlement

AUTRES VARIABLES

- Lien du suspect avec la victime
- Méthode du crime
- Mobile apparent
- Autres circonstances
- Lieu de crime
- Région géographique
- Date du crime
- Existence de la peine capitale
- Taux de meurtres par rapport à l'année précédente

Selon le modèle de régression logistique ordinaire, $P = Pr(Y=1)$ est prédit comme $1/[1+\exp(-\hat{\alpha}-X\hat{\beta})]$ où Y est la variable dépendante (qui prend la valeur 0 ou 1), X est un vecteur ligne formé de valeurs de variables indépendantes, $\hat{\alpha}$ est l'ordonnée à l'origine ajustée et $\hat{\beta}$ est le vecteur colonne formé des coefficients de régression de X ajustés. En termes mathématiques, cela revient à supposer que le logarithme naturel de la probabilité relative que Y soit égale à 1 est une fonction linéaire des variables indépendantes : $\ln[P/(1-P)] = \alpha + X\beta$. Dans cet article, les résultats de la régression sont exprimés en termes du rapport de probabilités $\exp(\hat{\beta}_j)$ où $\hat{\beta}_j$ est le coefficient ajusté d'une variable auxiliaire. Le rapport de probabilités (RP) est le facteur par lequel on multiplie le rapport estimé $\hat{P}/(1-\hat{P})$ lorsque la valeur de la variable auxiliaire passe de 0 à 1.

Le programme d'ordinateur qui a servi à l'exécution des régressions logistiques de cette étude est décrit dans SAS (1983). Après avoir ajusté un modèle intégral, nous avons appliqué une méthode efficiente d'élimination à rebours "rapide" afin de choisir des variables pour un modèle final "réduit"; l'algorithme utilisé était inspiré de Lawlen et Singhal (1978). Nous avons supprimé une à une les variables, sauf l' ordonnée à l'origine, jusqu'à ce que le modèle ne contienne plus que des variables avec un seuil de signification inférieur à 0.05. (Les variables qui avaient été extraites du modèle pouvaient par la suite y être remises si, une fois dans le modèle, elles avaient un seuil de signification inférieur à 0.05.) Nous avons vérifié l'ajustement total d'une régression logistique au moyen de l'approximation chi carré habituelle du rapport des vraisemblances et de la cote. La fraction des paires concordantes (FPC) a été calculée pour chaque régression comme un autre indice de l'efficacité prédictive globale du modèle ajusté. La FPC calculée par le programme décrit dans SAS (1983) est définie comme la proportion des paires d'estimations de la variable dépendante qui sont ordonnées correctement, c'est-à-dire des paires qui sont disposées dans le même ordre que les observations correspondantes.

Cet article présente les résultats de deux genres de régression logistique. Dans la section 2, nous décrivons une régression logistique où la variable dépendante indique si la personne est condamnée ou acquittée et où les variables indépendantes sont tirées de toutes les autres variables énumérées au tableau 1 sauf "sévérité de la sentence". Dans la section 3, il s'agit d'une régression logistique où la variable dépendante indique si la peine capitale est en vigueur ou non et où les variables indépendantes sont tirées de toutes les autres variables énumérées au tableau 1 sauf la date de l'infraction et le taux de meurtres par rapport à l'année précédente. Ces résultats seront exposés en détail un peu plus loin.

2. RÉGRESSION PERMETTANT D'ESTIMER LA PROBABILITÉ DE CONDAMNATION

Pour déterminer quels genres de meurtres ont été influencés par les variations de la probabilité de condamnation et pour vérifier si l'existence de la peine capitale (EPC), les variables chronologiques et les variables fractionnaires ont eu une incidence quelconque, nous avons effectué des régressions logistiques où la variable dépendante indique si la personne est condamnée ou acquittée. Nous avons aussi utilisé pour ces régressions une variable dépendante polychotomique, soit la "sévérité de la sentence". Afin de tenir compte des décisions des tribunaux, nous avons défini les variables chronologiques de ces régressions (EPC, date et taux) en fonction de la date de règlement. Une régression dont la variable dépendante indique s'il y a eu condamnation ou acquittement peut être considérée comme un moyen d'estimer la probabilité qu'un suspect soit déclaré coupable par un jury. La régression à variable dépendante polychotomique permet de prévoir la sévérité de la sentence (étant donné une condamnation), laquelle est choisie par le juge parmi une série de sentences étroitement liées au genre de condamnation prononcée par le jury. Le juge prend aussi en considération les rapports présentenciels de même que les recommandations du procureur, du jury et de la défense. Par conséquent, les résultats de

la régression polychotomique reflètent l'attitude du juge conditionnée par ces autres facteurs.

Les conséquences de l'attitude du juge ou du jury peuvent être confondues avec celles des méthodes de mise en accusation pratiquées par les services policiers. Par exemple, si pour une raison ou pour une autre les autorités policières ont déposé un nombre disproportionné d'accusations de meurtre contre des hommes innocents, on serait porté à dire des jurys qui ont acquitté ces personnes qu'ils ont acquitté un nombre disproportionné d'hommes. De la même façon, on risque de confondre le comportement des juges avec celui des suspects. Par exemple, si les hommes tendent à commettre des homicides plus punissables que ceux commis par les femmes, un juge impartial aura tendance à imposer une sentence plus sévère aux hommes.

Le tableau 2 donne les résultats d'une régression où la variable dépendante indique si la personne est condamnée ou acquittée. On y trouve également les rapports de probabilités pour certaines variables indépendantes utilisées dans un modèle de régression intégral. Les deux variables de sexe sont très significatives (P=.00 à deux chiffres significatifs). Pour les suspects de sexe masculin (par opposition aux suspects de sexe féminin), la probabilité relative de condamnation est multipliée par un facteur de plus de 1.8, tandis que pour les suspects qui ont été accusés d'avoir tué une personne de sexe masculin (plutôt qu'une personne de sexe féminin), la probabilité relative de condamnation est réduite par un facteur de .65. Les quatre combinaisons formées par le "sexe du suspect" et le "sexe de la victime" ont été classées par ordre décroissant de probabilité de condamnation : 1) homme accusé d'avoir tué une femme (probabilité de condamnation la plus forte); 2) homme accusé d'avoir tué un autre homme; 3) femme accusée d'avoir tué une autre femme; 4) femme accusée d'avoir tué un homme (probabilité de condamnation la plus faible). Un autre modèle de régression qui mettait en corrélation le "sexe du suspect" et le "sexe de la victime" a produit les mêmes résultats.

Tableau 2

Résultats partiels de la régression logistique
Variable dépendante = condamné ou acquitté
Nombre d'observations (nombre de suspects) = 6350

VARIABLE INDÉPENDANTE	RAPPORT DE PROBABILITÉS	VALEUR-P
Sexe du suspect = femme	1.00	---
Sexe du suspect = homme	1.88	.00
Sexe de la victime = femme	1.00	---
Sexe de la victime = homme	.65	.00
Origine raciale du suspect = Caucasienne	1.00	---
Orignie raciale du suspect = Canadienne	1.34	.03
Origine raciale de la victime = Caucasienne	1.00	---
Origine raciale de la victime = Canadienne	.78	.07
Lien du suspect avec la victime		
Victime = conjoint du suspect	1.00	---
Victime = parent du suspect	.39	.00
Victime = enfant du suspect	1.50	.06
Existence de la peine capitale (EPC)		
Peine capitale non en vigueur	1.00	---
Peine capitale en vigueur	.71	.00

La probabilité de condamnation était beaucoup plus élevée pour les suspects d'origine canadienne (RP=1.34, P=.03) que pour les suspects d'origine caucasienne.

Les personnes accusées d'avoir tué des Canadiens d'origine avaient un peu moins de chances d'être condamnées (RP=.78) que les personnes accusées d'avoir tué des Caucasiens. Bien que cette variable indépendante ait eu une valeur P élevée (.07) dans le modèle de régression intégral, elle a été retenue dans le modèle réduit, où elle avait un rapport de probabilités de .76 et une valeur P de .04.

Le groupe de référence pour la variable "lien du suspect avec la victime" est formé des personnes accusées d'avoir tué leur conjoint. Les personnes accusées d'avoir tué leur parent avaient le moins de chances d'être condamnées (RP=.39, P=.00) tandis que les personnes accusées d'avoir tué leur enfant avaient le plus de chances d'être condamnées (RP=1.50). Malgré une valeur P élevée (.06), cette dernière variable a été retenue dans le modèle réduit, où elle avait un rapport de probabilités de 1.66 et une valeur P de .00.

Dans l'estimation de la probabilité relative de condamnation, l'existence de la peine capitale (EPC) a été la plus importante des trois variables chronologiques (EPC, date et taux). Elle était aussi la seule variable significative (RP=.71, P=.00); la date et le taux avaient un niveau de signification élevé dans le modèle intégral (.28 et .47 respectivement) mais ne faisaient pas partie du modèle réduit. Ces résultats montrent que, peu importe qu'il y ait une correction pour la date et le taux, la probabilité de condamnation était beaucoup moins élevée lorsque la peine capitale était en vigueur. On pourrait voir dans ces résultats une hésitation de la part des jurys à prononcer un verdict de culpabilité lorsque la peine capitale était du moins une possibilité théorique.

3. RÉGRESSION PERMETTANT DE PRÉVOIR L'EXISTENCE DE LA PEINE CAPITALE

De nombreuses études ont été faites au cours des trente dernières années pour tenter d'établir un lien entre l'existence de la peine capitale et le taux d'homicide. Les conclusions tirées de cette volumineuse recherche sont souvent divergentes; on n'a pu démontrer de façon fiable et convaincante que la peine capitale avait un effet dissuasif. Néanmoins, l'efficacité de la peine capitale sur ce plan fait encore l'objet de nombreuses discussions sur la place publique. La présente étude vise à analyser objectivement les données canadiennes sur l'homicide afin de déterminer, s'il en est, les caractéristiques de meurtres qui sont subordonnées à l'existence ou à l'inexistence de la peine capitale. À cette fin, nous avons effectué des régressions logistiques où nous nous sommes servis de la variable dichotomique "existence de la peine capitale" comme variable dépendante. Pour tenir compte du comportement du suspect, nous avons défini l'existence de la peine capitale" en fonction de la date de l'infraction : si l'infraction a eu lieu avant juillet 1976, EPC est égale à 1; autrement, EPC est égale à 0. Ces régressions permettent donc d'établir un rapport entre les caractéristiques et le comportement du suspect d'une part et la date du meurtre d'autre part.

Nous nous devons ici de faire une mise en garde importante. L'abolition de la peine de mort en juillet 1976 est survenue à l'issue d'une période durant laquelle on s'était livré à de multiples révisions de la définition et de la classification des meurtres et des peines qui s'y rattachent (voir la perspective historique à l'appendice V de Statistique Canada, 1986b), et au cours de laquelle avaient été décrétés des moratoires **de droit** et **de fait** à la suite de la dernière exécution en 1962; l'abolition de la peine de mort est également survenue à un moment où le droit pénal canadien subissait des modifications profondes, dont certaines, comme la législation touchant le contrôle des armes à feu et adoptée en janvier 1978, devaient sûrement avoir des répercussions sur le taux de meurtres et le système judiciaire. Ainsi, dire que juillet 1976 marque brusquement la fin de l'existence de la peine capitale équivaut à simplifier la question et risque de camoufler les effets de

divers facteurs. En outre, le fait de devoir analyser les données d'une expérience incontrôlée nous réduit à faire des associations entre variables plutôt qu'une analyse des causes et effets.

Le tableau 3 donne certains résultats d'un modèle de régression intégral où l'existence de la peine capitale" sert de variable dépendante. Les meurtres commis avec un couteau (RP = .59, P = .00) sont liés beaucoup moins étroitement à l'EPC que ne le sont les meurtres commis avec une arme à feu, lesquels constituent la catégorie de référence.

La variable "autres circonstances" est tirée d'une section descriptive du Rapport sur les homicides. À partir des renseignements fournis par les services policiers, les codeurs de Statistique Canada enregistrent l'un ou l'autre des éléments suivants : boissons alcooliques, stupéfiants, milieu ou aucune circonstance particulière. Comme ces éléments ne s'excluent pas mutuellement et que l'interprétation de cette variable soulève d'autres problèmes, les résultats pertinents devraient être considérés avec prudence. Selon le tableau 3, les meurtres liés à la consommation d'alcool avaient une relation beaucoup plus étroite avec l'EPC que les meurtres commis dans d'autres circonstances (RP variant de .47 à .64 et toutes les valeurs P égales à .00).

Le sexe du suspect ou le sexe de la victime ne modifiaient pas réellement la probabilité estimée que le meurtre soit commis pendant que la peine capitale était en vigueur.

La variable "mobile apparent" est une présomption établie par les enquêteurs qui indiquent pourquoi, selon eux, le suspect aurait commis l'homicide. Par rapport à la catégorie de référence (colère, haine, dispute ou querelle), les autres catégories du "mobile apparent" avaient toutes des rapports de probabilités plus élevés et pour la plupart significatifs. Les homicides commis vraisemblablement en état de légitime défense étaient ceux qui présentaient le plus fort rapport de probabilités (RP = 3.52, P = .00).

Le tableau 3 peut être analysé dans l'optique de ceux qui prétendent que la peine capitale empêche les meurtres et qu'elle décourage plus les meurtres prémédités que les meurtres sans préméditation. Ce n'est pas chose facile de déterminer si un meurtre a été prémédité et dans quelle mesure il l'a été mais quelques-unes des catégories de meurtre énumérées au tableau 3 peuvent probablement être considérées comme relativement préméditées ou relativement involontaires. Selon la théorie de la dissuasion, un meurtre sans préméditation devrait avoir un rapport de probabilités plus élevé que le meurtre avec préméditation car le meurtre qui a été prévenu serait lié moins étroitement à l'existence de la peine capitale. On peut prétendre que le tableau 3 confirme cette théorie en ce qui a trait aux meurtres liés à la consommation d'alcool (par opposition à ceux qui ne le sont pas) et aux meurtres dont le mobile apparent est la légitime défense (par opposition à ceux qui, par exemple, ont été commis par vengance), si bien sûr on considère les meurtres liés à la consommation d'alcool et les meurtres commis en état de légitime défense comme relativement involontaires. On ne peut dire la même chose des meurtres qui ont été commis à l'aide d'un couteau (par opposition à ceux qui ont été commis à l'aide d'une arme à feu), si l'on considère que les premiers sont "moins prémédités" que les seconds. Toutefois, Scarff (1983) a laissé à entendre que le couteau se serait substitué dans une certaine mesure à l'arme à feu par suite de l'adoption de lois sur le contrôle des armes à feu en 1978.

Tableau 3

Résultats partiels de la régression logistique
Variable dépendante = existence de la peine capitale
Nombre d'observations (nombre de suspect) = 6369

VARIABLE INDÉPENDANTE	RAPPORT DE PROBABILITÉS	VALEUR-P
Méthode du crime = arme à feu	1.00	---
Méthode du crime = couteau	.59	.00
Autres circonstances =		
Boissons alcooliques	1.00	---
Circonstances autres que boissons alcooliques	.47-.64	.00
Sexe du suspet = femme	1.00	---
Sexe du suspect = homme	NS	NS
Sexe de la victime = femme	1.00	---
Sexe de la victime = homme	NS	NS
Mobile apparent =		
Colère, haine, dispute, querelle	1.00	---
Attentat à la pudeur, vol, vol qualifié, etc.	NS	NS
Vengance	1.37	.00
Jalousie	2.20	.00
Inconnu	2.20	.00
Aliénation mentale	2.23	.00
Autre	2.57	.00
Légitime défense	3.52	.00

NS = Rapport de probabilités ne diffère pas sensiblement de 1.00 au niveau .05.

BIBLIOGRAPHIE

Gentleman, Jane F., et Dixon, Daniel P. (1985). Assessment of the Analytic Usefulness of Statistics Canada Homicide Data. Rapport interne de Statistique Canada. Division des études sociales et économiques.

Gentleman, Jane F., et Dixon, Daniel P. (1986). Comments on Revised Coding Instructions for Statistics Canada Homicide Data. Rapport interne de Statistique Canada. Division des études sociales et économiques.

Gentleman, Jane F., et Dixon, Daniel P. (1987). Comments on Proposed Changes to the Homicide Return and to the Homicide Data Coding System (1987). Rapport interne de Statistique Canada. Division des études sociales et économiques.

Lawless, J.F., et Singhal, K. (1978). Efficient Screening of Nonnormal Regression Models. *Biometrics*, 34, 318-327.

Nabata, Tony (1986). The Homicide Development Project - The Homicide Return. Rapport interne de Statistique Canada. Centre canadien de la statistique juridique.

SAS Institute Inc. (1983). SUGI Supplemental Library User's Guide, édition de 1983. (Description de la méthode LOGIST par Frank E. Harrell Jr.), SAS Institute Inc., Caroline du nord, 181-202.

Scarff, Elisabeth (1983). *L'évaluation des mesures législatives canadiennes relatives au contrôle des armes à feu : rapport final.* Solliciteur général Canada.

Statistique Canada (1986a). *Statistique de la criminalité du Canada*, 1985. Numéro 85-205 au répertoire, annuel. Centre canadien de la statistique juridique.

Statistique Canada (1986b). *L'homicide au Canada, 1984, Perspective statistique.* Numéro 85-209 au répertoire, annuel (intitulé *Statistique de l'homicide* pour les données de 1981 et des années précédentes). Centre canadien de la statistique juridique.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

UTILISATIONS DES FICHIERS ADMINISTRATIFS AU CANADA POUR L'ÉTABLISSEMENT D'ESTIMATIONS DE LA POPULATION ET DES COMPOSANTES DE L'ACCROISSEMENT DÉMOGRAPHIQUE

RAVI B.P. VERMA ET RONALD RABY[1]

RÉSUMÉ

Les auteurs établissent dans quelle mesure l'utilisation du fichier des allocations familiales et du fichier d'impôt de Revenu Canada en vue de la production d'estimations des émigrants du Canada et des migrants interprovinciaux se révèle satisfaisante. Ils comparent l'application de ces fichiers pour fins d'estimation de la population totale du Canada, des provinces et des territoires aux chiffres du recensement de 1986. On démontre que ces deux fichiers administratifs offrent des séries de données cohérentes et raisonnablement précises sur l'émigration et la migration interprovinciale pour la période de 1981 à 1986. Enfin, il semble que l'estimation des émigrants produite à partir du fichier des allocations familiales serait plus précise si l'on utilisait, non plus le fichier d'impôt de Revenu Canada mais plutôt celui de l'immigration, dans le calcul du rapport des taux d'émigration des adultes à ceux des enfants.

1. INTRODUCTION

Les estimations de la population pour le Canada, les provinces et les territoires, les divisions de recensement et les régions métropolitaines de recensement sont basées sur les effectifs du dernier recensement ainsi que sur les données provenant de plusieurs fichiers administratifs: les fichiers d'impôt de Revenu Canada et des allocations familiales dans le cas de la migration, les registres de la statistique de l'état civil dans le cas des naissances et des décès et le registre des visas d'immigrant et les fiches relatives au droit d'établissement en ce qui concerne l'immigration. Les avantages et les désavantages que comporte l'utilisation de ces fichiers administratifs en lieu et place des données du recensement pour estimer la population et la migration, ont fait l'objet d'autres ouvrages (Statistique Canada, no 91-528F au catalogue, 1987; Verma et Parent, 1985; et Norris, Britton et Verma, 1982). Dans le présent document, la précision avec laquelle les sources de données sur les allocations familiales et celles de Revenu Canada permettent d'établir des estimations démographiques pour les provinces et les territoires sera évaluée au moyen de comparaisons avec les données du recensement de 1986. La performance de ces fichiers administratifs en 1986 sera de plus mise en lumière dans une comparaison rétrospective basée sur les recensements de 1971, 1976 et 1981.

[1] Ravi B.P. Verma et Ronald Raby, Division de la démographie, Statistique Canada, 4-A Édifice Jean Talon, Ottawa, Ontario K1A 0T6.

2. SOURCES DE DONNÉES ET MÉTHODES D'ESTIMATION

Les fichiers des allocations familiales et de l'impôt ne couvrant pas l'ensemble des populations-cibles, les données qu'ils fournissent devront être ajustées de façon à estimer de façon adéquate i) la migration interne; ii) l'émigration et iii) la population totale.

i) Migration interne

Pour produire les estimations annuelles et trimestrielles de la migration interprovinciale, on utilise les deux fichiers administratifs précités. Les estimations provisoires sont établies au moyen du fichier des allocations familiales tandis que les estimations définitives le sont à l'aide du fichier d'impôt de Revenu Canada.

Estimations provisoires

Les bénéficiaires d'allocations familiales doivent aviser le ministère de la Santé et du Bien-être social de tout changement d'adresse. Ces changements, enregistrés sur le fichier pertinent et relevés tous les mois, permettent d'étudier les mouvements selon la province d'origine et de destination et la taille de la famille (c'est-à-dire le nombre d'enfants par famille recevant l'allocation). La couverture de la population au moyen des allocations familiales est comparable à celle du recensement (Statistique Canada, no 91-528F au catalogue, 1987:46). On estime le nombre d'adultes (personnes âgées de 18 ans et plus) migrants en pondérant les données sur la migration des enfants (personnes âgées de 0 à 17 ans inclusivement) fournies par le fichier d'allocations familiales par les rapports des taux de migration des adultes à ceux des enfants ($f_{j,k}$), obtenus du fichier d'impôt de Revenu Canada le plus récent, antérieur habituellement d'un an ou deux à la période de référence de l'estimation. Par suite, une simple addition des adultes et des enfants migrants génère l'effectif total des migrants interprovinciaux. Les calculs se font comme suit:

$$\hat{M}_{(j,k),18+} = \frac{M_{(j,k),0-17}}{P_{j,0-17}} \cdot f_{(j,k)} \cdot P_{j,18+} \qquad (1)$$

$$f_{(j,k)} = \frac{M'_{(j,k),\ 18+}}{\hat{P}_{j,18+}} \div \frac{M'_{(j,k),\ 0-17}}{\hat{P}_{j,0-17}} \qquad (2)$$

$$\hat{M}_{(j,k),0+} = \hat{M}_{(j,k),18+} + M_{(j,k),0-17} \qquad (3)$$

où:

$\hat{M}_{(j,k),0+}$ = nombre total estimé de personnes quittant la province j à destination de la province k

$\hat{M}_{(j,k),18+}$ = nombre estimé d'adultes (personnes âgées de 18 ans et plus) quittant la province j à destination de la province k

$M'_{(j,k),18+}$ = nombre d'adultes quittant la province j à destination de la province k selon le fichier d'impôt de Revenu Canada

$M'_{(j,k),0-17}$ = nombre d'enfants (personnes âgés de 0 à 17 ans) quitant la province j à destination de la province k selon le fichier d'impôt de Revenu Canada

$M_{(j,k),0-17}$ = nombre d'enfants quittant la province j à destination de la province k selon le fichier des allocations familiales

$P_{j,18+}$ = nombre estimé d'adultes dans la province j, calculé de façon résiduelle à partir de l'estimation de la population totale (Division de la démographie) et du nombre d'enfants selon le fichier des allocations familiales

$P_{j,0-17}$ = nombre total d'enfants dans la province j, pour lesquels une allocation familiale est versée

$f_{(j,k)}$ = facteur d'estimation des adultes quittant la province j à destination de la province k, selon les estimations des données sur la migration établies à partir du fichier d'impôt de Revenu Canada

$\hat{P}_{j,18+}$ = nombre d'adultes dans la province j; estimations de la Division de la démographie

$\hat{P}_{j,0-17}$ = nombre d'enfants dans la province j; estimations de la Division de la démographie

Estimations définitives

Les estimations définitives de la migration interprovinciale sont produites différemment, et se basent sur le seul fichier d'impôt de Revenu Canada. Toutes les personnes qui gagnent un revenu annuel supérieur à une somme minimale établie doivent remplir une déclaration de revenu aux fins d'impôt avant la fin du mois d'avril de chaque année. On peut isoler les contribuables migrants en comparant pour chaque déclarant l'adresse du domicile consignée dans les déclarations de deux années consécutives. Le nombre et l'âge des personnes à charge sont déduits du montant de l'exemption personnelle totale du contribuable. Les différents effectifs ainsi obtenus sont ensuite ajustés afin de prendre en compte les migrants qui ne remplissent pas de déclaration de revenu aux fins d'impôt et qui ne figurent pas parmi les dépendants sur la déclaration d'un autre contribuable (Norris et Standish, 1983; Statistique Canada, no 91-528F au catalogue, 1987).

ii) Émigration

L'utilisation des fichiers administratifs est essentielle dans l'estimation de l'émigration, le Canada ne disposant d'aucun système de collecte de données sur les émigrants. Le fichier d'impôt de Revenu Canada permet de retracer les émigrants puisqu'ils y sont définis par une adresse "hors Canada" sur la déclaration d'une année donnée et une adresse "au Canada" pour l'année précédente. Le fichier des allocations familiales, quant à lui, permet l'identification des enfants émigrants par le biais des changements d'adresse des bénéficiaires. Les deux fichiers administratifs sont donc utilisés conjointement pour estimer les effectifs provisoires et définitifs d'émigrants. La méthode d'estimation (identifiée comme la méthode des allocations familiales) est semblable à celle utilisée pour l'estimation provisoire de la migration interprovinciale et est définie comme suit:

$$E_j = \left| \frac{E_{j,0-17}}{P_{j,0-17}} \cdot F_c \cdot P_{j,18+} \right| + E_{j,0-17} \qquad (4)$$

$$F_c = \frac{\hat{E}_{c,18+}}{\hat{P}_{c,18+}} \div \frac{\hat{E}_{c,0-17}}{\hat{P}_{c,0-17}} \qquad (5)$$

$$E_c = \sum_{j=1}^{12} \left| E_j \right| \qquad (6)$$

où:

E_j = nombre annuel estimé d'émigrants de la province j

E_c = nombre annuel estimé d'émigrants du Canada

$E_{j,0-17}$ = nombre d'émigrants de la province j, âgés de 0 à 17 ans inclusivement et admissibles à l'allocation familiale

$P_{j,0-17}$ = nombre d'enfants admissibles à l'allocation familiale dans la province j

$P_{j,18+}$ = population adulte estimée de la province j, obtenue en soustrayant le nombre d'enfants admissibles à l'allocation familiale de la population totale estimée

F_c = facteur d'ajustement annuel, servant à estimer l'émigration des adultes selon le fichier d'impôt de Revenu Canada

$\hat{E}_{c,18+}$ et $\hat{E}_{c,0-17}$ = nombres estimés d'adultes et d'enfants émigrants du Canada, selon le fichier d'impôt de Revenu Canada

$\hat{P}_{c,18+}$ et $\hat{P}_{c,0-17}$ = populations estimées des adultes et des enfants pour le Canada, au 1er juin, Division de la démographie.

iii) Population totale

On obtient les estimations trimestrielles et annuelles de la population totale du Canada, des provinces et des territoires ainsi que les estimations annuelles pour les divisions de recensement et les régions métropolitaines de recensement par la méthode des composantes. Au niveau national, le nombre de naissances et d'immigrants est ajouté à la population de base (soit celle du dernier recensement du Canada), et le nombre de décès et d'émigrants en est retranché. Aux niveaux provincial et local, on tient également compte des estimations de la migration interne.

3. ÉVALUATION DE LA QUALITÉ DES ESTIMATIONS DE L'ÉMIGRATION ET DE LA MIGRATION INTERPROVINCIALE PRODUITES AU MOYEN DES FICHIERS ADMINISTRATIFS

Chacune des composantes de l'accroissement démographique (les naissances, les décès, les migrations internationales et interprovinciales) est susceptible de présenter certaines erreurs. Toutefois, on peut considérer que les données relatives aux naissances, aux décès et même à l'immigration sont assez précises, tandis que celles relatives à l'émigration et à la migration interprovinciale le sont moins, même si les méthodes à la base de leur calcul ont été remises à jour en 1982 (voir Statistique Canada, no. 91-528F au catalogue, 1987). Le degré de précision des estimations d'émigrants et de migrants interprovinciaux

obtenues par la méthodologie présentée dans la section précédente au moyen des données des fichiers administratifs doit donc être évalué.

Données sur l'émigration

Le tableau 1 compare les estimations des émigrants du Canada établies selon diverses méthodes et à partir de différentes sources de données, pour les périodes 1976-1981 et 1981-1986. Pour la période 1981-1986, l'estimation des émigrants par la méthode résiduelle, sans ajustement des effectifs recensés pour le sous-dénombrement, est de beaucoup supérieure à l'estimation dérivée de la méthode des allocations familiales. Comme les données sur les naissances, les décès et l'immigration sont considérées comme des renseignements précis, la plus forte estimation de l'émigration par la méthode résiduelle ne peut être attribuable qu'à la différence dans les taux de sous-dénombrement des recensements de 1981 et 1986. Après correction des effectifs, de 2.01% pour le recensement de 1981 et 3.21% pour celui de 1986, le nombre d'émigrants estimé de façon résiduelle s'établit à 134,857. Ce nombre est le plus faible effectif estimé d'émigrants obtenu pour l'ensemble des méthodes (235,481 selon la méthode des allocations familiales et 165,272 selon les estimations basées sur le fichier d'impôt de Revenu Canada). Ce résultat est probablement imputable au fait que lorsqu'on estime le nombre d'émigrants de façon résiduelle, on ne tient pas compte du surdénombrement différentiel des recensements impliqués.

Tableau 1

Estimations des émigrants selon différentes méthodes, Canada, 1976-81 et 1981-86

Méthode	1976-81	1981-86
Résiduelle*		
(a) sans ajustement	277,558	476,373
(b) avec ajustement pour le sous-dénombrement	196,955 (1)	134,857 (1)
(c) avec ajustement pour le sous-dénombrement net	194,155 (2)	218,148 (2)
Fichier d'impôt de Revenu Canada**	207,420	165,272
Méthode des allocations familiales	278,624	235,481
Contre-vérification des dossiers***	296,724	288,376

* Méthode résiduelle:
Émigrants = ([naissances - décès] + [immigrants]) - accroissement intercensitaire de la population entre t et t+5.

** Méthode directe, sans recours au fichier des allocations familiales.
L'effectif estimé d'émigrants est alors égal à la somme de $\hat{E}_{C,18+}$ et $\hat{E}_{C,0-17}$, tels que définis à la page 441.

*** Émigrants retracés dans l'échantillon de la contre-vérification des dossiers du recensement, pondérés à l'ensemble de la population (voir le Bulletin d'information à l'intention des utilisateurs, no. 2, juillet 1988).

(1) Ajustement des chiffres des recensements en fonction des taux de sous-dénombrement: 2.04% pour le recensement de 1976, 2.01% pour celui de 1981 et 3.21% pour celui de 1986.

(2) Ajustement des chiffres des recensements de 1976, 1981 et 1986 en fonction des taux de sous-dénombrement net s'établissant à 1.53%, 1.51% et 2.40% respectivement. Les taux de sous-dénombrement net représentent environ 75% des personnes non recensées estimées au moyen de la contre-vérification des dossiers.

Source: Division de la démographie, Statistique Canada.

Le taux de surdénombrement des recensements de 1981 et 1986 est inconnu. Cependant, on peut supposer qu'il est similaire à celui observé aux États-Unis et qu'il s'établit à 25% du taux de sous-dénombrement. Après correction des chiffres des recensements de 1981 et de 1986 pour les taux de couverture nette, correction de 1.51% et de 2.40% respectivement, l'estimation des émigrants est voisine de celle dérivée de la méthode des allocations familiales, soit 218,148 relativement à 235,481.

Les estimations produites au moyen des mêmes méthodes, pour la période 1976-1981, ne corroborent pas ces conclusions. Loin d'être similaires, les effectifs estimés par la méthode résiduelle avec ajustement pour le sous-dénombrement net (194,155), voisins de ceux basés sur le fichier d'impôt de Revenu Canada (207,420), sont de beaucoup inférieurs aux émigrants estimés par la méthode des allocations familiales (278,624) ou par la contre-vérification des dossiers (296,724).

Une des sources possibles d'erreur dans la méthode des allocations familiales est liée au facteur F_C, c'est-à-dire le rapport du taux d'émigration des adultes à celui des enfants, calculé à partir du fichier d'impôt de Revenu Canada.

Le tableau 2 présente les estimations des émigrants du Canada pour la période 1981-86, selon la méthode des allocations familiales dans laquelle le facteur F_C a été calculé, annuellement pour cinq ans, à partir de diverses sources de données. On peut remarquer que les facteurs F_C dérivés du fichier d'impôt de Revenu Canada sont inférieurs à 1, tandis que ceux calculés d'après les trois autres sources de données (c'est-à-dire les données sur la migration interprovinciale tirées du fichier d'impôt sur le revenu, les renseignements contenus dans les fichiers d'immigration ainsi que les données sur les Canadiens émigrant aux États-Unis) sont supérieurs à 1. En conséquence, le nombre d'émigrants estimé selon ces dernières sources de données est supérieur au nombre officiel d'émigrants (235,481).

Pour chaque source utilisée, la valeur du facteur F_C varie d'une année à l'autre. Le sens des variations dans le temps n'est pas toujours similaire, mais, règle générale, l'ordre des sources est stable.

Les valeurs du facteur F_C pour les Canadiens émigrant aux États-Unis sont relativement élevées: le nombre d'adultes émigrant vers ce pays est de 23% à 42% supérieur au nombre d'enfants émigrants. Cette observation n'est pas surprenante, puisque les états du sud des États-Unis ont toujours attiré les retraités canadiens. Par conséquent, la valeur du facteur F_C établie d'après les données sur l'émigration vers les États-Unis pourrait ne pas convenir à l'estimation des émigrants quittant le Canada vers des pays autres que les États-Unis.

Tableau 2

Estimations des émigrants du Canada selon la méthode des allocations familiales, avec F_C (rapport du nombre d'adultes émigrants au nombre d'enfants émigrants) basé sur différentes sources de données, 1981-1986

Source de données	Valeur du facteur F_C					Nombre d'émigrants estimés 1981-86
	1981-82	1982-83	1983-84	1984-85	1985-86	
1. Fichier d'impôt de Revenu Canada	0.8698	0.8768	0.9052	0.8592	0.8592	235,481
2. Données sur la migration interprovinciale tirées du fichier d'impôt de Revenu Canada*	1.0760	1.1000	1.0664	1.0290	1.0029	265,816
3. Données sur l'immigration*	1.0801	1.0926	1.1723	1.1254	1.0694	275,762
4. Données sur les émigrants canadiens aux É.-U.	1.2300	1.2774	1.3196	1.3745	1.4232	316,268

* Sous l'hypothèse que la structure par âge des émigrants est identique à celle des immigrants ou des migrants interprovinciaux, selon le cas.

Source: Division de la démographie, Statistique Canada;

Les valeurs du facteur F_C basé sur les données de migration interprovinciale obtenues du fichier d'impôt de Revenu Canada, laissent quant à elles supposer que le nombre d'adultes migrants a dépassé de 0% à 10% le nombre d'enfants migrants entre 1981 et 1986. Chez ces adultes migrants, il pourrait y avoir eu une plus forte proportion de jeunes adultes. Par conséquent, cette source de données est aussi très spécifique et ne convient pas au calcul du facteur F_C.

Puisque, selon certains auteurs (Beaujot et Rappak, 1988), il y a corrélation entre les flux d'émigrants et d'immigrants, on peut calculer le facteur F_C au moyen du fichier d'immigration. Les valeurs du facteur F_C ainsi établies se situent entre les valeurs basées sur les données de migration interprovinciale et celles correspondant aux émigrants vers les États-Unis. Cependant, contrairement à ces deux dernières sources qui se sont révélées non convenables pour le calcul de F_C, le fichier d'immigration est une source tout à fait adéquate. Son utilisation en lieu et place du fichier d'impôt de Revenu Canada dans le calcul de F_C fait passer le nombre d'émigrants estimé selon la méthode des allocations familiales à 275,762, effectif voisin de celui estimé par la méthode de contre-vérification des dossiers de 1986 (288,376), et par conséquent, réduit la population totale estimée en 1986 de 34,739 individus. Il en résulte que l'erreur en fin de période pour 1986, soit la différence entre l'estimation de la population et les chiffres du recensement, pour le Canada, diminue de 0.95% à 0.79%.

En résumé, il semble qu'on pourrait améliorer les estimations de l'émigration en basant le calcul du facteur F_C non plus sur le fichier d'impôt de Revenu Canada mais plutôt sur les données d'immigration du Ministère de l'Emploi et de l'Immigration du Canada.

Données sur la migration interprovinciale

Pour vérifier la précision des estimations définitives de la migration interprovinciale tirées du fichier d'impôt de Revenu Canada, on peut effectuer deux genres d'évaluation: i) l'analyse de la cohérence des deux séries de données sur la migration interprovinciale,

celle basée sur le fichier d'impôt de Revenu Canada et celle basée sur le fichier des allocations familiales; ii) la comparaison entre les chiffres du recensement de 1986 et la population totale de 1986 estimée au moyen de l'une et l'autre des séries.

Le tableau 3 compare les estimations de la migration interprovinciale nette produites selon trois méthodes: celle utilisant le fichier d'impôt de Revenu Canada, celle basée sur le fichier des allocations familiales et la méthode résiduelle. Pour chacune des provinces, les estimations produites au moyen du fichier d'impôt de Revenu Canada et du fichier d'allocations familiales sont cohérentes, la migration nette variant toujours dans le même sens. Pour les deux séries de données, la migration nette pour la période 1981-1986 est positive pour les quatre mêmes provinces et négative pour les autres.

Tableau 3

Estimations de la migration interprovinciale nette d'après le fichier des allocations familiales, le fichier d'impôt et la méthode résiduelle,* Canada et provinces, 1981-1986

Région géographique	Allocations familiales	Impôt sur le revenu	Estimations résiduelles
CANADA	0	0	-238,178
T.-N.	-14,837	-15,051	-26,111
Î.-P.-É.	293	751	-509
N.-É.	5,204	6,895	-4,095
N.-B.	-2,239	-65	-11,212
Qué.	-76,040	-81,254	-167,286
Ont.	115,497	121,767	57,147
Man.	-3,700	-2,634	-8,180
Sask.	-668	-2,974	-13,564
Alta.	-34,073	-31,676	-50,811
C.-B.	13,289	7,382	-12,418
Yukon	-2,381	-2,775	-1,643
T.N.-O.	-345	-366	504

* La méthode résiduelle d'estimation de la migration interprovinciale nette est la suivante:

Migration nette = accroissement de la population recensée entre la période t et la période t+5 - [(naissances - décès) + (immigration -émigration)]

Source: Division de la démographie, Statistique Canada.

Les estimations de la migration interprovinciale nette calculées au moyen des fichiers d'allocations familiales et d'impôt de Revenu Canada ne sont pas nécessairement comparables à celles obtenues de façon résiduelle. Par définition, au Canada, la somme de la migration interprovinciale nette devrait être égale à zéro, ce qui est effectivement le cas pour les estimations basées sur les fichiers administratifs. Toutefois, cette somme s'élève à 238,178 lorsqu'on utilise la méthode résiduelle. De plus, la différence entre la migration interprovinciale nette estimée de façon résiduelle et les estimations produites au moyen des fichiers d'impôt de Revenu Canada et des allocations familiales est considérable dans cinq provinces: Terre-Neuve, le Nouveau-Brunswick, le Québec, l'Ontario et l'Alberta.

La comparaison entre les chiffres des recensements et la population totale estimée aux mêmes dates en considérant les migrants interprovinciaux basés sur l'un ou l'autre des fichiers administratifs considérés, est riche d'enseignements.

On a utilisé le coefficient de variation (le rapport entre l'erreur-type de l'erreur absolue moyenne en fin de période pour les provinces et l'erreur absolue moyenne en fin de période) pour mesurer la précision relative des données sur la migration interne, en supposant que les autres composantes de l'accroissement démographique sont exactes. Sur le plan statistique, un coefficient de variation se situant entre 20% et 30% est considéré comme acceptable.

Le tableau 4 indique le coefficient de variation (calculé d'après les chiffres du tableau 5) entre d'une part les estimations de la population fondées sur les deux séries de données sur la migration interne et d'autre part les chiffres des recensements de 1971, 1976, 1981 et 1986. Avant la période de 1976-1981, les coefficients de variation relatifs aux données sur la migration tirées du fichier d'impôt étaient supérieurs de 50% à ceux correspondant au fichier d'allocations familiales. Il fallait s'attendre à cette observation puisque la méthode d'estimation de la migration à partir du fichier d'impôt n'en était, à ce moment-là, qu'à l'étape d'élaboration et que, par conséquent, les estimations des données sur la migration étaient produites à titre expérimental. De plus, le facteur f_j (rapport du nombre d'adultes émigrants à celui des enfants) servant à l'estimation du nombre de migrants interprovinciaux était basé sur les données de mobilité tirées du recensement. Cette méthode d'estimation a été jugée depuis lors moins satisfaisante que la méthode actuelle. C'est pourquoi on constate, pour les périodes 1976-81 et 1981-86, une nette diminution des écarts entre les coefficients de variation correspondant aux deux types de fichiers.

Tableau 4

Coefficients de variation de l'erreur absolue moyenne en fin de période entre les estimations de la population et les chiffres de recensement chez les provinces (n=10), selon la source des estimations de la migration interprovinciale, 1966-1971, 1971-1976, 1976-1981 et 1981-1986

Période (t,t+5)	Source	EAM (t+5)	Écart-type	C.V.
		(1)	(2)	(3)=[(2)-(1)] x 100
1966-1971	Impôt	0.91	0.2863	31
	A.F.	1.33	0.2642	20
1971-1976	Impôt	0.44	0.1317	30
	A.F.	0.97	0.2135	22
1976-1981	Impôt	0.69	0.2463	36
	A.F.	0.86	0.2855	33
1981-1986	Impôt	1.07	0.1496	14
	A.F.	1.01	0.1570	16

Nota: EAM : Erreur absolue moyenne en fin de période
C.V. : Coefficient de variation
Impôt : Fichier de données sur la migration interprovinciale (Fichier d'impôt de Revenu Canada)
A.F. : Fichier de données sur la migration interprovinciale (Fichier des allocations familiales)

Source: Calculé à partir des chiffres du tableau 5.

Tableau 5

Erreur en fin de période entre les divers genres d'estimations de la population et les chiffres du recensement, selon la province et le territoire, pour les estimations établies au moyen de deux séries de données sur la migration interprovinciale, 1971, 1976, 1981 et 1986

Région géographique	Erreur en fin de période en pourcentage(1)							
	1971		1976		1981		1986	
	Impôt	A.F.	Impôt	A.F.	Impôt	A.F.	Impôt	A.F.
Terre-Neuve	-2.08	-1.64	0.49	1.34	1.63	2.30	1.97	2.01
Île-du-Prince-Édouard	-2.09	-2.01	0.17	2.11	-0.05	1.02	0.99	0.63
Nouvelle-Écosse	-1.68	-2.39	-0.20	1.18	0.30	0.40	1.24	1.04
Nouveau-Brunswick	-1.93	-2.65	-1.29	1.81	0.13	0.54	1.58	1.04
Québec	-0.33	-0.97	-0.05	-0.18	-0.30	-0.07	1.32	1.40
Ontario	0.11	0.99	0.15	0.16	0.64	0.37	0.72	0.65
Manitoba	0.29	0.38	-0.27	0.39	1.07	0.87	0.51	0.41
Saskatchewan	0.44	-0.33	0.45	0.37	-0.31	0.28	1.08	1.31
Alberta	-0.14	0.52	-1.07	-1.11	-2.39	-2.64	0.73	0.63
Colombie-Britannique	0.01	-1.34	0.28	-1.10	0.03	-0.07	0.59	0.79
Yukon	-5.36	-5.99	-0.87	3.79	-1.98	2.06	-4.78	-3.10
Territoires du Nord-Ouest	-2.12	2.64	-12.98	-3.39	-7.08	0.43	-1.44	-1.40
Erreur absolue moyenne								
10 provinces	0.91	1.33	0.44	0.97	0.69	0.86	1.07	1.01
Provinces et territoires	1.38	1.82	1.52	1.41	1.33	0.92	1.41	1.22

Nota: De 1976 à 1981, les données de Revenu Canada pour les enfants n'étaient disponibles que pour le groupe d'âge de 0 à 15 ans. Par conséquent, les facteurs $f_{(j,k)}$ ont été calculés au moyen du nombre de migrants faisant partie des groupes d'âge de 0 à 15 ans et de 16 ans et plus au lieu du nombre de migrants âgés de 0 à 17 ans et de 18 ans et plus.

(1) L'erreur en fin de période est calculée au moyen de l'équation suivante:

$$\text{Erreur en fin de période} = \frac{\text{Estimation - recensement}}{\text{recensement}} \times 100$$

Sources: Estimations de la migration interprovinciale selon les données sur l'impôt, (Impôt), Division des données régionales et administratives, Statistique Canada.
Estimations de la migration interprovinciale selon le fichier des allocations familiales (A.F.), Division de la démographie, Statistique Canada.

En 1981, le coefficient de variation correspondant au fichier d'impôt était supérieur de 9% à celui relatif au fichier des allocations familiales, tandis qu'en 1986, il lui était inférieur de 12%. Comme ces différences sont minimes, nous pouvons affirmer que les deux séries de données sur la migration sont tout à fait semblables. Elles permettent de produire des estimations provinciales de population et de présenter des erreurs en fin de période variant de façon similaire d'une province à l'autre. De plus, le fait que les coefficients de variation soient inférieurs à 20% en 1986 pour les deux types de fichiers, indique clairement que chacune des deux sources offre des données acceptables sur la migration interne.

Au niveau provincial, on observe une plus faible variation dans les erreurs en fin de période.

4. CONCLUSION ET DISCUSSION

Les fichiers d'allocations familiales et d'impôt de Revenu Canada jouent un rôle important dans l'établissement de séries cohérentes de données sur l'émigration et la migration interne, pour le Canada, les provinces et les territoires. Les estimations des émigrants et des migrants interprovinciaux produites à partir de ces fichiers pour la période 1981-1986 sont acceptables pour fins d'estimation de la population totale.

Pourtant, un problème demeure. Au niveau national, l'erreur en fin de période pour 1986 était supérieure à celle relevée pour les trois années de recensement antérieures, soit de 1971 à 1981. De plus, toujours en 1986, pour toutes les provinces, les estimations de la population étaient supérieures aux chiffres du recensement. Comment expliquer ces irrégularités?

La réponse à cette question est étroitement liée à la couverture de la population lors du recensement de 1981, qui a servi de population repère, et lors du recensement de 1986. Pour le Canada, le taux de sous-dénombrement en 1981, d'après la contre-vérification des dossiers, était estimé à 2.01%. Le taux correspondant pour le recensement de 1986 est, à 3.21%, beaucoup plus élevé. Par conséquent, le sous-dénombrement devient un facteur important d'erreur en fin de période.

BIBLIOGRAPHIE

Beaujot, R. et Rappak, J.P. (1988). *L'émigration du Canada: importance et interprétation.* Ottawa: Emploi et Immigration Canada.

Norris, D., Britton, M. et Verma, Ravi B.P. (1982). The Use of Administrative Records for Estimating Migration and Population. *Statistics of Income and Related Administrative Record Research: 1982*, Washington, D.C., Department of the Treasury, Internal Revenue Service.

Norris, D. et Standish, L.D. (1983). *Rapport technique sur la production de données migratoires à partir des dossiers d'impôt.* Rapport technique. Ottawa: Division de l'exploitation des données administratives, Statistique Canada.

Statistique Canada, no 91-528F au catalogue (1987). *Méthodes d'estimation de la population, Canada.* Ottawa: Ministère des Approvisionnements et Services Canada.

Statistique Canada (1988). Taux de sous-dénombrement provenant de la contre-vérification des dossiers de 1986. *Bulletin d'information à l'intention des utilisateurs*, no. 2.

Verma, Ravi B.P. et Parent, Pierre (1985). Vue d'ensemble des avantages et inconvénients des fichiers de données administratives choisis. *Techniques d'enquête*, Vol. 11, No. 2, 193-202.

Statistique Canada Symposium
sur les utilisations statistiques
des données administratives
novembre 1987

STATISTIQUES FONDÉES SUR LES DOSSIERS ADMINISTRATIFS AU MEXIQUE UNE ANALYSE DE LA SITUATION ACTUELLE

MA. ELENA FIGUEROA MARQUEZ[1]

RÉSUMÉ

L'Institut national de la statistique, de la géographie et de l'informatique (Instituto Nacional des Estadistica, Geografia e Informatica - INEGI) du Mexique a été créé en 1983. L'une de ses fonctions principale consiste à coordonner les activités reliées à la production de données statistiques au Mexique. À la fin de 1983, l'Institut débuta un processus de décentralisation avec la création de directions régionales, situées dans dix états fédéraux.

Actuellement, les statistiques de l'état civil sont recueillies et traitées à l'échelle régionale, et les données nationales sont calculées dans les bureaux centraux. On propose d'utiliser pour les statistiques sur la santé, l'éducation, culture, la justice et les relations de travail la même procédure que pour les statistiques de l'état civil. Plusieurs problèmes doivent encore être résolus: la mise à jour continue des répertoires d'informateurs, la formation continue du personel régional en collecte et codage des données, la formation en matière d'analyse et d'interprétation des données produites dans les bureaux régionaux, la coordination avec les institutions sociales pour la décentralisation de l'information qu'elles produisent, le contrôle de la qualité en ce qui a trait à la couverture et au traitement des données, la création de méthodes efficaces de coordination avec les bureaux régionaux. Le présent document traitera de ces problèmes et de l'orientation de ce domaine.

1. INTRODUCTION

Au Mexique, des statistiques sont établies depuis 1882 par le Bureau du Recensement et de la Statistique du Mexique, qui relève actuellement de l'Institut National de la Statistique, de la Géographie et de l'Informatique.

Bien que le Mexique ait dû faire face à divers problèmes résultant de l'insuffisance des ressources humaines et financières nécessaires à l'établissement de statistiques, il a quand même été possible d'effectuer dix recensements de la population et de produire une grande variété de données économiques et sociales à partir de différents systèmes de dossiers administratifs.

[1] MA. Elena Figueroa Marquez, Institut National de la Statistique de la Géographie et de l'Informatique, Insurgentes Sur 795, 120. PISO, Colonia Napoles, C.P. 03810, Mexique

Les recensements sont devenus plus perfectionnés car on les améliore tous les dix ans afin de tenir compte des résultats de recherches nationales et internationales, de l'évolution dans le domaine de l'informatique et des besoins en données fiables et à jour sur la population et d'autres sujets d'intérêt.

La production de statistiques au Mexique est une activité extrêmement complexe et coûteuse à cause de la grande superficie du pays et de ses caractéristiques géographiques, de son taux de croissance démographique élevé et des problèmes administratifs causés, notamment, par la centralisation du processus de production statistique.

Les dossiers administratifs n'ont donc pas beaucoup été utilisés comme source de statistiques, pour ces raisons et aussi parce qu'on accordait la priorité à la collecte de données par recensement.

En 1983, l'INEGI (Institut National de la Statistique, de la Géographie et de l'Informatique) a établi dix bureaux régionaux à travers le Mexique afin de promouvoir le développement de statistiques à l'échelle nationale et régionale.

Chaque bureau régional a la responsabilité d'environ trois états et vise surtout à rendre les services d'information sur les données accessibles aux utilisateurs locaux et à accélérer le processus de collecte et de dépouillement des données.

Les activités de chaque bureau régional sont coordonnées par un bureau central afin d'assurer une liaison complète pour les projets nationaux et régionaux.

Le bureau central de la statistique dépouille et traite les données des recensements démographiques et économiques, des enquêtes économiques et des dossiers administratifs portant sur des questions démographiques, sociales et économiques.

À cette fin, le bureau central a réparti ses services en cinq domaines d'intérêt.

Ces domaines sont les suivants:

1. Recensements
2. Statistiques à court terme
3. Comptes nationaux et statistiques économiques
4. Statistiques sociales et démographiques
5. Aide technique

Le personnel responsable de la collecte et du traitement des données tirées des dossiers administratifs est le responsable de l'établissement des statistiques sociales et démographiques. La division des statistiques démographiques produit des statistiques relatives aux naissances, aux décès foetaux, aux mariages, aux divorces, aux migrations et au tourisme international. La division des statistiques sociales est responsable de la production de statistiques dans les domaines suivants:

1. Culture: cinéma, musées et spectacles (théâtres, stades, etc.).
2. Relations de travail: conflits de travail et contrats de travail, mandats de grève et grèves.
3. Santé et sécurité sociale: services médicaux (publics et privés) et bien-être social.
4. Sécurité publique: délits et suicides.

On n'a pas commencé à produire ces statistiques au cours d'une même période. Les statistiques de l'état civil sont celles qui ont été établies les premières, soit en 1893.

En 1983, la division des statistiques démographiques et sociales a effectué une evaluation du processus de production statistique. Cette analyse a permis de cerner les problèmes suivants:

1. Les données de plusieurs sources sont fournies en retard.
2. Certaines sources des répertoires fournissent des données désuètes et incomplètes.
3. La centralisation des services donne lieu à un volume élevé de tâches de codage et, conséquemment, la production est ralentie.
4. La conception des questionnaires statistiques est à revoir et les instructions concernant le traitement électronique des données sont dépassées.
5. Les programmes de traitement sont désuets et donnent lieu à des problèmes d'exploitation.

Afin de régler quelques-uns des problèmes mentionnés ci-dessus, la division des statistiques démographiques et sociales a décidé de décentraliser les activités de production statistique en établissant des bureaux régionaux soumis aux normes et à la surveillance d'un organisme central. Ce processus de décentralisation a d'abord visé la production des statistiques de l'état civil en raison de leur volume et de leur importance.

Deux principaux secteurs s'occupent du processus de décentralisation de la production. Leurs fonctions sont les suivantes:

1. Secteur central - chargé de la normalisation du processus de production régionale et de l'intégration nationale des données.
2. secteur régional - chargé des opérations manuelles et électroniques ainsi que de l'analyse et de la diffusion des données à l'échelle nationale et régionale.

Le diagramme ci-joint illustre le processus actuel de production des données démographiques.

Une méthode de décentralisation semblable sera utilisée dans le cas des statistiques sociales. Divers organismes officiels ont été consultés afin de déterminer les modalités de transmission des données aux bureaux régionaux.

Afin d'améliorer et de faciliter le processus de décentralisation les aspects suivants nécessiteront une attention particulière :

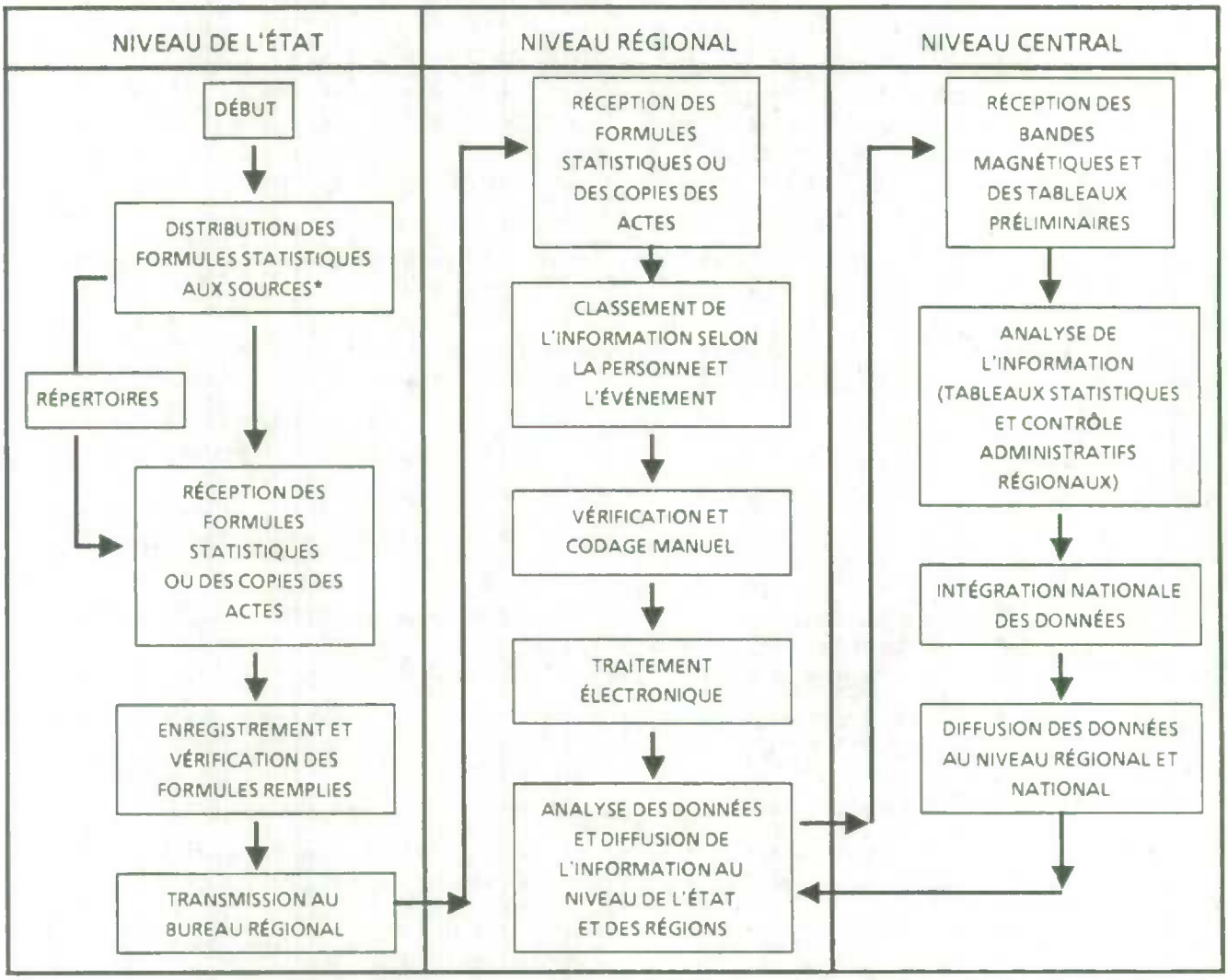
1. Mettre et maintenir à jour les répertoires de sources d'information.
2. Motiver le personnel des bureaux régionaux et assurer sa formation permanente, particulièrement en ce qui a trait à la collecte et au codage des données.
3. Former le personnel des bureaux régionaux en matière d'analyse et d'interprétation des données statistiques.
4. Effectuer des contrôles qualitatifs qui portent tant sur la couverture que sur le traitement des données.
5. Établir des méthodes de coordination efficaces avec le concours des bureaux régionaux.

6. Coordonner les efforts des établissements afin d'assurer leur participation efficace au processus de décentralisation des statistiques sociales.

Afin de répondre dans des délais raisonnables à la demande de données statistiques et, par conséquent, de publier des données plus à jour, la division des statistiques démographiques et sociales s'intéresse à l'élaboration d'indicateurs et de modèles statistiques qui permettraient d'établir des estimations des mesures requises pour satisfaire à ces besoins.

On accorde aussi de l'importance à l'évaluation des statistiques actuelles et à leur couplage avec des données d'autres sources telles que les recensements et les enquêtes.

PROCESSUS DE PRODUCTION DE DONNÉES POUR L'ÉTABLISSEMENT DE STATISTIQUES DE L'ÉTAT CIVIL
- DE 1985 JUSQU'À PRÉSENT -



* SEULEMENT POUR LES DÉCÈS, LES DÉCÈS FOETAUX ET LES DIVORCES. AU DÉBUT DE 1988, LES DONNÉES SUR LES DÉCÈS SERONT TIRÉES DES DOCUMENTS ET ACTES OFFICIELS DE DÉCÈS QUI SERONT ENVOYÉS.

SESSION VIII: COMMUNICATIONS OFFERTES

Président: John Coombs, Statistique Canada

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

MISE À JOUR DES PROBABILITÉS DE SÉLECTION DES DÉCLARATIONS D'IMPÔT DANS LE CADRE DU PROGRAMME DE LA STATISTIQUE DU REVENU DES SOCIÉTÉS

SUSAN HINKINS, HOMER JONES et FRITZ SCHEUREN[1]

RÉSUMÉ

Dans le cadre du programme de la statistique du revenu des sociétés (Corporate Statistics of Income Program - CSIP), la taille de l'échantillon de déclarations d'impôt se situe essentiellement autour de 90,000 annuellement. À cause des variations de la population des sociétés et de la croissance globale, les probabilités de sélection sont systématiquement révisées à la baisse à chaque année. Il est alors permis de s'interroger sur les répercussions de ces rajustements sur les estimations transversales, les estimations des faibles variations d'une année à l'autre et la composition de l'échantillon longitudinal. Le présent document explique tout d'abord l'origine du plan de sondage utilisé actuellement pour le CSIP et montre comment ce plan a été appliqué au fil des années. Les auteurs évaluent ensuite diverses méthodes de mise à jour et les comparent à la méthode actuelle.

1. INTRODUCTION

Depuis 1951, l'Internal Revenue Service des États-Unis échantillonne des déclarations d'impôt de sociétés pour produire des estimations annuelles des variables économiques et fiscales. Au cours des années, l'IRS a dû modifier ses méthodes de collecte de données et ses méthodes d'estimation à cause des variations de la population des sociétés (tendances économiques) et des modifications apportées aux lois fiscales.

Le perfectionnement des techniques informatiques et statistiques a aussi amené une modification des méthodes. Malheureusement, à cause d'un certain nombre de contraintes pratiques -- croissance soutenue de la population de déclarants, limitation du nombre de déclarations à dépouiller à cause des compressions budgétaires, et brève durée de la période de dépouillement des déclarations échantillonnées -- les taux de sondage n'ont cessé de diminuer depuis le début. Cette diminution constante des taux de sondage fait qu'il est plus difficile de garder les mêmes sociétés dans l'échantillon pendant un certain nombre d'années et de mesurer avec précision les variations d'une année à l'autre.

Ce document a pour but d'exposer les modifications qui ont été apportées récemment au plan de sondage et qui seraient susceptibles d'améliorer les estimations de la variation d'une année à l'autre. Quelques-unes des modifications proposées ont déjà été appliquées et d'autres sont envisagées pour l'avenir. Nous allons aussi examiner brièvement plusieurs caractéristiques pertinentes du plan de sondage du CSIP et les modifications récentes dont

[1] Susan Hinkins, Homer Jones et Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Ave., NW, Washington, DC 20224, États-Unis.

il a fait l'objet de même que le problème soulevé par l'estimation des variations d'une année à l'autre. Nous ne pouvons évidemment considérer les modifications apportées au plan de sondage sans tenir compte des effets que cela peut avoir sur les estimations transversales. Nous analysons ces effets dans la troisième section et faisons certaines réflexions sur la recherche future dans la dernière partie.

2. CONTEXTE

L'objectif primordial de l'échantillonnage de déclarations d'impôt de sociétés est de produire des données qui permettent d'estimer avec précision des variables économiques à chaque année. Par exemple, des estimations transversales ont été produites récemment à propos des variables suivantes:

- total de l'actif des sociétés enregistrées en 1984 ($11.1 billions);
- total des revenus des sociétés enregistrées en 1984 et dont l'actif est inférieur à $100,000 ($357.1 milliards);
- total du bénéfice net (moins les pertes) des sociétés minières enregistrées en 1984 et dont l'actif total est de $50,000,000 ou plus ($779.4 millions).

À l'origine, toutes les déclarations d'impôt des sociétés étaient prélevées avec une probabilité égale à 1. Puis avec la prolifération des sociétés, l'échantillonnage est devenu nécessaire. Au début, l'échantillonnage était effectué manuellement dans quelque 64 bureaux de district par de nombreuses personnes qui avaient peu de connaissances en statistique. Le plan de sondage devait évidemment être aussi simple que possible. Avec l'arrivée de l'ordinateur et la réduction du nombre de centres d'échantillonnage (une dizaine actuellement dans le réseau de l'Internal Revenue Service), on a pu accroître le nombre de variables de stratification et le degré de complexité du plan de sondage.

L'univers des déclarations d'impôt des sociétés est très asymétrique étant donné qu'un nombre relativement faible de grandes sociétés représentent à elles seules plus de la moitié de l'actif total et du bénéfice net de toutes les sociétés. En 1984, par exemple, 56% des sociétés les moins importantes ne représentaient que 0.5% de l'actif total aux États-Unis tandis que 0.11% des sociétés les plus importantes représentaient 75% de ce même actif. L'IRS utilise donc un plan de sondage stratifié et les très grandes sociétés sont prélevées avec une probabilité égale à 1.

On trouvera une documentation abondante mais très condensée sur le plan de sondage actuel dans **Statistics of Income — 1984 Corporation Income Tax Returns.** Harte (1982), Jones et McMahon (1984) et Oh et Scheuren (1987) donnent également une description du plan et des méthodes d'estimation.

Le plan actuel (1985) utilise deux variables de stratification: actif total (AT) et bénéfice (ou perte) net (BN). La première variable reflète la valeur de l'actif total et d'autres items du bilan pouvant figurer sur la déclaration d'impôt tandis que la seconde reflète la valeur des composantes du revenu qui servent à établir le revenu total et les déductions totales. À chaque variable de stratification correspond une série de classes propre et les strates sont définies par la classe la plus importante dans chaque cas. Le genre d'entreprise sert également à définir les strates, surtout en ce qui concerne les entreprises de taille plus élevée, parce qu'un grand nombre d'entreprises financières ont un actif total disproportionné par rapport à celui des entreprises non financières. On détermine ensuite les taux de sondage à l'aide de la répartition optimum de Neyman. Nous traitons l'échantillon comme s'il s'agissait d'un échantillon aléatoire stratifié mais il se rapproche plus en fait d'un échantillon de Poisson stratifié (comme le définit Sunter, 1986).

3. ESTIMATION DE LA VARIATION D'UNE ANNÉE À L'AUTRE

Les caractéristiques du plan de sondage que nous venons de décrire se rapportent uniquement à l'objectif primordial de ce plan, soit produire des estimations annuelles exactes; cependant, comme un nouvel échantillon est prélevé chaque année, il serait également souhaitable d'estimer les variations d'une année à l'autre. Le plan de sondage est déjà conçu de manière à permettre ce genre d'estimation et nous verrons dans les paragraphes qui suivent comment se fait cette estimation et quelles sont les façons inédites d'améliorer les estimations ainsi obtenues.

À titre d'exemple, définissons la "variation de l'actif total d'une année à l'autre" comme la variable d'intérêt qui nous permettra de mesurer la variation et désignons-la par ΔAT. Pour simplifier le problème, nous ne considérons que la partie de la population qui est soumise à l'échantillonnage (c'est-à-dire les sociétés de taille moyenne ou petite) et ne reconnaissons en l'occurrence qu'une seule variable de stratification, soit AT. Le tableau 1 donne les strates selon l'actif total et les taux de sondage correspondants pour 1984 et 1985. (Ces taux ont été arrondis à un chiffre significatif.)

Tableau 1

Taux de sondage fondés sur l'actif total (AT)
(arrondis à un chiffre significatif)

Strates	AT (actif total)	Taux de sondage	
		1984	1985
(1)	(2)	(3)	(4)
1	0 - $50,	.002	.002
2	$50, - $100,	.002	.002
3	$100, - $250,	.007	.006
4	$250, - $500,	.01	.01
5	$500, - $1,000,	.02	.02
6	$1,000, - $2,500,	.07	.06
7	$2,500, - $5,000,	.1	.1
8	$5,000, - $10,000,	.2	.2
9	$10,000, - $25,000,	.7	.6

Pour estimer ΔAT, nous nous reportons aux classes (ou cases) d'un tableau à double entrée qui met en rapport les strates de 1984 et celles de 1985 (tableau 2). La première ligne du tableau indique les créations d'entreprises, ou sociétés qui ont vu le jour en 1985. La première colonne indique les disparitions d'entreprises, c'est-à-dire les sociétés qui n'étaient plus en affaires en 1985 ou qui ont été absorbées. Le corps du tableau est formé des classes de variation pour les sociétés incluses dans la population les deux années. Les cases de la diagonale représentent les sociétés qui étaient dans la même strate en 1984 et en 1985. Les autres cases du tableau représentent les sociétés qui ont changé de strate entre 1984 et 1985.

Tableau 2

Tableau à double entrée de l'actif total (AT) mettant en rapport les strates de 1984 et celles de 1985

Strates 1984 \ Strates 1985	1	2	3	4	5	6	7	8	9	Taux de sondage 1984
Créations										s.o.
1	....									.002
2		....								.002
3			....							.007
4				....						.01
5					....					.02
6						....				.07
7							....			.1
8								....		.2
9									....	.7
Taux de sondage 1985	s.o.	.002	.002	.006	.01	.02	.06	.1	.2	.6	s.o.

Note: s.o. = sans objet

Si nous concevions le plan de sondage de 1985 pour estimer ΔAT, nous le ferions de manière à obtenir un échantillon de nouvelles entreprises représentatif et un échantillon de sociétés existantes aussi semblable que possible à celui de l'année précédente. Pour mesurer le degré de chevauchement des deux échantillons, il faudrait idéalement tirer les mêmes sociétés d'une année à l'autre. S'il n'était pas possible d'utiliser le même échantillon qu'en 1984, à tout le moins nous chercherions probablement à échantillonner des sociétés qui ne se trouvent pas dans les cases de la diagonale (c'est-à-dire des sociétés qui auraient connu une forte variation de leur actif total).

Le plan tranversal donne un échantillon représentatif en ce qui concerne les entreprises nouvellement créées; toutefois, pour ce qui a trait aux sociétés qui étaient en affaires au cours des deux années, il y a très peu de chances de retrouver les mêmes sociétés dans l'échantillon d'une année à l'autre si l'on procède de façon aléatoire, les seules exceptions étant les grandes sociétés stables qui sont prélevées au complet. Si l'on considère les échantillonnages annuels comme des événements indépendants, le taux de sondage réel pour l'échantillonnage d'une société deux années consécutives est égal au produit des taux de sondage des deux années. Si nous prenons par exemple la case qui

représente les sociétés qui appartenaient à la strate 1 en 1984 (AT inférieur à $50,000) et à la strate 3 en 1985 (AT supérieur à $100,000 mais inférieur à $250,000), le taux réel pour l'échantillonnage de telles sociétés deux années consécutives serait de .000012 = (.002)*(.006).

Le plan de sondage utilisé pour le CSIP permet d'estimer la variation d'une année à l'autre en accroissant toujours plus chaque année le degré de chevauchement des échantillons. Un numéro d'identification de l'employeur (NIE) est attribué à chaque société. Chaque NIE s'accompagne d'un nombre pseudo-aléatoire. Si ce nombre est inférieur à une constante qui détermine le taux de sondage, la société est échantillonnée. Par conséquent, si la société est échantillonnée une première fois, elle le sera une deuxième fois si le taux de sondage est au moins aussi élevé qu'à la première occasion. (Cette méthode est analysée dans Harte, 1986). Ainsi, le taux de sondage réel pour l'échantillonnage d'une société deux années consécutives est égal au moindre des taux de sondage des deux années si l'on utilise le NIE. Si nous reprenons l'exemple précédent, nous voyons qu'en utilisant le NIE, le taux de sondage réel serait le minimum de (.002, .006), soit .002; ce taux diffère considérablement de celui calculé précédemment (.000012) dans le cas où l'on procède de façon aléatoire.

Sunter (1986) désigne cette méthode comme l'échantillonnage longitudinal implicite et montre qu'elle maximise le degré de chevauchement des échantillons après plusieurs répétitions; toutefois, lorsque la plupart des variations sont faibles, la grande majorité des sociétés qui se retrouvent dans l'échantillon une deuxième année consécutive sont celles qui sont demeurées dans la même strate. Comme l'indique le tableau 3, l'analyse ne met pas l'accent sur les sociétés qui ont connu une forte variation de leur actif total.

Les cases situées au-dessus de la diagonale représentent les sociétés dont l'actif total a augmenté de 1984 à 1985. Le degré de chevauchement des échantillons dans ce cas est faible car les taux de sondage en 1984 étaient moins élevés. Pour accroître le degré de chevauchement des échantillons, il aurait fallu prévoir en 1984 quelles sociétés devaient connaître une croissance de leur actif dans les années suivantes. Il est permis de douter que cet exercice de prévision puisse être réellement fructueux un jour.

Les cases situées au-dessous de la diagonale représentent les sociétés qui ont connu une baisse de leur AT de 1984 à 1985; c'est pourquoi le taux de sondage pour 1985 est inférieur au taux pour 1984. Bon nombre des sociétés représentées par ces cases seraient donc incluses dans l'échantillon de 1984 mais non dans celui de 1985. Une façon d'accroître le degré de chevauchement des échantillons en l'occurrence serait de réexaminer les résultats de 1984 avant de faire l'échantillonnage pour 1985.

Au cours des dernières années, cette méthode a permis d'ajouter des variables de stratification dans le plan de sondage afin d'accroître le nombre de sociétés dans les deux échantillons. Ainsi, par suite d'une modification récente du plan de sondage, on utilise comme variable de stratification non plus l'actif total mais le plus élevé de l'actif total ou de l'actif au début de l'année. On se trouve ainsi à "revenir" sur 1984 -- puisque l'actif au début de 1985 est normalement égal à l'actif à la fin de 1984 (actif total) --et à accroître par conséquent le degré de chevauchement des échantillons en ce qui concerne les sociétés situées au-dessous de la diagonale.

Tableau 3

Taux de sondage réels pour la formation d'échantillon identiques d'une année à l'autre; en utilisant le NIE comme valeur de départ

Strates 1984	Strates 1985									Taux de sondage 1984
	1	2	3	4	5	6	7	8	9	
1	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002
2	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002
3	.002	.002	.006	.007	.007	.007	.007	.007	.007	.007
4	.002	.002	006	.01	01	.01	.01	.01	.01	.01
5	.002	.002	.006	.01	.02	.02	02	.02	.02	.02
6	.002	.002	.006	.01	.02	.06	.07	.07	.07	.07
7	.002	.002	.006	.01	.02	.06	.1	.1	.1	.1
8	.002	.002	.006	.01	.02	.06	.1	.2	.2	.2
9	.002	.002	.006	.01	.02	.06	.1	.2	.6	.7
Taux de sondage 1985	.002	.002	.006	.01	.02	.06	.1	.2	.6	s.o.

Notes: s.o. = sans objet.
Les taux de la diagonale sont soulignés

4. EFFETS DU PLAN

Comme nous l'avons déjà souligné, des modifications ont été faites pour améliorer les estimations de la variation d'une année à l'autre et d'autres correctifs sont envisagés pour les années à venir. Lorsqu'on examine les diverses solutions (tant pratiques que théoriques) proposées pour la modification du plan de sondage, il faut s'interroger nécessairement sur les effets que pourraient avoir ces modifications sur les estimations (annuelles) transversales. De combien augmenterait la variance des estimations annuelles de AT et de BN?

Dans cette section, nous calculons les effets de plan selon le modèle classique de la stratification en fonction de plusieurs variables, comme le décrit Cochran (1977). Trois variables sont considérées: AT, BN et ΔAT (variation de l'actif total de 1984 à 1985). Rappelons que les deux premières sont les variables de stratification utilisées dans le plan transversal actuel.

On estime que l'utilisation du bénéfice (perte) net (BN) comme variable de stratification peut déjà contrôler parliellement ΔAT puisqu'il est censé exister une forte

corrélation entre BN et ΔAT. Le degré de cette corrélation n'est toutefois pas encore établi et des recherches seront nécessaires à cette fin. Pour les besoins de notre analyse, nous avons supposé un degré de corrélation modeste ($r \approx 0.5$); les résultats qui seraient conformes au plan actuel pourraient donc ne pas être aussi réalistes qu'on le croit. Il est tout aussi important d'examiner ces calculs selon d'autres modèles.

Dans notre analyse, nous avons déterminé les taux de sondage selon la méthode utilisée pour la formation de l'échantillon de déclarations d'impôt de sociétés. On suppose que l'écart type dans chaque strate est proportionnel à l'intervalle de valeurs de la strate et on applique la répartition optimum de Neyman. Lorsque AT et BN sont utilisées concurremment, la stratification se fait selon la variable qui détermine la classe la plus importante; lorsque la stratification se fait selon plusieurs variables, on fait la moyenne des taux de sondage calculés à l'aide des écarts types.

Nous avons examiné les sept plans de sondage qui peuvent découler des trois variables de stratification considérées (variables prises individuellement ou combinées). La figure 1 illustre les plans de sondage qu'il est possible d'obtenir. Les sommets du triangle représentent les plans de sondage qui reposent sur une seule variable de stratification; le sommet du bas correspond au plan de sondage fondé uniquement sur ΔAT, etc. Les points situés sur les droites liant les sommets correspondent aux plans fondés sur une combinaison de ces variables de stratification. Les plans qui nous intéressent le plus sont:

- le plan fondé sur les trois variables (AT, BN et ΔAT), qui est le cas à l'étude dans cette section;
- le plan fondé sur AT et BN, qui correspond au plan actuel;
- les trois plans fondés respectivement sur une variable, qui représentent le plan optimal pour chaque variable.

À l'aide de ce diagramme, nous analysons les effets du plan sur les trois variables, c'est-à-dire le rapport de variances selon chaque plan pour l'estimation de AT, de BN et de ΔAT. (Les variances en question sont des variances conditionnelles, qui sont subordonnées à la taille des strates obtenue.) Les plans sont comparés au plan de sondage considéré, c'est-à-dire celui qui utilise les trois variables de stratification; par définition, l'effet de ce plan est donc toujours égal à 1.00.

La figure 2 montre les effets de plan (e.p.) pour l'estimation de AT. En l'occurrence, le plan "optimal" est celui où la stratification se fait uniquement en fonction de AT (effet de plan le moins élevé: 0.96). L'effet de plan découlant de l'utilisation de la seule variable AT comme variable de stratification est défini comme le rapport entre la variance de l'estimation de AT calculée selon le plan fondé sur AT et la variance de l'estimation de AT calculée selon le plan fondé sur AT, BN et ΔAT. Le plan actuel (fondé sur AT et BN) n'est pas optimal et se traduit par une augmentation de la variance de AT (e.p. = 1.04). On remarquera que cette dernière valeur est supérieure à l'effet de plan calculé pour le plan fondé sur les trois variables de stratification à la fois (1.00). Ainsi, le fait d'ajouter ΔAT au plan de sondage actuel pourrait améliorer l'estimation de AT.

La figure 3 montre les effets du plan pour l'estimation de BN. L'addition de ΔAT au plan actuel ne semble pas avoir beaucoup d'effet sur la variance de l'estimation de BN; les effets de plan sont essentiellement les mêmes.

La figure 4 montre les effets du plan pour l'estimation de ΔAT. Le plan optimal en l'occurrence serait celui où la stratification se fait uniquement en fonction de ΔAT puisque c'est le cas où l'effet de plan est le moins élevé. Il convient de souligner le rendement du plan actuel: une variance qui n'est que de 11% supérieure à celle observée pour le plan optimal. (On suppose bien sûr dans les circonstances un chevauchement parfait des échantillons pour les deux années.) Ce rapport dépend largement de l'hypothèse

d'une corrélation modérée entre BN et ΔAT; si cette corrélation est exagérée, l'effet du plan actuel est sous-estimé et la variance de l'estimation de ΔAT peut être beaucoup plus élevée. Le plan fondé sur les trois variables à la fois donne une variance qui n'est que de 3.5% supérieure à celle obtenue par le plan optimal; il constitue une amélioration par rapport au plan actuel.

Ces résultats montrent que nous pouvons envisager des moyens d'améliorer les estimations de ΔAT (variation d'une année à l'autre) sans pour autant modifier les estimations annuelles transversales. De fait, dans les trois cas d'estimation que nous venons de voir, le plan considéré était aussi bon sinon meilleur que le plan actuel.

Figure 1. - - Plan de sondage possibles

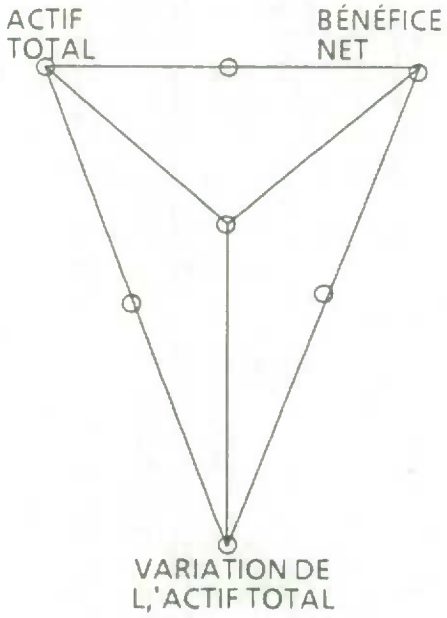


Figure 2. - - Effets du plan pour l'estimation de AT

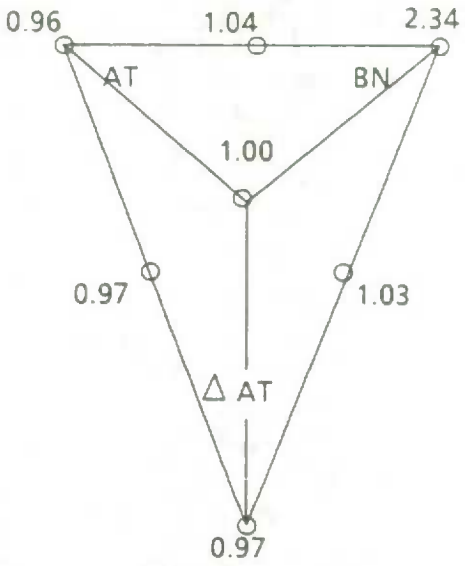


Figure 3. - - Effets du plan pour l'estimation de BN

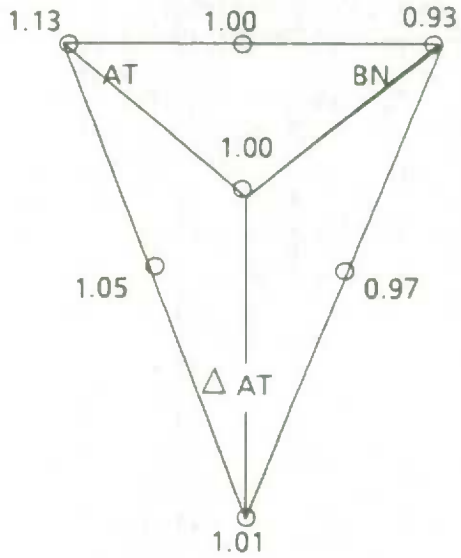
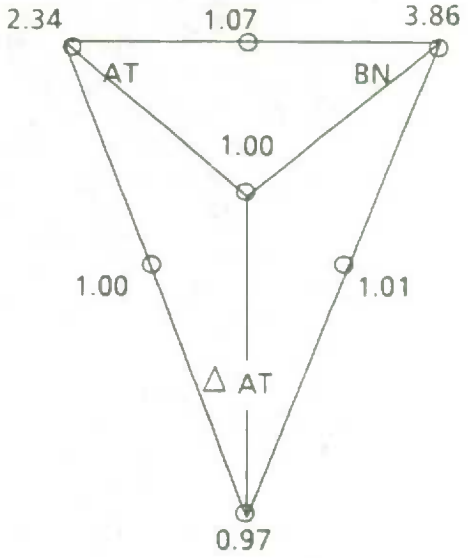


Figure 4. - - Effets du plan pour l'estimation de Δ AT



5. CONCLUSIONS ET SUJETS DE RECHERCHE

L'analyse précédente n'a porté que sur une population statique, c'est-à-dire que nous n'avons pas considéré les effets de l'inflation ou de la croissance réelle. Les limites des strates ne varient pas vraiment au cours des années puisque des estimations sont publiées annuellement pour les classes de population définies par ces strates.

Néanmoins, un taux d'inflation de 3% par exemple provoquerait des déplacements d'une strate à l'autre qui ne reflèteraient pas une variation réelle de l'actif de sociétés. On a donc envisagé de modifier les limites de strates de manière que celles-ci représentent le plus fidèlement possible la même partie de population d'année en année et que les déplacements d'une strate à l'autre soient plus représentatifs de variations réelles. Cette modification aurait aussi pour effet d'améliorer les estimations annuelles globales; en revanche, elle pourrait avoir des effets négatifs sur les estimations en dollars constants des sous-classes.

Nous venons de donner un très bref aperçu des propositions actuellement à l'étude dans le but d'améliorer le plan de sondage utilisé pour le CSIP; nous avons passé sous silence de nombreux problèmes d'application tels que les effets des erreurs non dues au sondage, le contrôle des tailles d'échantillon, etc. Des modifications ont été apportées au plan de sondage du CSIP dans le but d'améliorer l'estimation des variations d'une année à l'autre et nous croyons qu'il est possible d'apporter d'autres améliorations sans que cela n'ait d'effet négatif sur les estimations transversales. Nous sommes en train d'approfondir la question et souhaitons pouvoir donner un compte rendu de nos recherches à la prochaine assemblée (1988) de l'American Statistical Association qui se tiendra à la Nouvelle-Orléans.

BIBLIOGRAPHIE

Cochran, W.G. (1977). *Sampling Techniques*, (3è éd.) New York, John Wiley, 85.

Harte, J.M. (1982). Post-Stratification Approaches in the Corporation Statistics of Income Program, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 250-253.

Harte, J.M. (1986). Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 603-608.

Internal Revenue Service (1987). Description of the Sample and Limitations of the Data, *Statistics of Income - 1984 Corporation Income Tax Returns*, publ. no 16, Washington, D.C., 7-14.

Jones, H.W., et McMahon, P. (1984). Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 437-442.

Oh, H.L., et Scheuren, F.J. (1987). Variantes de la méthode itérative du quotient. *Techniques d'enquête*, vol. 13, no. 2, Statistique Canada.

Sunter, A.B. (1986). Implicit Longitudinal Files: A Useful Technique, *Journal of Official Statistics*, vol. 2, no. 2, Statistics Sweden, 161-168. Voir particulièrement 164.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

UTILISATION DE DONNÉES ADMINISTRATIVES POUR L'ÉTABLISSEMENT DES PROFILS INITIAUX ET ULTÉRIEURS DES ENTITÉS ÉCONOMIQUES

COLLEEN CLARK et ROBERT LUSSIER[1]

RÉSUMÉ

Statistique Canada s'emploie actuellement à remanier son registre central des entités économiques. Dans le nouveau registre, chaque entité économique est considérée comme un réseau d'entités juridiques et exploitantes qui définissent des entités statistiques. On obtient l'image de ce réseau, c'est-à-dire le profil par le processus dit d'"établissement des profils", qui suppose des contacts avec l'entité économique. En 1986, on s'est servi d'une liste de toutes les entités avec lesquelles il fallait entrer en contact afin d'obtenir les profils permettant de constituer le nouveau registre. Pour dresser cette liste, on a eu recours à des données administratives. À l'avenir, les données administratives serviront de source de renseignements sur les changements qui auront pu se produire dans les entités économiques. Elles pourront donc être utilisées pour demander qu'on révise et mette à jour les profils.

L'article porte d'abord sur les objectifs du processus d'établissement des profils. On présente ensuite les procédures de construction de la base de sondage servant au processus d'établissement des profils initiaux à l'aide de plusieurs sources de données administratives. Ces procédures comprennent l'application de concepts, la détection des cas de chevauchement entre les sources et l'évaluation de la qualité des données. On examine ensuite le rôle des données administratives comme source de renseignements sur les changements qui peuvent s'être produits dans les entités économiques et comme source de données sur lesquelles on peut s'appuyer pour demander de vérifier les profils. Suit une analyse des résultats d'une étude de simulation visant à évaluer ce rôle. L'exposé s'achève par une série de questions sur la méthodologie relative à l'utilisation de données administratives en vue de l'établissement de profils de mise à jour.

1. INTRODUCTION

Statistique Canada s'emploie actuellement à réorganiser son programme d'enquêtes économiques. Dans le nouveau programme, on utilisera davantage les données administratives. Ces dernières feront partie intégrante d'une Base de données du registre central (BDRC) d'où les enquêtes économiques tireront leur échantillon.

[1] Colleen Clark, DMES, Statistique Canada, 4-C1, Immeuble Jean Talon, Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6; Robert Lussier, DMEE, 11-M, Immeuble R.H. Coats, Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6.

Les données administratives serviront aussi à mettre à jour la BDRC. Cet aspect et d'autres aspects de la stratégie de réorganisation sont présentés dans Colledge et Lussier (1985). Certains résultats de la mise en pratique de la stratégie sont contenus dans Colledge (1987).

Une des premières étapes de la réorganisation a été de définir les unités de la BDRC. L'unité fondamentale est l'entité commerciale.[2] Statistique Canada (voir Statistique Canada, 1987) définit une entité commerciale comme "un agent économique ayant la responsabilité et le pouvoir d'affecter des ressources à la production de biens et/ou de service et, de ce fait, de gérer et de contrôler la façon de recevoir et d'employer les recettes, d'accumuler des avoirs, d'emprunter et de prêter des capitaux et de tenir à jour des états financiers complets de ses activités" (traduction).

La Base de données du registre central actuellement élaborée par Statistique Canada est une entreprise de représentation de la structure de l'économie canadienne. Elle tient compte du fait que cette économie est dominée par un petit nombre de grandes entités commerciales qui représentent le gros de l'activité dans cette économie. La BDRC est divisée en deux parties pour traduire cette dichotomie.

Une des composantes, la partie intégrée (PI), couvre le petit nombre d'entités commerciales importantes par la taille ou selon d'autres critères, tandis que l'autre composante, la partie non intégrée (PNI), couvre les autres entités commerciales, c'est-à-dire le grand nombre de petites entités. Les entités de la première composante sont plus complexes. Aussi faut-il un certain effort pour déterminer, dans une entité commerciale complexe, quels éléments peuvent être intéressants pour une enquête donnée.

La partie intégrée (PI) de la BDRC est un moyen de représentation de la structure complexe des entités commerciales à l'aide d'un modèle d'information. Le modèle se compose de cinq structures liées entre elles qui décrivent une entité commerciale. Ces structures permettent de définir avec précision les populations d'enquête. Les entités de trois de ces structures ne sont pas controlées par Statistique Canada, tandis que les deux autres sont produites par le Bureau à des fins de collecte, de vérification, d'estimation et de totalisation de données économiques. Les cinq structures sont les suivantes.

i. La **structure juridique,** qui décrit comment l'entité commerciale est organisée juridiquement. Elle représente les entités juridiques et les relations de propriété et de contrôle qu'elles ont entre elles. Entre autres exemples d'entités juridiques, notons les entreprises constituées en société en vertu de chartes fédérales ou provinciales.

ii. La **structure opérationnelle,** qui décrit comment l'entité commerciale fonctionne et comment elle organise son système comptable. Elle se compose des entités exploitantes. C'est elle qui organise et contrôle la production des biens et/ou des services. Il s'agit d'un moyen de structurer l'entité commerciale de la façon dont cette dernière se perçoit. Entre autres exemples d'entités exploitantes, notons les divisions, les centres de profit et les usines.

[2] L'expression "entité commerciale" correspond ici à "business entity". Le lecteur doit savoir qu'aux États-Unis l'emploi du terme "business" pour désigner une entreprise est souvent réservé au secteur commercial. Ce n'est pas le cas au Canada, où ce terme sert pour tous les secteurs de l'économie, y compris le secteur manufacturier, les transports et les professions. Le terme "économique" a été utilisé dans le titre pour éviter toute confusion possible avec le sens du terme américain "business", mais c'est le terme "commercial" qui est toujours utilisé dans la suite du texte.

iii. La **structure statistique,** qui représente les entités statistiques classées selon un ordre hiérarchique. Les entités statistiques sont constituées à partir de la structure opérationnelle correspondante suivant les unités de la structure opérationnelle pour lesquelles un ensemble particulier de données est tenu à jour.

iv. La **structure déclarante,** qui représente les modalités de déclaration définies pour chacune des entités statistiques choisies, par enquête. Les données du système comptable de l'entité commerciale sont communiquées par les entités déclarantes.

v. La **structure administrative,** qui contient des données administratives comme les données fiscales recueillies auprès des entités juridiques et les données des comptes de retenue sur la paye recueillies auprès des entités exploitantes.

Le processus complexe de délimitation des frontières de chacune des entités commerciales et de détermination de ses cinq structures et des relations existant entre ces structures est appelé "établissement de profils". Cette représentation de l'entité commerciale comme réseau est le "profil". Les données servant à construire un profil sont obtenues par contact avec l'entité commerciale ou avec une de ses composantes. Les éléments d'information sur la structure juridique et sur la structure opérationnelle des entités commerciales ainsi que certains éléments d'information sur leur structure administrative sont obtenues ou révisées et mis à jour au cours de l'interview. La structure statistique est ensuite produite ou mise à jour automatiquement à partir de la nouvelle structure opérationnelle. Finalement, des entités déclarantes implicites sont créées pour chaque entité statistique nouvellement choisie à l'aide des données de certaines zones tirées des structures juridique, opérationnelle ou administrative. Ces entités peuvent par la suite être mises à jour au moyen des renseignements obtenus à l'occasion du premier contact avec les répondants.

Le type de contact qui est pris en vue de l'établissement des profils dépend de la complexité des entités et de toute modalité de déclaration spéciale. Pour ce qui est des données relatives aux entités les plus complexes et les plus importantes, elles seront recueillies par le personnel du bureau central ou d'un bureau régional par interview sur place. Les données relatives aux autres entités seront recueillies par interview téléphonique. Les entités seront contactées une fois tous les deux ans ou plus souvent, selon la rapidité avec laquelle leurs structures changent.

La méthode d'établissement cyclique des profils, par laquelle les entités commerciales sont contactées périodiquement, est une des méthodes qui seront utilisées pour mettre à jour la PI de la BDRC. On utilisera également des renseignements tirés d'enquêtes et des données de sources administratives.

On a prévu pour la conception et la construction de la BDRC une période de trois ans qui devrait aboutir à la création d'une base de données à intégrer dans les programmes des enquêtes en avril 1988. En avril 1988, la plupart des données relatives à la partie intégrée de la BDRC auront été obtenues au moyen du processus d'établissement des profils qui a commencé en avril 1986. Mais à ce moment-là, il n'y avait pas de liste unique d'entités commerciales dont on pouvait établir le profil.

Les données administratives ont joué un rôle important dans l'amorce du processus d'établissement des profils. Statistique Canada s'en est servi comme point de départ pour construire son registre des entités commerciales. Une liste des entités commerciales pouvant faire l'objet d'un profil initial a été dressée à partir de sources de données administratives. La partie 2 du présent document décrit comment cela s'est passé. La partie 2.1 présente les éléments nécessaires à la création de la base de sondage. Une description des sources de données utilisées pour construire la base suit dans la partie 2.2, tandis que la partie 2.3 montre comment on a déterminé l'unité de sondage et comment on a combiné les diverses sources de données pour construire la base de sondage.

La partie 3 décrit comment les données administratives seront utilisées pour détecter les changements qui se seraient produits dans une entité commerciale et pour lancer ensuite le processus de mise à jour des profils. Sont ensuites présentés les résultats d'une étude de simulation effectuée pour quantifier le degré d'utilisation proposé des sources de données administratives. Enfin, le document se termine par une analyse de certains points soulevés dans cette étude.

2. UTILISATION DE DONNÉES ADMINISTRATIVES POUR L'ÉTABLISSEMENT DE PROFILS INITIAUX

2.1 Éléments nécessaires à la création de la base de sondage

La première étape de la construction de la base de sondage devant servir à l'établissement des profils initiaux a été de définir l'unité de sondage. L'unité idéale aurait été l'entité commerciale. Toutefois, cette entité ne pouvait être obtenue à partir de sources ni intérieures ni extérieures à Statistique Canada. Les seules entités qu'il nous était possible d'obtenir étaient essentiellement des entités juridiques. Il a donc fallu regrouper les entités juridiques pour reproduire approximativement des entités commerciales. L'unité de sondage a été définie comme un agrégat d'entités juridiques assujetties aux contraintes suivantes.

i. La définition de l'entité commerciale suppose que cette dernière englobe toutes les entités juridiques liées entre elles par des liens de contrôle. Une façon de contrôler une entité juridique est de posséder plus de 50% de ses actions avec droit de vote. Le groupement d'entités juridiques par cette forme de contrôle est limité à un seul niveau de contrôle étranger à l'extérieur du Canada.

ii. Il faut qu'une seule entité juridique canadienne contrôle toutes les autres entités juridiques canadiennes faisant partie de l'entité commerciale. Cette condition est indispensable parce que les contacts en vue de l'établissement des profils avec l'entité commerciale pouvaient être pris au Canada seulement.

L'étape suivante a été de déterminer quelles unités de sondage composeraient la base de sondage et quelles données il fallait pour chacune. La base dont les entités commerciales seraient tirées pour un premier contact en vue de l'établissement d'un profil initial et à partir de laquelle on pourrait produire une première image de l'entité commerciale devait contenir toutes les entités commerciales pouvant être contactées.

Les entités peuvent être contactées en vue de l'établissement d'un profil si elles remplissent les conditions requises pour faire partie de la partie intégrée de la BDRC. Les critères déterminant si une entité commerciale doit ou non être incluse dans la partie intégrée de la BDRC sont appliqués à la structure juridique, qui décrit comment l'entité commerciale est organisée juridiquement.

Les entités juridiques peuvent faire partie de la partie intégrée de deux façons. Premièrement, si l'entité commerciale se compose d'une seule entité juridique, celle-ci entrera dans la partie intégrée si son revenu au cours de l'exercice financier considéré est supérieur à une valeur prédéterminée. Cette limite dépend de l'activité principale de l'entité juridique et de l'endroit où se trouve son siège social. Deuxièmement, si la structure juridique comprend plus d'une entité juridique, les entités juridiques feront toutes partie globalement de la partie intégrée si au moins une d'entre elles a un revenu supérieur à la limite établie pour ce genre d'entité.

Aussi, pour déterminer quelles entités commerciales pouvaient être contactées, il a fallu recueillir les renseignements suivants pour chaque entité juridique.

i. Les relations de propriété entre les entités juridiques.

ii. Le revenu de l'exercice financier considéré, l'activité principale et l'endroit où se trouve le siège social.

Pour ce qui est des entités commerciales qui remplissaient les conditions requises pour faire partie de la base de sondage et donc pour être contactées en vue de l'établissement d'un profil initial, il fallait des renseignements pour pouvoir les choisir et les contacter. Pour les choisir, il fallait les renseignements suivants:

i. La liste de toutes les industries dans lesquelles l'entité commerciale avait des activités, pour pouvoir contacter d'abord les entités du secteur du commerce de gros et/ou du commerce de détail. Les enquêtes auprès de ces entités ont nécessité, avant les autres enquêtes, un ensemble d'entités statistiques produit à partir de contacts en vue de l'établissement de profils.

ii. Le nombre de locaux d'affaires de toutes les entités commerciales comprenant une seule entité juridique ou deux entités juridiques si le propriétaire est un étranger. Cet élément d'information a déterminé le type de contact à prendre en vue de l'établissement de profils, qui pouvait être une interview téléphonique réalisée par le personnel d'un bureau régional ou une interview sur place réalisée par le personnel du bureau central ou d'un bureau régional.

iii. Le province de résidence du principal propriétaire canadien. La province a été utilisée pour répartir entre les bureaux régionaux selon leur capacité la charge de travail en termes des contacts à prendre en vue de l'établissement des profils.

Pour contacter les entités commerciales, il fallait le nom et l'adresse de l'entité juridique à la tête (excluant les propriétaires étrangers) de l'entité commerciale. Il était souhaitable de disposer en plus des données de contact et des données sur toute modalité de déclaration spéciale ayant été utilisées dans des enquêtes récentes.

2.2 Sources de données

Les sources de données qui pouvaient être utilisées étaient limitées principalement par le champ des unités que devait couvrir la base de sondage. Cette condition éliminait les listes d'éléments d'échantillons et beaucoup de listes propres à des industries précises comme des bases de sondage. Les seules sources de données qui pouvaient être considérées étaient celles qui représentaient des listes de toutes les entités juridiques susceptibles d'entrer dans le champ des unités qui pouvaient être contactées en vue de l'établissement d'un profil et qui pouvaient produire au moins en partie certains des éléments d'information requis. Les sources de données retenues ont été les suivantes:

i. La **Base de données sur les liens de parenté entre firmes** (LPF), qui est une liste de toutes les entités juridiques ayant des activités au Canada et appartenant à une entité juridique étrangère ou canadienne, et ses propriétaires. Les entités juridiques étrangères sont couvertes jusqu'au niveau qu'il faut pour remonter jusqu'au propriétaire qui en dernière analyse possède le contrôle.

ii. Le **Registre des entreprises** (RE) actuel, qui est principalement une liste de toutes les entités juridiques qui sont des employeurs. C'est dans ce registre qu'on peut trouver le nombre de locaux d'affaires d'une entité juridique, les données de contact (adresse et modalités de déclaration) utilisées pour les enquêtes et les industries dans lesquelles l'entité juridique est active.

iii. La **Base des unités de l'impôt sur le revenu des sociétés** (IRS), qui est une liste de toutes les entités juridiques qui ont produit une déclaration d'impôt sur le revenu

des sociétés auprès de Revenu Canada, Impôt au cours d'une année donnée. L'activité principale, l'endroit où se trouve le siège social et le revenu de l'exercice financier sont indiqués dans cette source de données.

iv. La **Base des unités de l'impôt sur le revenu des particuliers** (IRP), qui est une liste de tous les particuliers qui ont produit une déclaration d'impôt sur le revenu auprès de Revenu Canada-Impôt au cours d'une année donnée. Les particuliers qui déclarent un revenu provenant d'un travail autonome sur leur déclaration sont des entités juridiques au sens des enquêtes économiques effectuées par Statistique Canada. On peut obtenir de cette base des données sur l'activité principale et des données de contact pour chaque particulier qui déclare un revenu provenant d'un travail autonome.

Les deux sources de données de l'impôt sur le revenu (IRS et IRP) sont des fichiers de données administratives. Les données administratives reçues tous les mois de Revenu Canada-Impôt concernant les retenues sur la paye d'un employeur sont utilisées pour mettre à jour le RE. La source de données LPF est un fichier de réponses à une enquête par recensement.

Aucune de ces sources de données ne couvre complètement l'univers des unités ni ne fournit tous les éléments d'information requis. La seule façon de couvrir tout l'univers est de les combiner. Cela vaut aussi pour certains éléments d'information requis tandis que pour les autres éléments, plus d'une source peuvent les fournir. La stratégie utilisée pour combiner ces sources de données afin d'obtenir le meilleur taux de couverture possible et la meilleure qualité des données possible est présentée dans la prochaine partie.

2.3 Procédures de création de la base de sondage

Le problème que pose la création de la base de sondage servant aux contacts à prendre en vue de l'établissement des profils initiaux est d'intégrer quatre sources de données qui n'ont pas été conçues pour les mêmes objectifs et qui n'ont jamais été intégrées dans cette mesure auparavant. Cette difficulté est commune à tous les utilisateurs de données administratives. La tâche était encore compliquée par le fait que c'était la première fois qu'on appliquait en même temps autant de concepts établis pour la BDRC.

Des contraintes de temps et de ressources ont obligé l'équipe chargée du projet à faire certaines hypothèses au moment de créer la base de sondage. Les hypothèses formulées étaient toutefois justifiables puisqu'on pourrait par la suite corriger la base de sondage à l'aide des données obtenues du processus d'établissement des profils. La façon dont on a procédé pour créer la base de sondage est décrite en termes simples dans les paragraphes qui suivent.

La création de la base de sondage comportait trois étapes que nous allons décrire plus en détail un peu plus loin.

i. Construire une liste de toutes les unités pouvant éventuellement faire partie de la base de sondage.

ii. Déterminer quelles unités remplissaient les conditions requises pour faire partie de la base de sondage.

iii. Obtenir des données de sélection et de contact.

2.3.1 Création des unités susceptibles de faire partie de la base de sondage

Les unités de sondage ont été construites en groupant les entités juridiques pour créer des entités commerciales de la façon suivante. Les entités juridiques ont d'abord été

groupées en structures juridiques. Une structure juridique se composait de l'ensemble des entités juridiques liées entre elles par la propriété de plus de 50 % des titres de propriété. Les relations comportant des entités juridiques étrangères étaient acceptées seulement si l'entité juridique étrangère possédait une entité juridique canadienne ou appartenait à une entité juridique canadienne. Dans les cas où une entité étrangère possédait plus d'une entité canadienne, la structure juridique a été divisée en autant d'entités commerciales qu'il y avait d'entités canadiennes appartenant directement à l'entité étrangère. Ainsi, un contact en vue de l'établissement d'un profil pouvait être pris avec le propriétaire canadien qui possède le contrôle de chaque entité commerciale résultante. Des exemples de cette façon de procéder sont donnés dans le diagramme 1 à la fin de la partie 2.3.1.

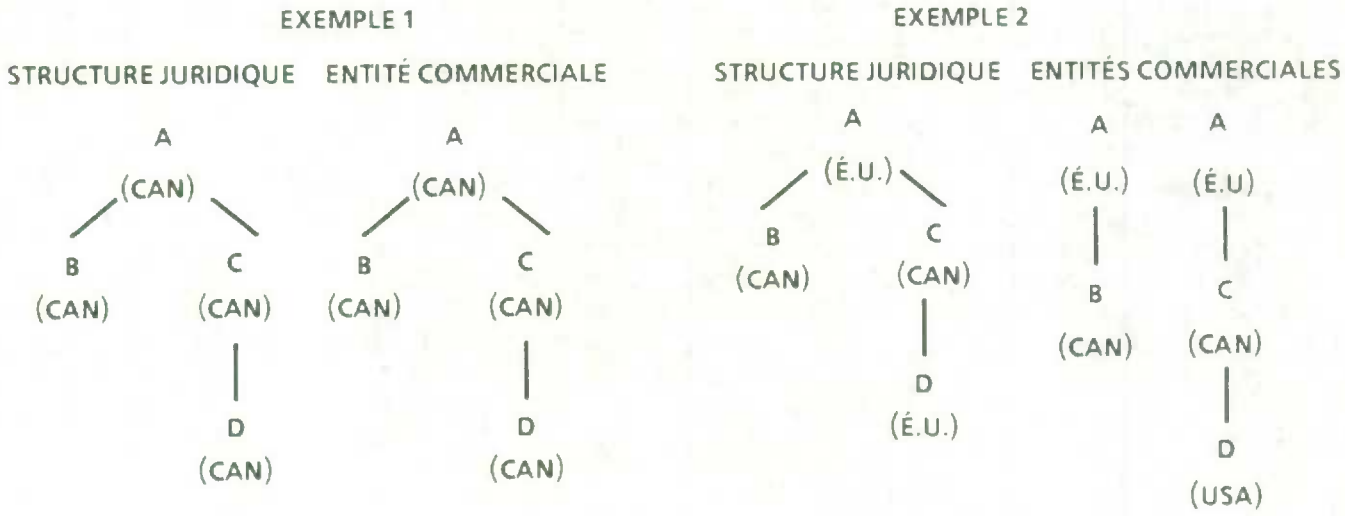
Les particuliers ayant déclaré un revenu provenant d'un travail autonome étaient considérés comme une structure juridique contenant une seule entité juridique. Pour la construction des entités commerciales, on n'a pas tenu compte du fait que des sociétés puissent appartenir à des particuliers, ni des relations de coentreprise entre des sociétés en coparticipation.

Par conséquent, on peut considérer que l'ensemble des entités commerciales faisant partie du champ des unités susceptibles d'être contactées en vue de l'établissement d'un profil initial est formé de deux sous-ensembles mutuellement exclusifs. Le premier groupe est formé des entités juridiques qui représentent des particuliers ayant déclaré un revenu provenant d'un travail autonome. La base des unités de l'impôt sur le revenu des particuliers (IRP) contient une liste de toutes les unités de ce groupe pouvant éventuellement faire partie de la base de sondage.

Le deuxième groupe se compose des entités juridiques qui représentent les sociétés ayant des activités au Canada. La source de données sur les liens de parenté entre firmes (LPF) a été traitée de manière à fournir une liste des sociétés appartenant aux structures juridiques contenant plus d'une entité juridique. On a dressé une liste de toutes les entités juridiques n'appartenant pas à une autre entité juridique en éliminant de la base des unités de l'impôt sur le revenu des sociétés les entités juridiques qui appartenaient à une autre entité juridique ou qui étaient elles-mêmes propriétaires d'autres entités. Pour cela, il a fallu comparer les données de la base de données LPF aux données de la base IRS pour déterminer dans quelle mesure les deux bases se recoupaient. De cette façon, on a pu déterminer quelles entités juridiques figuraient dans les deux bases pour faire en sorte qu'elles n'apparaissent qu'une fois dans la base de sondage. Le couplage des entités des deux sources n'a pas été facile à effectuer et comportait une partie manuelle parce que souvent il n'existait pas de numéro d'identification commun aux enregistrements des deux sources.

DIAGRAMME 1

DÉFINITION DES ENTITÉS COMERCIALES



2.3.2 Détermination des unités susceptibles de faire partie de la base de sondage

Les données dont on avait besoin pour déterminer si les particuliers ayant déclaré un revenu provenant d'un travail autonome pouvaient faire partie de la base de sondage figuraient dans la base IRP. Il a été très simple de déterminer si le revenu d'une entité juridique était supérieur ou non au seuil fixé au préalable pour cette entité.

La situation était plus compliquée pour les sociétés. Le couplage des données de la base de données LPF et des données de la base IRS nous a fourni les données dont nous avions besoin pour appliquer la règle du seuil limite. Toutefois, environ 20% des sociétés figurant dans la base de données LPF n'ont pu être liées à des sociétés de la base IRS. Dans ce cas, il a fallu faire une hypothèse qui a entraîné une surestimation du nombre d'entités commerciales pouvant faire partie du champ des unités susceptibles d'être contactées en vue de l'établissement d'un profil initial. L'hypothèse était que les structures juridiques qui contenaient au moins une société non appariée satisfaisaient aux conditions d'inclusion dans la base de sondage. Sinon, les structures juridiques et, par conséquent, les entités juridiques qu'elles contenaient faisaient automatiquement partie de la base de sondage si au moins une de leurs sociétés satisfaisait à la règle du seuil limite.

2.3.3 Acquisition de données de sélection et de données de contact

L'étape précédente a abouti à une liste par approximation de toutes les entités commerciales pouvant éventuellement faire partie du champ des unités susceptibles d'être contactées en vue de l'établissement d'un profil initial. Les données de sélection et de contact décrites dans la partie 2.1 qui n'étaient pas déjà dans la base de sondage étaient disponibles du RE. La base de sondage et le RE se recoupent en partie parce que la majorité des unités de la partie de la base de sondage représentant les sociétés et une plus faible proportion des unités de la partie de la base de sondage représentant les particuliers étaient des employeurs. Il a fallu apparier les données de la base de sondage et les données du RE pour pouvoir ajouter à la base de sondage les données tirées du RE quand on constatait que des unités figuraient dans les deux sources. Autrement dit, il a fallu trouver quelles unités figuraient dans les deux sources.

Il a été encore plus difficile d'apparier ces deux sources de données que les données de la base de données LPF et les données de la base IRS. Cela était attribuable non seulement au fait qu'on ne retrouvait souvent pas de numéros d'identification communs dans les deux sources, comme dans le cas LPF-IRS, mais aussi au fait que le RE ressemble plus à la structure opérationnelle des entités commerciales qu'à leur structure juridique. Le nom et l'adresse tirés du RE ont été utilisés pour apparier les enregistrements quand il n'y a avait pas de numéro d'identification commun. Toutefois, les noms et les adresses figurant dans le RE s'appliquent souvent aux locaux "commerciaux" ou "exploitants" qui ont parfois des noms et des adresses différents des noms et des adresses "juridiques" figurant dans la base de données LPF et la base des unités IRS. Quand cela se produit, il est difficile d'établir un lien et donc d'éliminer les doubles comptes.

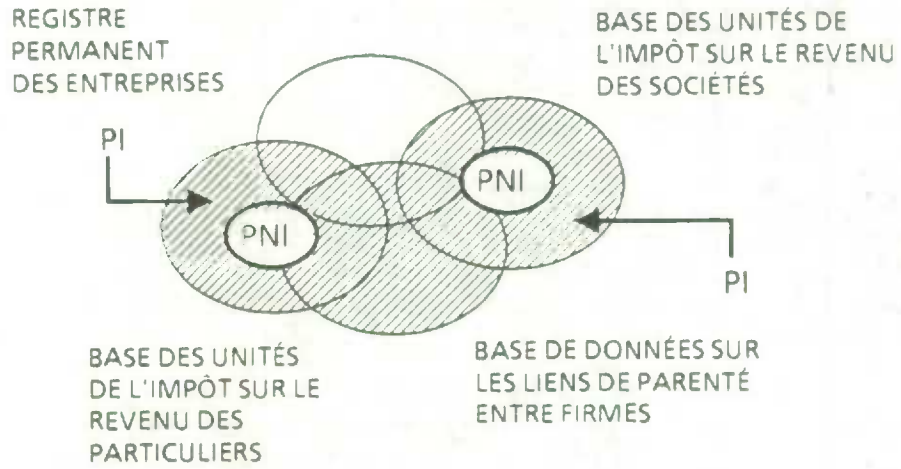
La base de sondage contenait certaines unités pour lesquelles on n'a pas établi de lien avec le RE soit parce que ces unités n'étaient pas des employeurs, et donc ne figuraient pas dans le RE, soit parce que les procédures de couplage n'ont pas permis d'établir le lien. Pour ces cas-là, les étapes ultérieures du processus d'établissement des profils initiaux ont été modifiées de manière à tenir compte des conditions imposées par la base de sondage. Des données de contact moins bonnes ont été tirées de la base des unités de l'impôt. Les critères de sélection ont également été modifiés de manière à tenir compte du fait qu'il puisse ne pas y avoir de données sur la répartition industrielle et les locaux d'affaires de ces entités juridiques.

Quand une entité juridique était active dans plus d'une industrie, l'activité principale pouvait être déterminée à partir des bases des unités de l'impôt sur le revenu et du RE. Il a fallu alors concilier les éléments d'information communs quand ils différaient. Dans ce cas, l'industrie indiquée dans le RE a été utilisée étant donné que l'on jugeait cette source de renseignements plus fiable.

Le diagramme 2 illustre par une figure (qui n'est pas à l'échelle) la base de sondage ainsi créée.

DIAGRAMME 2

BASE DE SONDAGE EN VUE DE L'ÉTABLISSEMENT DE PROFILS INITIAUX



2.3.4 Évaluation de la base de sondage

La qualité de la base de sondage obtenue a été évaluée de trois façons. On a d'abord vérifié que la base de sondage concordait avec les spécifications développées pour sa création.

Deuxièmement, on a comparé diverses distributions des entités juridiques figurant dans la base de sondage avec les mêmes distributions produites de façon indépendante à partir d'une simulation de la partie intégrée. Les distributions ne différaient pas de façon significative.

Enfin, on a évalué la base de sondage en la comparant au RE. Un échantillon de 30 des plus grandes unités du RE a été apparié à la base de sondage créée en vue de l'établissement des profils initiaux. On a retrouvé toutes les unités, quoique très difficilement, parce que les deux sources n'utilisent pas les mêmes concepts.

2.4 Conclusion

La stratégie élaborée pour créer la base de sondage qu'on vient de décrire reposait sur des hypothèses simples au sujet de la couverture, de la qualité des données et de la façon dont fonctionnent les entités commerciales. On a souvent utilisé des "raccourcis" pour remplir les conditions requises par la base de sondage. On a estimé que cette approche était parfaitement justifiée étant donné qu'on jugeait que le rôle de la base de sondage était de donner des entités commerciales une première image qui pouvait ensuite être

mise à jour pendant le processus d'établissement des profils. Les conséquences de ces hypothèses sont analysées dans les paragraphes qui suivent.

La population des entités commerciales pouvant faire partie du champ des entités susceptibles d'être contactées en vue de l'établissement d'un profil initial peut contenir des unités comptées deux fois et des unités hors du champ. Si c'est le cas, on prendra plus de contacts qu'il n'est nécessaire. Cela augmentera les coûts de production de Statistique Canada et accroîtra indûment le fardeau de réponse de certains répondants en les obligeant à répondre deux fois aux mêmes questions. Enfin, l'image de Statistique Canada pourrait être ternie un peu.

La population pourrait être sous-estimée. Dans ce cas-là, on pourra établir le profil des unités manquantes un peu plus tard. Cela risque de retarder l'introduction de nouvelles grandes unités dans la partie intégrée de la BDRC. Entre-temps, les unités manquantes seraient couvertes par la partie non intégrée et non par la partie intégrée.

L'imprécision des données de sélection et/ou de contact pourrait compliquer ou retarder les contacts jusqu'à ce qu'on obtienne des données précises. Dans ces cas-là, il en résulte également que la BDRC sera imprécise jusqu'à ce que tous les profils aient été achevés.

Ces divers résultats montrent les complications que peut entraîner l'utilisation de données administratives. Ils montrent également qu'il faut bien s'assurer que les données administratives utilisées concordent avec les objectifs fixés. Des exemples ont été donnés des types de compromis qu'il faut faire quand il est impossible d'obtenir une compatibilité raisonnable.

3. UTILISATION DE DONNÉES ADMINISTRATIVES POUR L'ÉTABLISSEMENT DES PROFILS ULTÉRIEURS

3.1 Établissement cyclique et établissement ponctuel de profils

Il y aura deux types de mise à jour de profils initiaux, à savoir l'établissement cyclique de profils et l'établissement ponctuel de profils. Chaque type est expliqué ci-dessous.

L'établissement cyclique de profils est le processus par lequel on s'assure que le profil de toutes les entités commerciales de la population est refait au bout d'un certain temps. Compte tenu des prévisions budgétaires actuelles, on prévoit que cette période sera de deux ans. Le temps écoulé depuis le dernier profil sera le premier critère utilisé pour déterminer si cette entité doit être soumise au processus d'établissement cyclique des profils. On tiendra compte d'autres facteurs pour classer par ordre de priorité les entités dont il faut refaire le profil.

L'établissement ponctuel de profils est le processus d'établissement de profils des entités commerciales qui est déclenché parce qu'on a reçu d'une autre source des renseignements selon lesquels des changements se seraient produits dans l'entité commerciale et que l'image statistique de l'entité commerciale dans le registre peut n'être plus valide. Le processus d'établissement ponctuel de profils fera que la BDRC sera plus à jour que si l'on utilisait seulement le mécanisme d'établissement cyclique des profils. Les divers fichiers de données administratives reçus régulièrement à Statistique Canada font partie des autres sources d'information sur les changements.

3.2 Sources de données administratives pouvant être utilisées

Les trois sources de données administratives que Statistique Canada peut utiliser pour mettre à jour son registre central et dont nous parlons dans le présent document sont:

- la Base des unités de l'impôt sur le revenu des particuliers,
- la Base des unités de l'impôt sur le revenu des sociétés, et
- les données sur les comptes de retenue sur la paye saisies par les administrations fiscales.

En général, les particuliers et les sociétés produisent une seule déclaration d'impôt pour une année de référence donnée. Il se peut toutefois qu'ils en produisent plus d'une parce que, par exemple, une société aurait changé la fin de son exercice financier avec l'approbation des administrations fiscales. On peut quand même dire que les déclarations d'impôt sur le revenu sont une source de données annuelles sur les changements.

Ce n'est pas seulement une fois par année que les bases des unités de l'impôt sur le revenu sont communiquées à Statistique Canada. En fait, Statistique Canada reçoit régulièrement pendant deux ans les fichiers de données fiscales se rapportant à une année de référence donnée. Par conséquent, on pourrait mettre à jour tous les mois le registre à partir des données fiscales, mais chaque enregistrement du registre serait généralement mis à jour une fois par année seulement.

Par ailleurs, on s'attend qu'en règle générale les employeurs s'acquittent de leurs comptes de retenue sur la paye une fois par mois. Ainsi, Statistique Canada reçoit un fichier de données des comptes de retenue sur la paye une fois par mois. Par conséquent, on peut mettre à jour le registre tous les mois à partir des données des comptes de retenue sur la paye, et chaque enregistrement du registre peut en théorie être modifié tous les mois.

Il convient de noter que d'autres sources de données administratives peuvent aussi être utilisées. On n'en parle pas dans le présent document parce qu'elles ne couvrent pas toute la population, ni sur une base régulière. Il convient toutefois de les citer. Ce sont:

- l'information limitée recueillie par les administrations fiscales sur les sociétés qui n'ont pas produit de déclaration d'impôt mais dont on pense qu'elles sont actives,
- d'autres données recueillies d'un échantillon de déclarations d'impôt sur le revenu par Statistique Canada, et
- les données recueillies par Statistique Canada à même les formules fiscales remplies par les employeurs demandant qu'on leur ouvre un compte de retenue sur la paye.

3.3 Signaux de changement

On a mis au point un système de signaux de changement qui proviendraient de chacune des sources de données administratives décrites dans la partie précédente. Ces signaux, rapportés à des enregistrements administratifs, indiquent quelles entités statistiques ont pu subir des changements. Ils indiquent également aux responsables du registre qu'il peut-être souhaitable de refaire le profil de ces entités statistiques à l'aide du processus d'établissement ponctuel de profils pour tenir le registre à jour.

Les signaux dépendent de la source administrative. Dans chacune des trois sources énumérées dans la partie 3.2, les signaux sont des tests de comparaison entre les nouvelles données reçues pour un enregistrement donné et les données précédentes reçues de la même source pour l'enregistrement en question. Les tests peuvent porter sur une seule zone ou sur un groupe de zones et être conditionnel à une ou plus d'une zone. Ces tests de comparaison tentent de refléter les événements du monde réel qui influent sur les entités statistiques et non pas seulement ceux qui influent sur les entités administratives. Il faut se rappeler que les entités statistiques ont été créées aux fins des programmes de statistiques économiques et souvent diffèrent complètement des entités juridiques ou administratives. Aussi, ces tests de comparaison devraient permettre le plus possible de trouver les changements survenus dans les données administratives qui reflètent un

changement dans les entités statistiques. Par exemple, un changement de propriétaire d'une usine de fabrication peut signifier qu'un enregistrement administratif disparaîtra et qu'un autre apparaîtra. Cela peut par contre n'avoir aucun effet sur les entités statistiques étant donné que le même établissement avec le même degré de capacité de fournir les données requises peut continuer d'exister.

Si la base de sondage était mise à jour directement à partir des changements observés dans les enregistrements administratifs sans aucun autre contrôle, il y aurait une forte proportion de créations et de disparitions apparentes d'entités et un risque proportionnel de couverture incomplète ou de couverture en double. C'est pourquoi il faut aussi contacter les répondants ou au moins faire des recherches internes à l'aide de tous les documents qu'on peut obtenir pour trouver ce qui est arrivé aux entités statistiques dans le cas des enregistrements administratifs signalés. Le processus de "traduction" n'est pas ce qu'il y a de plus simple et la recherche d'une solution constitue justement l'objectif du processus d'établissement ponctuel de profils.

Le nombre de signaux retenus pour chaque source et quelques exemples de signaux sont présentés dans le Tableau 1. Il faut toutefois noter les points suivants quand on regarde les données sur le nombre de signaux. Certains signaux sont très raffinés, d'autres pas. Dans bien des cas, on a décidé de diviser le signal initial en deux sous-signaux mutuellement exclusifs parce qu'on estimait que cela pouvait contribuer davantage à déterminer les bonnes mesures à prendre à partir du signal. L'exemple le plus simple concerne les comptes de retenue sur la paye. Dix-huit des quarante signaux représentent des changements dans le nombre estimatif d'employés couverts par le compte. Les 18 signaux permettent de distinguer entre des augmentations et des diminutions du nombre estimatif et tiennent compte de l'ampleur des augmentations et des diminutions. On a estimé que cette façon de procéder aiderait à classer par ordre de priorité les travaux manuels à accomplir. On pourrait néanmoins considérer tous ces signaux comme un seul signal.

Tableau 1

Signaux par source administrative

Source administrative	Nombre de signaux distincts	Exemples
Déclarations annuelles d'impôt sur le revenu des particuliers	50	Passage d'une seule à plus d'une province d'imposition
Déclarations annuelles d'impôt sur le revenu des sociétés	49	Début d'une coentreprise
Comptes mensuels de retenue sur la paye	38	Nouveau compte avec la description du nom qui identifie une société

Même si les déclarations d'impôt sur le revenu sont dépouillées régulièrement, on s'attend qu'une déclaration donnée produise en général des signaux au maximum une fois par année de référence, tandis qu'un compte de retenue sur la paye pourrait produire un ou plus d'un signal tous les mois. Mais ce qui nous intéresse le plus, ce n'est pas le nombre de signaux que peut transmettre une source, mais le nombre d'enregistrements que ces signaux permettent de trouver. Cela donnerait une idée des ressources en personnel de

bureau à investir pour mettre à jour le registre à partir de sources administratives. On a donc fait une étude de simulation pour répondre à cette question.

3.4 Étude de simulation

L'étude de simulation a consisté à appliquer les signaux décrits précédemment aux populations suivantes:

- les déclarations d'impôt sur le revenu des particuliers pour les exercices financiers prenant fin en 1984, pour détecter les changements survenus durant ces exercices financiers;
- les déclarations d'impôt sur le revenu des sociétés pour les exercices financiers prenant fin en 1984, pour détecter les changements survenus durant ces exercices financiers;
- les comptes de retenue sur la paye du début d'octobre 1985, pour détecter les changements survenus depuis le début de septembre 1985.

Les résultats de l'étude de simulation sont présentés dans le Tableau 2. On peut faire les observations suivantes sur les résultats obtenus.

- Beaucoup de déclarations d'impôt sur le revenu produisent des signaux: seulement un huitième environ des déclarations d'impôt sur le revenu des particuliers et un cinquième des déclarations d'impôt sur le revenu des sociétés n'ont pas produit de signal.
- Durant la période d'un mois observée, 8,258 comptes de retenue sur la paye ont produit un signal pour le mois considéré. Si l'on suppose que les signaux émis par les comptes de retenue sur la paye sont distribués uniformément d'un mois à l'autre, cela ferait presque 100,000 comptes signalés par année. Il convient toutefois de remarquer qu'il est fort possible qu'un même compte soit signalé plus d'un mois et qu'on risque donc de compter au moins deux fois les signaux si on les additionne sur une période d'un an.
- Si l'on additionne tous les enregistrements signalés au cours de l'année, cela donne au total 244,269 enregistrements signalés. Il est toutefois évident qu'un même signal peut se trouver dans plus d'une source administrative. Par exemple, un changement de raison sociale d'une entreprise peut être indiqué dans la déclaration d'impôt produite par l'entreprise et dans chacun de ces deux comptes de retenue sur la paye.

Tableau 2

Résultats de l'étude de simulation

Source administrative	Population totale	Nombre d'enregistrements signalés	Nombre d'enregistrements signalés en pourcentage de la population totale
Déclarations d'impôt sur le revenu des particuliers	72,190	63,446	87.9
Déclarations d'impôt sur le revenu des sociétés	102,688	81,727	79.6
Comptes de retenue sur la paye	134,973	8,258	6.1

3.5 Questions soulevées

Les résultats de l'étude de simulation et l'analyse du rôle des signaux soulèvent un certain nombre de questions au sujet de l'établissement des profils.

Six de ces questions sont présentées ci-dessous.

Dans quelle mesure les signaux permettent de détecter les changements survenus dans les entités statistiques

Les signaux indiqueront plus ou moins fidèlement quelles entités juridiques et/ou exploitantes ont subi des changements réels qui ont un effet sur les entités statistiques. Il faudra ensuite mettre à jour le registre central pour maintenir la qualité des produits statistiques. Les signaux reflètent-ils réellement les événements du monde réel qui influent sur les entités statistiques ou est-ce qu'il y en a qui ne correspondent à aucun effet? Si certains n'ont pas de signification, cela va engendrer du travail inutile.

Une enquête à petite échelle a été menée en 1986 pour déterminer dans quelle mesure les signaux permettaient de déceler les changements survenus dans les entités statistiques. Toutefois, pour diverses raisons, les seuls signaux qu'on a pu utiliser étaient ceux de l'étude de simulation. Ces signaux ont trait aux changements survenus dans les déclarations d'impôt entre les années d'imposition 1983 et 1984. Mais le décalage entre la période de référence des signaux et la période d'enquête (1986) a posé des problèmes de mémoire aux répondants. C'est ainsi que les répondants ont inclus des événements qui ont eu lieu après la période de référence ou ont oublié des événements qui ont eu lieu pendant la période de référence. On a donc jugé que l'enquête n'était pas concluante, et aucun autre essai n'a été tenté depuis ce temps-là.

Répétition de signaux

Les signaux seront reçus à plusieurs moments et de sources différentes indépendantes les unes des autres. Les déclarations d'impôt sur le revenu, en particulier, entraînent des délais sensibles. Au moment où un signal est reçu, la BDRC pourra déjà avoir été mise à jour pour tenir compte de l'événement du monde réel indiqué par ce signal. Cette mise à jour pourra avoir été faite à l'occasion du traitement d'un signal venu d'une autre source ou du processus cyclique d'établissement des profils ou encore de l'incorporation de renseignements tirés d'autres enquêtes. Par conséquent, on ne peut pas traiter les signaux indépendamment de la BDRC pour décider de procéder à l'établissement ponctuel d'un profil. Toutefois, comment devrait-on vérifier un signal en se référant à la BDRC pour voir si la BDRC a déjà été mise à jour? Par exemple, dans le cas où une déclaration d'impôt sur le revenu des sociétés indique une grosse augmentation de revenu, comment peut-on vérifier que la BDRC a déjà été mise à jour pour tenir compte de l'événement du monde réel sous-jacent à cette augmentation quand on ne sait même pas quel est cet événement?

Omission de signaux de changement

De même, certains enregistrements ne seront pas signalés. Est-ce que l'absence de signaux signifiera nécessairement qu'aucun événement du monde réel ne s'est produit qui oblige à mettre à jour la structure statistique? Devrait-on mettre au point d'autres signaux pour couvrir les omissions? Encore ici, l'étude mentionnée précédemment n'a pas donné de réponses concluantes à ces questions.

Possibilité d'obtenir des ressources en quantité suffisante pour traiter les enregistrements signalés

Comme l'étude de simulation l'a montré, beaucoup d'enregistrements seront signalés qu'il faudra traiter en partie à la main. Il est probable qu'il n'y aura pas assez de

ressources pour accomplir tout ce travail. Comment devrait-on procéder pour déterminer quelle proportion de l'ensemble des ressources devrait être consacrée à l'établissement ponctuel de profils et pour déterminer la façon d'utiliser les ressources pour traiter les enregistrements signalés? Si une rareté de ressources oblige à ignorer un certain nombre de signaux, comment allons-nous déterminer quels signaux il faut laisser tomber?

Fardeau de réponse

Les résultats de l'étude de simulation semblent indiquer qu'il faudra contacter les entreprises plus souvent qu'une fois tous les deux ans si l'on veut vérifier les changements à apporter à la base de sondage par d'autres moyens que les enquêtes régulières. Cela augmentera nécessairement le fardeau de réponse. Peut-on trouver un compromis entre l'obligation d'augmenter le fardeau de réponse pour tenir le registre à jour et la nécessité d'empêcher que le registre ne cesse d'être à jour? Quel devrait être ce compromis?

Rôle du processus d'établissement cyclique des profils

Le grand nombre d'enregistrements signalés par les déclarations d'impôt dans l'étude de simulation amène à s'interroger sur l'utilité du processus d'établissement cyclique des profils. On peut déduire des résultats de cette étude que très peu d'enregistrements devront être soumis au processus cyclique plutôt qu'au processus ponctuel. Supposons par exemple que les résultats de l'étude de simulation concernant le nombre d'enregistrements signalés continuent d'être vrais une deuxième année. Supposons ensuite que les enregistrements signalés la deuxième année ne sont pas tous les mêmes que les enregistrements signalés la première année, mais qu'un certain nombre de nouveaux enregistrements sont signalés et que certains enregistrements de la dernière année ne sont pas signalés la deuxième année. On peut alors supposer sans grand risque d'erreur que très peu d'enregistrements ne recevront aucun signal pendant deux ans. Il restera très peu d'enregistrements qui n'auront pas été signalés l'une ou l'autre année. Ce nombre représentera en fait le nombre maximum d'unités de la population à soumettre au processus d'établissement cyclique des profils. Faudra-t-il établir absolument le profil de ces entités, sachant qu'elles ne sont signalées ni par les comptes de retenue sur la paye, ni par les déclarations d'impôt?

4. CONCLUSION

Dans la partie 2, nous avons montré comment les données administratives ont été utilisées pour construire une base de sondage en vue de l'établissement de profils initiaux. Les données administratives offraient un taux de couverture très élevé. Toutefois, nous avons aussi vu que les différences conceptuelles existant entre nos besoins et les données administratives amenaient des complications qui nous ont obligés à faire des hypothèses et des compromis simplificateurs.

La base de sondage que nous avons créée incluait les éléments nécessaires à l'établissement des profils initiaux de toutes les entités commerciales à l'exception des plus complexes. Dans ces derniers cas, on ne pouvait pas accepter les approximations données par la base de sondage. Il a donc fallu faire beaucoup de recherches sur ces entités commerciales en utilisant d'autres sources d'information comme des rapports annuels publics et des réponses à des enquêtes.

La base de sondage a également joué un rôle important dans la construction et la mise en vigueur de la BDRC. Elle a servi avec le Registre des entreprises à déterminer quelles entités seraient incluses dans la partie intégrée.

La méthode suivant laquelle les données administratives seront utilisées pour établir les profils de mise à jour a été décrite dans la partie 3. Des signaux de changement seront

tirés de diverses sources de données administratives et produiront des demandes de vérification des profils. Beaucoup de questions ont été soulevées à cet égard. Les diverses équipes chargées de mettre en pratique la stratégie de mise à jour de la BDRC s'emploient actuellement à trouver des solutions à ces questions. Une des solutions à une partie des problèmes serait de classer les signaux par ordre de priorité selon, par exemple, la durée de la période depuis la dernière fois qu'on a établi le profil d'une entité. Une autre solution serait de mettre au point un système autodidacte. L'expérience dira quels signaux sont utiles et devront être conservés. Il reste donc encore beaucoup de travail à faire avant que le processus soit tout à fait opérationnel.

6. BIBLIOGRAPHIE

Colledge, M., et Lussier, R. (1985). "A Strategy for the Provision of Frame Data and Use of Tax Data for Economic Surveys", *Proceedings of the Section on Survey Methods*, 1985, American Statistical Association, Washington (D.C.).

Colledge, M. (1987). "Projet de remaniement des enquêtes-entreprises - Mise en pratique d'une nouvelle stratégie à Statistique Canada", document présenté à la *Bureau of the Census Third Annual Research Conference.*

Statistique Canada (1987). "Version 4.2 of the CFDB Data Dictionary", Projet de remaniement des enquêtes-entreprises, document de travail, 4 mars 1987.

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

L'INTÉGRATION DES DOSSIERS DES ÉTUDIANTS DANS UNE BASE DE DONNÉES DYNAMIQUE ET UN SYSTÈME DE DÉCLARATION STATISTIQUE

ANN E. HOLLINGS et BRIAN D. PETTIGREW[1]

RÉSUMÉ

À l'Université de Guelph, le Bureau du registraire représente une source importante de données pour la recherche entreprise par le Student-Environment Study Group (Groupe d'étude du milieu étudiant). La structure de ces données administratives vise à simplifier la tenue de dossiers, qui est confiée au Bureau du registraire, tout en permettant un accès facile à des dossiers particuliers. Toutefois, elle ne facilite pas l'extraction de renseignements descriptifs de la population visée : dans les faits, cette mine de données a été inaccessible aux chercheurs et aux planificateurs de l'Université. Le document décrit un système qui structure ces renseignements pour en faire une base de données intégrées et unifiées. Ce système permet de suivre les succès d'un étudiant à l'Université, de son admission à son départ, de procéder à n'importe quelle combinaison de dossiers du Bureau du registraire et de simplifier le tirage d'échantillons stratifiés aléatoires pour des enquêtes. Comme un lien est établi entre les résultats des enquêtes et les données de ce bureau, il n'est plus nécessaire de demander aux répondants de donner des renseignements déjà compris dans la base de données. Dans de nombreux cas, des réponses peuvent être apportées aux questions des chercheurs sans que l'on ait besoin de tenir une enquête, l'analyse de la sous-population visée se révélant suffisante.

1. INTRODUCTION

Dans les universités, les administrateurs et les dirigeants s'inquiètent de plus en plus du taux élevé d'abandon, surtout chez les étudiants de première année. Un nombre élevé d'abandons a un effet négatif important à plusieurs niveaux : autant pour l'établissement, lequel perd de l'argent, que pour l'étudiant, qui accuse une perte sur les plans personnel et financier (Gilbert et Gomme, 1986). À l'Université de Guelph, le SESG (Student-Environnement Study Group ou Groupe d'étude du milieu étudiant) a été mis sur pied pour examiner le milieu d'apprentissage des étudiants de premier cycle afin que soient relevés et mieux compris les problèmes pouvant se poser et que soient facilitées la formulation de directives et l'évaluation des programmes.

[1] A.E. Hollings et B.D. Pettigrew, Student-Environnement Study Group, University of Guelph, Guelph (Ontario), Canada, N1G 2W1

En règle générale, on examine le climat social chez les étudiants en menant une enquête auprès d'un échantillon pertinent : aux étudiants sélectionnés sont posées des questions sur leurs attitudes, leurs idées et leurs réactions à propos d'un certain nombre de questions. Compte tenu de la sélection des sous-populations pertinentes et de la présence, dans des enquêtes de ce genre, de facteurs semant une certaine confusion, il est souvent nécessaire de recueillir des données à caractère personnel et scolaire. Une des principales raisons de la mise sur pied du système de dépistage du SESG est le fait que la majeure partie de ces renseignements est déjà rassemblée par l'Universisté à des fins administratives. Le présent rapport précise le mécanisme qu'a instauré le SESG afin d'établir, à partir d'une base de données créée initialement pour la tenue de dossiers, une dynamique base de données unifiées permettant de suivre l'évolution d'un étudiant à l'Université, de son entrée à son départ (parce qu'il a obtenu son diplôme ou abandonné ses études).

2. SOURCE DES DONNÉES

La plupart des données stockées dans le système de dépistage proviennent des dossiers que le Bureau du registraire a établis sur les étudiants à des fins administratives. À l'origine, ces dossiers renferment des bribes distinctes de renseignements suivant le genre de données fournies; par exemple, tous les renseignements personnels, comme le nom, la date de naissance et le sexe, figurent dans un dossier personnel tandis que les renseignements relatifs à l'admission, comme la moyenne en treizième année, l'école secondaire fréquentée et le choix de programme, sont consignés dans un dossier d'admission. À mesure qu'un étudiant poursuit ses études à l'Université, certains dossiers sont ajoutés (par exemple, les notes obtenues à chaque semestre ou les demandes d'admission à un nouveau programme) et d'autres sont mis à jour (par exemple, l'adresse ou l'état matrimonial). L'utilisation directe de cette base de données pour des études rétrospectives ou longitudinales présente des problèmes, car ces renseignements sont souvent difficiles à extraire et à utiliser, et ce, même si la plus grande partie d'entre eux est conservée par le Bureau du registraire.

En réponse à diverses demandes interdépartementales, le Bureau du registraire offre un certain nombre d'ensembles de renseignements, lesquels renferment des genres particuliers de dossiers pour n'importe quel semestre. Présentés sous une forme transversale, ces ensembles contiennent les renseignements demandés pour un semestre particulier. Ces ensembles sont très utiles et suffisent à la plupart des utilisateurs; cependant, si l'on veut suivre l'évolution d'un étudiant à l'Université, l'emploi d'une base de données administratives pose des problèmes. Sous sa forme d'origine, la base de données permet d'obtenir des renseignements sur la situation actuelle de l'étudiant, par exemple, les cours auxquels il est présentement inscrit ou son adresse actuelle. Par contre, pour le dépistage, une telle structure de données pose de difficiles problèmes de nature logistique.

3. OBJECTIFS DU SYSTÈME DE DÉPISTAGE DU SESG

Système informatique composé de fichiers de données et de programmes, le système de dépistage constitue le fondement des ressources dont le SESG se sert régulièrement pour atteindre son principal objectif, soit suivre et mieux comprendre tous les aspects de la vie étudiante. Les objectifs précis du système de dépistage sont:

1. Établir une base de données rassemblant des renseignements de nature personnelle et scolaire sur tous les étudiants pendant qu'ils fréquentent l'Université.

2. Permettre l'enrichissement de la base de données établie par l'intégration des résultats d'enquête sur certains dossiers d'étudiant.

3. Suivre l'évolution de certains étudiants pendant leur passage à l'Université.

4. Permettre l'évaluation et la description de caractéristiques et de tendances de nature démographique.

5. Faciliter le tirage d'échantillons probabilistes et d'échantillons probabilistes stratifiés de la population étudiante.

Établies principalement à l'orginine pour la tenue de dossiers, les données recueillies par le Bureau du registraire constituent la ressource la plus importante à laquelle a recours le SESG pour atteindre ses objectifs. Grâce à la collaboration qui continue d'exister entre le personnel du SESG et celui du Bureau du registraire, le système de dépistage réussit à optimiser les possibilités offertes par les ressources existantes tout en réduisant au minimum la répétition des efforts déployés.

4. AVANTAGES DE L'EMPLOI DE DONNÉES ADMINISTRATIVES POUR UNE ENQUÊTE

Dans une enquête, il faut, en règle générale, demander directement à l'étudiant de donner des renseignements rétrospectifs et d'autres portant sur le programme d'études. Ci-dessous sont présentés quelques problèmes liés à cette démarche et certaines façons dont le système de dépistage peut les contourner.

1. Les données relatives à des faits dont le répondant se souvient sont souvent inexactes (par exemple, la moyenne en treizième année). Il arrive même que des données sur des faits courants soient inexactes; par exemple, il se peut qu'un étudiant déclare le programme d'étude qu'il a l'intention de suivre plutôt que celui auquel il est présentement inscrit. Dans le système de dépistage, les renseignements de ce genre figurent déjà dans un fichier et sont exacts.

2. Les répondants considèrent souvent les questions qui appellent des renseignements personnels comme une violation de leur vie privée, et ce, même si les données obtenues ne servent qu'à la production de statistiques. L'emploi d'un numéro de contrôle protège l'anonymat du répondant tout en permettant d'établir un lien important avec les renseignements demandés.

3. La répétition des efforts déployés est évitée, car on utilise des données déjà recueillies par des services administratifs de l'Université.

4. Le fardeau de réponse imposé au répondant est moins lourd, la quantité de renseignements détaillés à donner étant moins importante.

5. Des variables confondantes peuvent être intégrées aux résultats d'enquête: il est possible de relever un facteur imprévu ayant une incidence majeure sur les résultats obtenus. Dans des conditions normales, il serait impossible de demander des renseignements complémentaires aux répondants.

Des statistiques utiles peuvent donc être produites sans que l'on ait à mener une enquête auprès de la population visée. Par exemple, l'examen de la distribution des notes obtenues pour les cours de première année ou la totalisation du nombre de changements dans les principaux domaines d'étude peuvent donc être d'une très grande utilité aux planificateurs de programmes d'étude et aux organisateurs des départements. De tels renseignements peuvent être obtenus directements de la base de données.

5. SYSTÈME DE DÉPISTAGE

Le système de dépistage consiste en un ensemble de fichiers de données et de programmes informatiques, lequel vise à faciliter l'analyse statistique et la production de rapports décrivant la population étudiante qui fréquente l'Université.

La figure 1 représente schématiquement les sources de données et la structure des fichiers du système de dépistage. Au nombre de ces fichiers, il y a des fichiers principaux qui sont créés à raison d'un fichier au début de chaque semestre. Après quatre années d'exploitation, le système comptera 12 fichiers principaux (un pour chacun des trois semestres d'une année) qui renfermeront des renseignements sur tous les étudiants inscrits à l'Université. Chaque semestre qui suivra, le fichier principal le plus ancien sera envoyé aux archives afin de faire place au nouveau fichier principal créé.

Au commencement de chaque semestre, le SESG reçoit du Bureau du registraire une bande d'ordinateur contenant le nom de tous les nouveaux étudiants ainsi que des données relatives à leur admission. Il convertit ensuite en EBCDIC tous les renseignements originaux enregistrés en ASCII et les stocke sur une bande qu'il conserve dans sa bandothèque. De cette façon, les renseignements demeurent relativement accessibles (si on en a besoin pour une vérification) sans occuper la zone de manoeuvre de l'ordinateur.

Les renseignements qui intéressent tout particulièrement le SESG sont recueillis et stockés dans un fichier principal. À cette fin, un numéro de contrôle unique est attribué à chaque nouvelle entrée dans le système; c'est ce numéro, et non le numéro d'identification de l'étudiant, qui permet de relier les fichiers de données. Le fichier principal comprend donc le numéro de contrôle et certaines données tirées des dossiers du Bureau du registraire. Il compte aussi des zones en blanc correspondant à des indicateurs d'enquête relatifs à la participation à des enquêtes antérieures, ceux-ci renvoient plus particulièrement aux étudiants qui ont reçu le questionnaire de l'enquête X et à ceux qui y ont répondu. Il est ensuite très facile de voir si des étudiants ont déjà fait partie de l'échantillon d'une enquête afin d'éviter qu'une personne soit sollicitée pour trop d'enquêtes.

Les fichiers principaux sont stockés dans l'ordinateur sous la forme suivante : MasterSAA où S correspond au semestre et AA, à l'année; par exemple, le fichier MasterF87 renvoie au fichier renfermant les données relatives aux étudiants qui sont entrés à l'Université à l'automne (Fall en anglais) 1987. Dès que les notes obtenues chaque semestre sont connues, elles sont intégrées à un fichier Notes distinct comprenant des renseignements sur le semestre, la moyenne et le programme pour tous les étudiants. Il y a aussi un répertoire pour la consultation des fichiers : il indique le numéro de contrôle et le fichier principal qui renferme les dossiers de chaque étudiant.

Dans le système, il y a d'autres fichiers comprenant certains résultats d'enquête, habituellement des renseignements fournis par l'étudiant et identifiés par un numéro de contrôle.

La très grande utilité du système de dépistage est attribuable à la liberté de combinaison des fichiers qui y sont stockés. Il est possible de combiner tout ensemble de fichiers principaux et de fichiers d'enquête; de cette façon, on arrive facilement à verser dans un seul fichier de données toute la population visée, ou des groupes de celle-ci, ainsi que des renseignements sur les notes obtenues. Par exemple, on peut fusionner les fichiers principaux de l'hiver, du printemps et de l'automne 1987 pour former une cohorte de l'année civile ou on peut combiner les fichiers principaux du printemps pour 1987, 1986 et 1985 afin d'examiner les étudiants qui se sont inscrits à des cours pendant ce semestre. Dans les deux cas, il est possible de lier les numéros de contrôle et le fichier Notes en vue de brosser un tableau plus complet de la population visée.

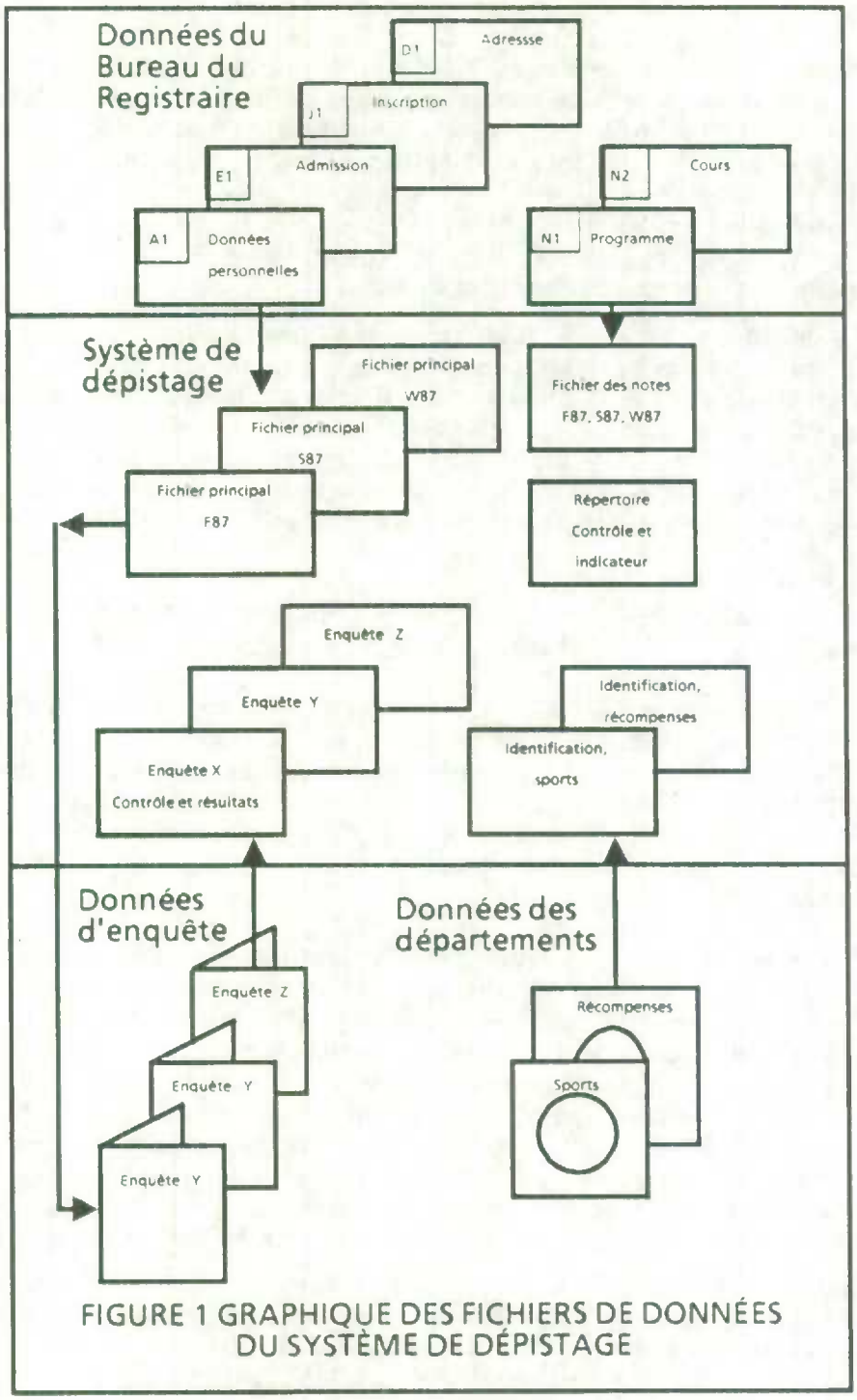


FIGURE 1 GRAPHIQUE DES FICHIERS DE DONNÉES DU SYSTÈME DE DÉPISTAGE

6. RESSOURCES

Comme le SESG exploite son système en se servant de diverses ressources informatiques, il jouit d'une grande autonomie par rapport au gros ordinateur de l'Université. La principale machine est un micro-ordinateur compatible qui a un disque dur de 30 méga-octets et dans lequel sont chargés divers progiciels, y compris LOTUS, PC-SAS, WordPerfect et un émulateur de terminal. Les données du système de dépistage sont stockées dans le gros ordinateur, et on peut y avoir accès par le micro-ordinateur en utilisant l'émulateur de terminal.

L'analyse des renseignements fournis à de nombreuses enquêtes, surtout celles portant sur de petits sous-ensembles de la population, peut facilement être réalisée avec le micro-ordinateur (au moyen de SAS), les résultats pouvant être stockés sur des disques souples. Les "données de base" de l'enquête sont téléchargées dans le gros ordinateur et stockées dans un fichier distinct. Le PC-SAS sert aussi à la production d'étiquettes et de listes pour l'envoi postal, ce qui réduit la dépendance envers les programmes utilitaires du gros ordinateur. Grâce à divers logiciels du gros ordinateur et du micro-ordinateur, un traceur permet de représenter graphiquement les données.

Pour la zone de manoeuvre du SESG réservée dans le gros ordinateur, il y a plusieurs programmes superviseurs pilotés par menu, lesquels simplifient considérablement certaines opérations, comme le transfert des données originales sur une bande du SESG, la sélection de sous-populations selon des critères précisés par l'utilisateur, la fusion des fichiers principaux et de renseignements à caractère scolaire, la sélection aléatoire d'un échantillon, la surveillance des questionnaires retournés pour les suivis et la mise à jour des dossiers (adresse actuelle, notes et documents d'enquête).

7. EXEMPLE 1 — L'ENQUÊTE UNIVERSITY RESIDENCES ENVIRONMENT SCALE (URES)

Une des enquêtes que le SESG a reproduites l'hiver dernier a été la University Residences Environment Scale (Moos et Gerst, 1974), laquelle vise à évaluer le climat social dans les résidences situées sur le campus. Les questions établies pour cette enquête sont posées à des étudiants habitant les résidences, et les résultats sont calculés pour chaque résidence. La Figure 2 montre schématiquement le mécanisme servant au tirage de l'échantillon, à l'inscription des renseignements indiqués sur les questionnaires et à l'analyse des résultats.

Étape 1: Détermination de la population à échantillonner. Quatre résidences (pavillons A, B, C et D) ont été choisies pour l'étude. Dans le système de dépistage, tous les fichiers principaux ont été fusionnés afin que tous les étudiants habitant ces résidences fassent partie de la population initiale. Pour établir les bonnes sous-populations, il a fallu se servir de renseignements tirés du dossier portant sur l'adresse actuelle. On a traité chque résidence de façon distincte afin qu'un échantillon probabiliste stratifié soit tiré. Au besoin, il aurait été possible de choisir, par exemple, seulement des étudiants ayant moins d'un certain âge, seulement des femmes inscrites à un programme d'étude particulier ou seulement des personnes qui n'avaient pas participé à l'enquête X.

Étape 2: Sélection aléatoire. À mesure que le système de dépistage établit une sous-population intéressant le SESG, il attribue un numéro d'observation à chaque personne, soit de 1 à n. La sélection aléatoire qu'effectue le système de dépistage a permis de tirer un échantillon aléatoire avec 50 % des dossiers de la façon suivante : le système a produit un nombre particulier de numéros aléatoires uniques allant de 1 à n; ces numéros aléatoires ont ensuite été reliés aux numéros d'observation attribués lors de l'établissement de la sous-population; les étudiants ainsi sélectionnés ont été inclus dans l'échantillon. Un échantillon aléatoire a été tiré pour chaque résidence.

Afin de faire ressortir ces étudiants dans les fichiers principaux, on a procédé à une mise à jour par laquelle un code alphabétique a, dans la zone correspondant à l'enquête URES, été attribué aux personnes retenues. Le nom et l'adresse des étudiants choisis ont été téléchargés dans le micro-ordinateur; le logiciel SAS a ensuite servi à produire des étiquettes et des listes de vérification pour l'envoi postal.

Étape 3: Collecte et analyse des données. Après que les délais prévus pour l'enquête eurent été échus, les données des questionnaires retournés ont été introduites dans le micro-ordinateur. Un programme SAS a servi à coter les résultats et à établir les statistiques sommaires. Au moyen de SAS et de LOTUS, le micro-ordinateur a effectué la cotation des résidences, l'analyse corrélationnelle et des graphiques illustrant, selon un certain nombre d'échelles, les cotes dégagées pour les résidences. Pour libérer la zone de manoeuvre du micro-ordinateur et rassembler en un seul endroit les données et les programmes pertinents afin de faciliter la consultation, on a stocké les programmes, les chiffriers, les fichiers de graphiques et les données brutes sur un disque souple qui a été rangé. Les cotes des résidences et les renseignements fournis par les répondants ont été téléchargés dans le gros ordinateur qui a procédé à une mise à jour des codes d'enquête des fichiers principaux afin que soient indiquées les personnes qui avaient répondu. Les codes d'enquête permettent de consigner, dans une seule zone du dossier, des renseignements particuliers sur les personnes comprises dans l'échantillon et sur celles qui ont répondu.

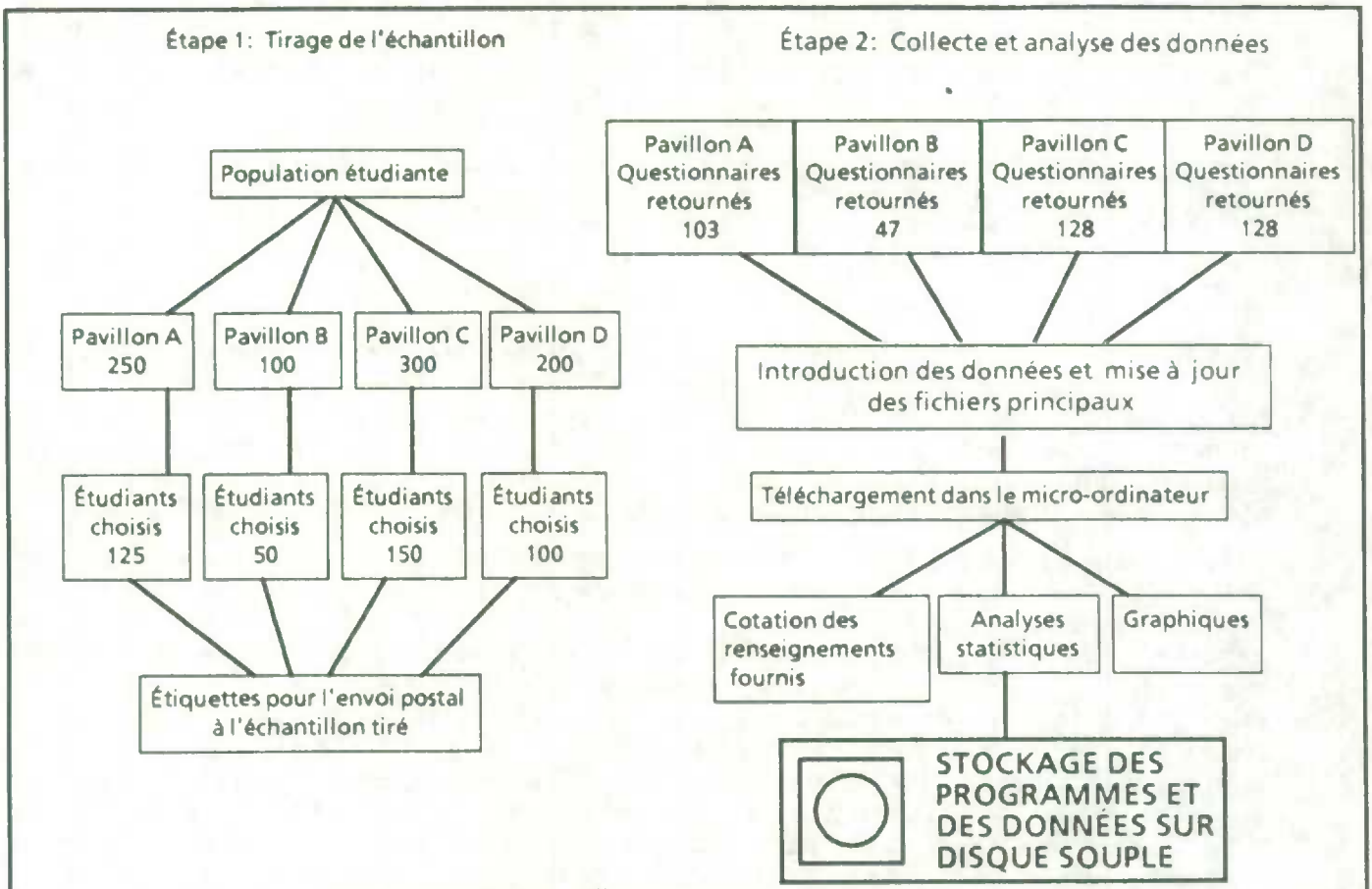


FIGURE 2 EXEMPLE DE L'ENQUÊTE UNIVERSITY RESIDENCES ENVIRONMENT SCALE

8. EXEMPLE 2 - RÉPARTITION DES NOTES OBTENUES LA PREMIÈRE ANNÉE

Le Learning Skills Task Group (Groupe de travail sur les facultés d'apprentissage) avait besoin de renseignements précis sur des cours parce qu'il participait à la mise sur pied d'un nouveau centre portant sur les facultés d'apprentissage. Il manifestait aussi un intérêt considérable à l'égard de la capacité des nouveaux étudiants à s'adapter à leur nouveau milieu scolaire, (capacité illustrée par la différence moyenne entre les notes obtenues en treizième année et celles enregistrées à leur premier semestre à l'Université).

Les renseignements tirés du système de dépistage ont été les suivants : numéro de contrôle de l'étudiant, numéro des cours et note obtenue, moyenne en treizième année, programme d'étude et sexe. Aucun nom n'était demandé, et le numéro de contrôle à servi à relier un certain nombre de dossiers de genres différents.

La répartition des notes pour tous les cours de première année a été réalisée, et les moyennes, les écarts-types et les taux d'échec ont été calculés. Pour cette partie de l'analyse, le numéro d'identification de l'étudiant n'était même pas nécessaire, le numéro des cours et la note obtenue étant suffisants. Cette analyse a permis d'examiner la répartition des notes d'une nouvelle manière : antérieurement, cet examen n'était possible que par une consultation directe des départements en jeu.

On a effectué une analyse plus poussée en recueillant, par numéro d'identification de l'étudiant, tous les renseignements sur les cours et en les reliant à ceux qui, dans les fichiers principaux, correspondent à la moyenne obtenue en treizième année. La moyenne des notes des cours suivis à la première année à l'Université et celle des notes obtenues en treizième année ont pu être comparées pour chaque étudiant (l'anonymat étant conservé), et on a aussi examiné la possibilité que le sexe et le programme d'étude aient une incidence sur la différence relevée.

9. ORIENTATIONS FUTURES

Même si, à l'origine, ce système visait à utiliser les données existantes pour le dépistage, des possibilités illimitées se sont offertes, une fois le système créé. Outre l'amélioration du menu du système, on vise à mettre à jour les renseignements tirés des données fournies par le Bureau du registraire, c'est-à-dire inclure les renseignements qui semblent maintenant utiles et éliminer ceux qui sont périmés (par exemple, les résultats obtenus à l'OTEA) ou qui ne sont pas aussi utiles que ce que l'on prévoyait au départ.

L'automne dernier, le SESG a mené une enquête qui portait sur les nouveaux étudiants (Astin, 1987) et qui a permis de dégager beaucoup de renseignements inédits sur les attentes ainsi que sur les ressources financières et personnelles des étudiants qui commencent leurs études universitaires. Cette enquête sera effectuée de façon régulière auprès des étudiants de première année, et les résultats obtenus relèveront les renseignements contenus dans les fichiers principaux. On projette aussi d'intégrer au système de dépistage les résultats de l'enquête portant sur les étudiants qui ont obtenu leur diplôme universitaire. Depuis 1976, on mène cette enquête auprès des anciens étudiants de l'Université de Guelph deux ans après l'obtention du diplôme afin de déterminer leur succès sur le marché du travail et de connaître leur évaluation de la formation universitaire reçue. Grâce au système de dépistage, il sera possible de comparer, d'une part, les succès enregistrés sur le marché du travail par les étudiants qui ont obtenu leur diplôme et, d'autre part, les attentes et les succès scolaires des nouveaux étudiants.

L'établissement d'un moyen permettant de suivre régulièrement les changements de programme et les changements de majeur dans un programme soulève aussi un intérêt. Simpson (1987) a relevé un certain nombre d'utilisations intéressantes des renseignements

de ce genre, y compris avertir assez tôt des départements que des problèmes se posent pour certains programmes lorsque, par exemple, de nombreux changements de département et de programme sont enregistrés. En ce qui concerne l'abandon des études, on manifeste aussi beaucoup d'intérêt à l'égard des majeurs, car l'incertitude à propos du programme d'étude est un indicateur d'insatisfaction et d'éventuel abandon. À mesure que s'accroît la connaissance des capacités du système, des départements qui n'avaient pas encore eu recours aux services du SESG présentent des demandes, ce qui permet à celui-ci de mieux connaître les intérêts des départements et les possibilités du système.

10. RÉSUMÉ

Le présent document décrit le système de dépistage du milieu étudiant, ses sources de données, ses ressources informatiques et certaines de ses possibilités. Il présente deux exemples qui permettent de voir la portée et l'utilité du système, et les orientations futures sont indiquées brièvement. Les points ci-dessous font état de certains avantages que présente le système et de sa dépendance envers une combinaison de données primaires et secondaires:

1. Exactitude des données à caractères personnel et scolaire.
2. Utilisation discrétionnaire de ces données pour des enquêtes et maintien de l'anonymat des étudiants observés.
3. Réduction du fardeau de réponse.
4. Utilisation efficace des données, les renseignements déjà rassemblés n'ayant pas à être recueillis.
5. Inclusion automatique de variables qui, à l'origine, n'étaient pas considérées comme importantes.
6. Obtention de renseignements utiles et récents sans avoir à mener une enquête auprès de la population visée.
7. Détermination des fins auxquelles servira le système par des groupes d'utilisateurs et d'après la documentation existante.
8. Rapidité du tirage de l'échantillon et de l'analyse des données grâce à la souplesse d'emploi des fichiers de données et des programmes informatiques.
9. Autonomie partielle d'un certain nombre de fonctions du système et faible importance des coûts garanties par un emploi judicieux des installations informatiques.

REMERCIEMENTS

Le Student-Environment Study Group tient à remercier de son apport considérable Gayle Jeffery qui a conçu la plupart des programmes du gros ordinateur s'appliquant au système de dépistage et qui a donné, pour l'établissement du système, d'utiles conseils au moment de la conceptualisation. Nous tenons aussi à exprimer notre reconnaissance à Jim Burgess et à Tom Rockola du Bureau du registraire, lesquels ont fait preuve d'une patience infinie en expliquant les caractéristiques des données originales.

BIBLIOGRAHPIE

Astin, A. (1987). The Cooperative Institutional Research Program Freshman Survey. Higher Education Research Institute, Graduate School of Education, University of California, Los Angeles CA.

Gilbert, S.N., et Gomme, I.M. (1986). Future directions in research on voluntary attrition from colleges and universities. *The Journal of the American Association of Collegiate Registrars and Admission Officers*, 61, 227-238.

Moos, R., et Gerst, M. (1974). University residence environment scale. *Consulting Psychologists Press*, Palo Alto, CA.

Simpson, W. (1987). Tracking students through majors: methodology and applications. *The Journal of Higher Education*, 58, 323-342.

SESSION IX: COMMUNICATIONS OFFERTES

Président: Geoff Lee, Australian Bureau of Statistics

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

CONTRÔLE AUTOMATISÉ DE LA QUALITÉ DES DONNÉES PROVENANT DE DOSSIERS ADMINISTRATIFS

JAMES R. JONAS et PAUL S. HANCZARYK[1]

RÉSUMÉ

Le Census Bureau se sert beaucoup de renseignements tirés de dossiers administratifs dans le cadre de ses différents programmes économiques. Même si le nombre de dossiers traités chaque année est élevé, le Bureau recevra encore plus de dossiers durant les années de recensement. Les ordinateurs centraux du Bureau appliquent des totalisations de contrôle de la qualité (CQ) aux données; toutefois, comme on a besoin d'un grand nombre de tableaux de CQ et que les ressources pour la programmation sont limitées et coûteuses, il est difficile d'établir un système complet de CQ articulé sur un ordinateur central. Si on ajoute à cela la nature délicate des données et les répercussions potentiellement très négatives de données erronées, il s'avère assez évident que l'on a besoin d'un système perfectionné d'assurance de la qualité articulé sur un micro-ordinateur. La Division des enquêtes économiques élabore actuellement un système de ce genre, qui sera mis en oeuvre pour le traitement des fichiers de données provenant des dossiers administratifs de 1987.

Le système automatisé d'assurance de la qualité intègre la technologie des micro-ordinateurs et celle des ordinateurs centraux. Les données des dossiers administratifs sont reçues chaque semaine et traitées d'abord à l'aide de programmes de CQ applicables sur un ordinateur central. Les sorties sont transférées sur un micro-ordinateur et mises dans le format approprié de façon à être transférables sur un chiffrier électronique Lotus 1-2-3. Avec le progiciel Lotus, les données sont soumises à une vérification systématique de la qualité, pendant que les fonctions d'examen des données, de détection des erreurs et de production des rapports sont effectuées automatiquement. En raison du passage d'un ordinateur central à un micro-ordinateur, le système permet de réduire la tâche du personnel de programmation, d'accroître la souplesse du personnel chargé de l'analyse, de réduire les coûts de traitement sur un ordinateur central et de fournir la composante complète d'assurance de la qualité des dossiers administratifs.

[1] James R. Jonas et Paul S. Hanczaryk, Economic Surveys Division, U.S. Bureau of the Census, Washington, D.C. 20233

1. INTRODUCTION

Aux fins de son programme de la statistique économique, le Census Bureau fait grande utilisation des données administratives, en particulier des données provenant des dossiers des déclarations d'impôt sur le revenu des entreprises du Internal Revenue Service (IRS) et, dans une moindre mesure, des dossiers de la Social Security Administration (SSA). Dans les années où ont lieu le recensement agricole et le recensement économique, le Census Bureau reçoit une quantité beaucoup plus considérable de données administratives. Ces données nous permettent de mener nos recensements économique et agricole de manière efficace, dans les meilleurs délais, de même que de réduire au minimum le fardeau de réponse des entreprises et des exploitants agricoles. La réussite de nos programmes de la statistique économique et de la statistique agricole dépend, dans une large mesure, de la qualité et de l'actualité de ces données administratives.

Il est indispensable que le Bureau vérifie la qualité de toutes les données qu'il reçoit. À cette fin, nous avons mis au point et utilisé au cours des derniers recensements économiques, des programmes de contrôle sur gros ordinateur. Toutefois, étant donné le très grand nombre de tableaux de vérification requis, et compte tenu du fait que les ressources nécessaires à la programmation sont limitées et coûteuses, il aurait été difficile d'adopter un système exhaustif de contrôle exploité entièrement sur gros ordinateur. En outre, vu le caractère confidentiel des données et les conséquences possibles des erreurs, nous avons conclu qu'il fallait mettre au point un système de contrôle plus complexe. Le Census Bureau a réussi à mettre au point un système complet pour gérer les différentes étapes du processus de vérification des données administratives. Ce système informatisé nous permettra d'effectuer une vérification plus approfondie des données malgré les contraintes budgétaires qui nous sont imposées et les ressources de programmation limitées dont nous disposons.

Le système intègre la technologie des gros ordinateurs à celle des micro-ordinateurs. Nous avons défini des critères ou exigences de base quant au genre de données administratives que nous désirons obtenir. Nous stockons ces critères dans la mémoire des micro-ordinateurs. Après exécution du programme de contrôle de la qualité sur gros ordinateur, nous transférons les résultats sur les micros-ordinateurs. Les modes de déclaration des données sont ensuite comparés à nos exigences. La vérification proprement dite (examen des données, détection des erreurs, production d'états sélectifs des erreurs) est faite sur micro-ordinateur. L'utilisation des micro-ordinateurs par opposition au recours exclusif à un gros ordinateur, permet d'alléger la charge des programmeurs, de faciliter le travail des analystes (plus de souplesse) et de réduire les coûts de traitement sur gros ordinateur. En outre, la composante "contrôle-vérification" du système assure une vérification rigoureuse et exhaustive des données. Enfin, bien que ce système ait été conçu spécialement pour la vérification des données provenant des déclarations d'impôt sur le revenu des entreprises aux fins du recensement économique, il peut être modifié (et il le sera à compter de 1988) pour effectuer la vérification de toutes les données administratives que nous recevons.

2. APERÇU DU SYSTÈME DANS UNE PERSPECTIVE DE GESTION

Le Census Bureau a beaucoup recours aux dossiers administratifs pour produire une source d'information dont l'importance n'a pas cessé de croître avec les années. Cet état de chose tient à la nécessité de produire des statistiques plus nombreuses et de meilleure qualité, de réduire au minimum le fardeau de réponse des entreprises privées et d'optimaliser l'utilisation de nos ressources humaines et financières.

Ces dernières années, les données provenant des dossiers administratifs ont été, dans l'ensemble jugées d'excellente qualité. Toutefois, les données tirées des déclarations d'impôt des entreprises qui nous ont été fournies en 1982 par l'IRS nous ont causé certains problèmes. Nous pensons, en particulier, à la qualité des principaux codes d'activité économique pour les entreprises à propriétaire unique. C'est en raison de ce problème, qu'au recensement de 1982, le Census Bureau n'a pu produire que des statistiques limitées sur les non-employeurs. Si nos programmes de contrôle de la qualité avaient été plus perfectionnés, nous aurions pu cerner plus rapidement les erreurs et en réduire les effets.

Lorsqu'est venu le moment de préparer le recensement de 1987, nous nous sommes rendu compte qu'il fallait prendre des mesures supplémentaires pour assurer la qualité des données administratives fournies par l'IRS. Ce dont nous avions besoin, c'était d'un système global de contrôle nous permettant de composer avec les trois principaux facteurs qui, dans le passé, ont compromis la qualité des ensembles de données administratives. Ces trois sources de problèmes sont les suivantes :

1. Volume élevé des données administratives

L'IRS nous fournira des enregistrements provenant des déclarations d'impôt produites par les entreprises pour 1987 (nous les recevrons en 1988). Ces enregistrements sont tirés des déclarations d'entreprises ayant différentes formes juridiques, notamment, des sociétés de capitaux, des sociétés dites "S Corporations" (petites sociétés), des sociétés étrangères, des sociétés de personnes, des entreprises à but non lucratif et des entreprises à propriétaire unique. En 1988, le Census Bureau prévoit recevoir au-delà de 75 millions d'enregistrements au total. On trouvera à l'annexe 1 le nombre approximatif d'enregistrements qui seront tirés, aux fins des recensements économique et agricole de 1987, des diverses formules d'impôt. Il est clair que le nombre d'enregistrements que nous recevrons est extrêmement élevé. Toutefois, la raison pour laquelle nous avons envisagé la mise au point d'un système de contrôle de la qualité aussi perfectionné ne tient pas uniquement à la quantité des données à traiter mais aussi à la complexité du traitement. Un enregistrement contient souvent plusieurs données élémentaires et chacune d'elles vient accroître la complexité (niveau de détail) de l'enregistrement lui-même et de l'ensemble du fichier. En outre, les différentes formules d'impôt ne contiennent pas toutes les mêmes éléments d'information et les revenus n'y sont pas tous déclarés de la même façon. Par conséquent, non seulement faut-il vérifier la qualité de plus de 75 millions d'enregistrements mais aussi la qualité des données élémentaires qu'ils contiennent. L'annexe 2 contient les éléments d'information demandés dans chacune des formules d'impôt sur le revenu des entreprises pour l'année 1987.

Soulignons enfin que les entreprises transmettent leurs déclarations d'impôt à l'un des dix centres de traitement de l'IRS. Chaque centre assure le traitement des déclarations qu'il reçoit et la qualité des données qui nous sont transmises peut varier. Le Bureau vérifie donc individuellement la qualité des données transmises par chacun d'eux.

2. Contraintes budgétaires

Les contraintes budgétaires sont un autre facteur important qui vient accroître la difficulté du contrôle de la qualité des données administratives. Compte tenu de la politique générale de restriction des dépenses du gouvernement, le Bureau essaie d'offrir des services plus nombreux à un moindre coût. La charge de travail en programmation est beaucoup plus lourde pendant les années de recensement, mais les programmeurs ne sont pas plus nombreux. La vérification de la qualité, qui dépend dans une large mesure des diverses ressources informatiques existantes, peut en souffrir. Il faut également souligner que dans le passé, la plupart des tâches de

vérification de la qualité ont été faites sur un gros ordinateur. L'utilisation de l'ordinateur central coûte cher, d'autant plus que les fichiers sont de plus en plus importants.

3. Manque de communication entre les organismes

Par le passé, le manque de communication entre les organismes intéressés ont nui à la qualité des données. Pour corriger ce problème, il était indispensable de définir clairement les axes de communication devant exister entre le Bureau et l'organisme qui lui fournit les données, et ceci, pour chacune des étapes du travail. Le Bureau doit d'abord définir les fichiers de données et les enregistrements dont il a besoin puis s'entendre avec l'organisme sur les données qui peuvent lui être fournies. Certaines des données demandées ne sont peut être pas disponibles ou encore peuvent être trop coûteuses à produire. Les difficultés doivent être aplanies au fur et à mesure car des retards dans la transmission des données pourraient en compromettre l'utilité. En outre, les organismes doivent s'entendre sur la qualité et la quantité des données désirées. Les exigences du Bureau quant à la quantité des données requises doivent être bien définies.

C'est pour apporter une solution globale au problème posé dans le passé par la qualité des données administratives que le Bureau a décidé d'élaborer et de mettre en oeuvre son système informatisé de contrôle de la qualité : le système permet de vérifier des fichiers importants et complexes de l'IRS, favorise les échanges et permet la détection immédiate des erreurs. La composante "vérification informatisée" est la plus importante du système. En gros, le Bureau définit les critères ou exigences de base auxquels doivent répondre les données fournies par l'IRS. Il compare ensuite les modes de déclaration aux critères établis puis procède, sur micro-ordinateur, à la vérification systématique des données. Le Bureau produit ensuite un rapport de la situation dans lequel il indique si les données répondent ou non aux exigences.

Le personnel du Bureau définit ses exigences bien avant que les données ne lui soient transmises. L'IRS dispose donc de tout le temps voulu pour déterminer s'il peut raisonnablement fournir les données demandées et, au besoin, pour modifier les demandes. Les exigences définies par le Bureau portent sur les deux points suivants : fréquence de transmission et qualité des données. Pour ce qui est de la fréquence de transmission, le Bureau établit une estimation du total des déclarations qu'il souhaite obtenir pour chaque type d'entreprise avec la date à laquelle il désire les recevoir. Pour ce qui est des exigences qualitatives, le Bureau définit les modes de déclaration désirés pour chaque type d'élément d'information.

Le système de vérification automatisé facilite l'analyse des données. Une série de tableaux de résultats sont produits afin de comparer les données fournies aux critères établis. Des indicateurs permettent de relever les données qui ne satisfont pas aux exigences. Cette façon de procéder réduit les risques d'omission pendant le processus d'analyse.

Chaque mois, le Bureau envoie à l'IRS des rapports de la situation dans lesquels les modes de déclaration des données sont comparés aux critères établis. Ces rapports constituent des sous-ensembles des tableaux détaillés des résultats et ne contiennent que les exigences de base pour les ensembles de données fournis par l'IRS. Ces rapports favorisent la communication entre le Bureau et l'IRS. Les données posant des problèmes y sont indiquées et le Bureau et l'IRS doivent décider immédiatement des mesures correctives à prendre (notamment, reprise des erreurs) afin d'éviter de compromettre les résultats du recensement. À cet égard, il est indispensable que les mesures nécessaires soient prises dans les plus brefs délais parce que l'IRS ne conserve pas ses bandes de données indéfiniment. Si les erreurs ne sont pas décelées rapidement et si les mesures

correctives ne sont pas prises à temps, la reprise risque d'être impossible ou encore d'être extrêmement coûteuse.

Le système de contrôle de la qualité ne nous permet pas de nous assurer que les données administratives ne poseront plus jamais de difficultés. Toutefois, il nous permet de définir très clairement nos exigences, de sorte que les caractéristiques des ensembles de données ne sont pas laissées au hasard. Il nous permet par ailleurs de déceler les erreurs le plus rapidement possible.

3. PARTICULARITÉS DU SYSTÈME

Les fichiers de données que le Bureau reçoit chaque semaine sont d'abord soumis à un contrôle qualitatif sur gros ordinateur. Les programmes utilisés sont mis au point bien avant l'arrivée des fichiers et ils servent à produire les premiers tableaux de contrôle de la qualité, qui constituent la pierre angulaire de tout le système. Auparavant, les programmeurs travaillant sur gros ordinateur étaient chargés de créer les tableaux dans leur ensemble, c'est-à-dire, y compris les cellules de données et le texte qui s'y rapporte (en-têtes et colonnes principales). Toutefois, dans les programmes de production de tableaux pour le recensement de 1987, ces deux composantes seront traitées séparément. La production des totalisations continuera de se faire sur gros ordinateur tandis que le texte des tableaux sera produit à l'aide de micro-ordinateurs par des employés qui ne sont pas des spécialistes de la programmation. Une procédure a été mise au point pour permettre la production de tableaux pour tous les fichiers de données administratives. Cette procédure a permis au Census Bureau de concevoir un programme pour micro-ordinateur permettant de construire des images de tableaux pour n'importe quel genre de fichier de données administratives. Une fois achevées, les images de tableaux sont transférées au gros ordinateur et sont utilisées pour aligner les données contenues dans les fichiers. Le travail de programmation qu'exigent les tableaux de contrôle de la qualité est de beaucoup simplifié du fait que la construction des images de tableaux est effectuée par des non-spécialistes. Les programmeurs travaillant sur gros ordinateur peuvent donc concentrer leurs efforts uniquement sur l'établissement des totalisations. On trouvera à l'annexe 3 un exemple des divers tableaux produits sur gros ordinateur pour chaque type d'entreprise. Ce tableau fait voir la répartition pondérée des formules 1040, annexe C, selon le centre de traitement et la tranche de revenu net.

Le gros ordinateur n'effectue que les totalisations de base (mise à jour des totalisations). Ces dernières sont ensuite transférées sur des micro-ordinateurs où toutes les autres opérations de vérification sont effectuées. Parmi celles-ci, mentionnons le calcul des pourcentages dans les nouvelles totalisations, la production de totalisations cumulatives, la vérification de données clés et la production de rapports sur l'état d'avancement des travaux de vérification. Cette méthode de traitement systématique, qui s'appuie essentiellement sur l'utilisation de micro-ordinateurs, offre plus de souplesse pour l'examen des données et permet de réduire la charge de travail des programmeurs.

Les totalisations produites sur gros ordinateur sont transférées sur micro, sur une feuille de calcul LOTUS 1-2-3. Cette feuille de travail contient également les critères (ou exigences) définis par le Census Bureau relativement aux ensembles de données administratives. Le micro-ordinateur effectue automatiquement la vérification des données du tableau. Les écarts entre les données et les exigences établies sont indiqués dans les tableaux de résultats. Les deux principaux avantages du système de vérification sont les suivants :

1. Il nous permet de déceler aisément les problèmes. Les données qui ne répondent pas aux normes sont signalées par un indicateur afin qu'un analyste les examine. On risque donc moins d'oublier des erreurs.

2. Il attire notre attention sur des données qu'il faudrait examiner de plus près. Souvent, bien que toutes nos exigences aient été remplies, les tableaux de résultats nous mettent sur la piste de certains problèmes: par exemple, le système nous permet d'examiner plus en détail certaines tendances inattendues. En fait, les tableaux de résultats nous permettent de concentrer notre attention sur les données qui peuvent poser des difficultés. Il peut s'agir de demander à un centre de traitement de l'IRS d'examiner à nouveau les données ou encore, de transférer les enregistrements présentant certaines caractéristiques. Ensuite, nous examinons ces enregistrements à la main afin de cerner les problèmes.

Comme nous l'avons souligné, nous avons défini les exigences qualitatives de base auxquelles doivent satisfaire les données pour les recensements économique et agricole de 1987. Notre système automatisé (comparaison des données fournies aux critères établis), nous permet de déterminer immédiatement si les données sont de qualité acceptable.

Après examen et vérification des données du cycle en cours, des tableaux cumulatifs sont produits sur micro-ordinateur. Le fait d'avoir recours à des micros plutôt qu'à un gros ordinateur pour la production de ces tableaux favorise une utilisation plus efficace de nos ressources. Premièrement, le système nous évite d'avoir à garder en mémoire du gros ordinateur des fichiers cumulatifs, ce qui réduit les coûts. Auparavant, ces fichiers cumulatifs étaient conservés dans le gros ordinateur et étaient combinés aux fichiers du cycle en cours pour produire de nouveaux tableaux cumulatifs. En utilisant les micro-ordinateurs, nous avons pu intégrer au progiciel LOTUS 1-2-3 des formules simples qui nous permettent de créer des tableaux cumulatifs à très peu de frais. Deuxièmement, la production de ces tableaux cumulatifs n'engage pas la création de programmes pour gros ordinateur. Des imprimés des tableaux cumulatifs sont produits et conservés pour analyse ou référence.

Outre cet ensemble complet de tableaux cumulatifs, nous produisons également un ensemble de tableaux de résultats. Ces tableaux contiennent des comparaisons détaillées de certaines données clés pour le cycle en cours. L'annexe 4 montre un des nombreux tableaux des résultats produits par le système. On y trouve le nombre et le pourcentage de formules 1040 - annexe F envoyées par chaque centre de traitement de même que le pourcentage attendu de formules. Comme on peut le voir, les chiffres cumulatifs sont raisonnables et répondent aux critères établis. Si des incohérences avaient été relevées, un indicateur aurait été affiché en regard du nom du centre de traitement concerné. Le dernier type de rapport produit par le système automatisé est le rapport de la situation qui est un bilan détaillé du nombre cumulatif de fichiers envoyés par l'IRS. Dans ce rapport, on compare la qualité globale des ensembles de données aux critères établis (délais pour la transmission des données et exigences qualitatives). Ces rapports sont transmis à l'IRS une fois par mois environ. Comme nous l'avons souligné, ces rapports, qui constituent un état récapitulatif de la qualité des données, favorisent les échanges entre le personnel du Bureau et celui de l'IRS.

4. RÉSULTATS DU CONTRÔLE DE LA QUALITÉ

Les rapports de la situation permettent au Census Bureau et à l'IRS de déceler les problèmes au fur et à mesure et ils facilitent la collaboration. Il est à noter cependant que lorsque nos critères ne sont pas respectés (délais ou qualité), la plupart du temps c'est en raison d'un changement dans les modes de déclaration des entreprises. Des cas semblables ne posent aucune difficulté pour l'IRS (il n'a pas de mesure corrective à

prendre) mais, pour le Bureau, ils peuvent entraîner des frais de traitement informatique supplémentaires. Aux recensements économique et agricole, le Bureau a relevé un cas semblable (voir le tableau présenté à l'annexe 5) : à la fin de mai 1987, le Census Bureau avait reçu environ 697,600 formules 1120 (sociétés); le nombre requis étant de 760,000. L'objectif n'a donc pas été atteint. Par contre, le Bureau a reçu un nombre beaucoup plus élevé de formules 1120S ("S Corporations") que le nombre exigé (environ 328,850 au lieu de 225,000). La cause de ce changement est la suivante : avec l'adoption de la nouvelle loi fiscale, il devenait plus avantageux pour les entreprises de remplir une formule 1120S qu'une formule 1120. Ainsi, bien qu'il n'y ait eu aucune erreur dans les données, il était important que le Bureau relève ce changement pour pouvoir modifier ses méthodes de traitement. Nous sommes en train de mettre au point des mesures pour tenir compte de cet accroissement du nombre des "S Corporations". L'annexe 6 présente un des tableaux que nous utilisons pour étudier la qualité des données. Comme on peut le voir, la qualité des données répond aux exigences. Si, pour un élément d'information donné, nos critères n'étaient pas respectés, un indicateur serait affiché pour qu'un analyste fasse les recherches nécessaires.

Le système informatisé de contrôle de la qualité sera tout à fait prêt pour effectuer le traitement des fichiers de l'IRS de 1987. Des systèmes-pilotes ont été utilisés (et continuent de l'être) pour évaluer les fichiers de 1985 et de 1986. C'est grâce à ces systèmes que le Bureau a pu mener à bien la vérification des données administratives transmises par l'IRS pour les recensements agricole et économique.

L'utilisation combinée de micro-ordinateurs et d'un gros ordinateur a permis au Census Bureau de vérifier, d'une manière efficace et exhaustive, les très gros fichiers de données de l'IRS et d'en assurer la qualité. En outre, le système a allégé la charge des programmeurs et il continuera de le faire parce qu'une bonne partie du travail peut être faite sur micro-ordinateurs, par des non-spécialistes. Par ailleurs, puisque le système a permis de réduire l'utilisation du gros ordinateur et la charge de travail des programmeurs, ces derniers peuvent concentrer leurs efforts sur les totalisations de base. Enfin, en ce qui concerne l'analyse des données, le système répond aux besoins de tout le personnel. Les états récapitulatifs permettent aux gestionnaires de déterminer rapidement et efficacement si les délais de transmission sont respectés et si la qualité des données est satisfaisante. Les analystes, pour leur part, consultent les rapports hebdomadaires plus détaillés qui leur indiquent, le cas échéant, les ensembles de données nécessitant un examen plus approfondi.

5. RÉSUMÉ

Le Census Bureau a mis au point un système global de contrôle de la qualité qui permet de déceler les problèmes risquant de nuire à la qualité des données. Avec le système, il est possible de traiter de très gros fichiers de données. En outre, comme le système favorise la collaboration, il permet au Census Bureau de régler rapidement les problèmes de manière à assurer la transmission adéquate des données. Les critères de qualité et les délais de transmission sont définis conjointement par le Bureau et l'IRS. Le respect de ces critères est vérifié par l'ordinateur. Par ailleurs, grâce à l'informatisation, le Bureau est en mesure d'utiliser au mieux les programmeurs et les ressources budgétaires limitées dont il dispose pour effectuer la vérification. De plus, le recours aux micro-ordinateurs, parce qu'il offre beaucoup de souplesse d'utilisation, a permis d'accroître le rôle des analystes et de diminuer la charge de travail des programmeurs. Les échanges entre le Census Bureau et l'IRS étant nombreux, les personnes intéressées

sont plus sensibles à l'importance de la qualité des données. Ces échanges permettent également de mettre au point des procédures efficaces de transmission des données. Enfin, grâce à cet effort collectif pour régler les problèmes, le Census Bureau est maintenant assuré de recevoir les données requises pour mener à bien les recensements économique et agricole de 1987.

Pièces jointes

Annexe 1
Nombre approximatif d'enregistrements tirés de dossiers administratifs utilisés aux recensements économique et agricole de 1987, selon le genre de formule et l'année d'imposition

Genre de fichier	Nombre d'enregistrements		
	1985	1986	1987
Déclaration d'impôt sur le revenu des entreprises:	2,617,000	20,051,000	30,881,000
Form ule 1040, annexe C	--	11,750,000	12,500,000
Formule 1040, annexe F	2,450,000	2,450,000	--
Formule 1040, annexe SE	--	--	10,000,000
Formule 1120	42,000	2,550,000	2,650,000
Formule 1120-A	--	200,000	210,000
Formule 1120F	--	11,000	11,000
Formule 1120S	17,000	900,000	950,000
Formule 1065	108,000	1,750,000	1,800,000
Formule 990	--	380,000	400,000
Formule 990-PF	--	35,000	35,000
Formule 990-T	--	25,000	25,000
Formule 1120S, annexe K-1	--	--	700,000
Formule 1065, annexe K-1	--	--	1,600,000
Fichiers des données fiscales (données annuelles):	41,950,000	43,500,000	45,050,000
IRS Fichier principal des entreprises	24,000,000	25,000,000	26,000,000
IRS Fichier de la paye et du personnel	17,000,000	17,500,000	18,000,000
SSA Fichier des nouvelles entreprises	950,000	1,000,000	1,050,000
Total	44,567,000	63,551,000	75,931,000

Annexe 2
Éléments d'information tirés de diverses formules d'impôt sur le revenu des entreprises (1987)

Élément d'information	Formule								
	1040 Annexe C	1040 Annexe SE	1065	1120 & 1120-A	1120 S	1120 F	990	990-PF	990-T
Nom et prénom du contribuable, nombre de caractères	X								
Adresse postale et ville du contribuable, nombre de caractères	X								
No de sécurité sociale du contribuable SSN	X	X							
No de sécurité sociale du conjoint SSN	X								
EIN	X		X	X	X	X	X	X	X
PIA	X		X	X	X	X			
Revenu brut (chiffre des ventes)	X		X 1/	X 1/	X 1/	X 1/			X 1/
Rendus et rabais sur ventes	X								
Exercice financier	X	X	X	X	X	X	X	X	X
Salaires	X								
Coût de la main-d'oeuvre	X								
Code de fin d'année	X		X		X				
Nombre de mois d'exploitation	X		X		X				
Indicateur 914	X								
Redevances brutes				X					
Revenus bruts déclarés							X		
Total des revenus								X	
Nom de famille et no de sécurité sociale des associés (maximum de 10) 2/			X						
Nom de famille, no de sécurité sociale det nombre de parts (jusqu'à 10 actionnaires) 2/					X				
Revenu net, exploitation agricole		X							
Revenu net, autre activité que l'exploitation agricole		X							
Revenu net total provenant du travail autonome		X							

1/ Moins les rendus et rabais sur ventes
2/ Données tirées de l'annexe K-1.

Annexe 3

Répartition pondérée des formules 1040, annexe C, selon le centre de traitement et la tranche de revenu net 1

Centre de traitement	Total	Tranche de revenu net										
		<0	en blanc ou 0	1-2,499	2,500-4,999	5,000-9,999	10,000-24,999	25,000-49,999	50,000-99,999	100,000-249,999	250,000-499,999	500,000 +
Ensemble des centres												
Atlanta												
Philadelphie												
Austin												
Cincinnati												
Kansas City												
Andover												
Ogden												
Brookhaven												
Memphis												
Fresno												
Autres												

1/ Revenus bruts moins les rendus et rabais sur les ventes.

Annexe 4

Répartition pondérée des formules 1040, annexe F (1986), en pourcentage, selon le centre de traitement

Année d'imposition	Annexes F reçues	Centre de traitement										
		Atlanta	Phila-delphie	Austin	Cin-cinnati	Kansas City	Andover	Ogden	Brook-haven	Memphis	Fresno	Autres
1986												
En nombre	2,087,200	176,700	71,600	374,900	262,100	358,600	118,800	343,200	40,300	288,100	52,500	400
En pourcentage	100.0	8.5	3.4	18.0	12.6	17.2	5.7	16.4	1.9	13.8	2.5	0.0
Objectif visé												
En pourcentage	100.0	8.5	3.0	18.5	11.5	17.5	5.5	16.5	2.0	14.0	2.5	0.0
Objectif non atteint 1/												

1/ Objectif atteint: écart de plus ou moins 2%

Annexe 5
Répartition pondérée des formules 1120 (1986), en nombre, selon la date

Date	Formules 1120		Objectif non atteint
	Nombre reçu	Nombre exigé	
Fin mars 1987	326,500	303,000	
Fin april 1987	697,600	760,000	Objectif non atteint
Fin mai 1987		988,000	
Fin juin 1987		1,190,000	
Fin juillet 1987		1,418,000	
Fin août 1987		1,621,000	
Fin january 1988		2,077,000	
Fin october 1988		2,533,000	

Répartition pondérée des formules 1120S (1986), en nombre, selon la date

Date	Formules 1120S		Objectif non atteint
	Nombre reçu	Nombre exigé	
Late March 1987	103,350	90,000	
Late April 1987	328,850	225,000	
Late May 1987		292,000	
Late June 1987		352,000	
Late July 1987		420,000	
Late August 1987		480,000	
Late January 1988		615,000	
Late October 1988		750,000	

Annexe 6
Modes de déclaration des éléments d'information
des formules 1120S, 1986 (chiffres pondérés)

Eléments d'information	Pourcentage de formules 1120S		Objectif non atteint
	reçu	exigé	
EIN			
Espaces laissés en blanc, zéros, données non numériques	0.0	Moins de 1.0	
Code IRD invalide	0.0	Moins de 1.0	
CODE PBA	0.0	Moins de 6.0	
Espaces laissés en blanc ou données non numériques	11.5	Moins de 18.0	
Espaces laissés en blanc, données non numériques, codes non classés ou codes PBA invalides	20.9	Moins de 40.0	
REVENUS OU CHIFFRE DES VENTES BRUTS MOINS RENDUS ET RABAIS			
Espaces laissés en blanc, zéros ou données non numériques	45.7	30.0 - 60.0	
Enregistrements pour lesquels un chiffre est fourni, pourcentage dans chacune des tranches ci-après:			
- moins de $100,000	36.9	20.0 - 50.0	
- $100,000 ou plus mais moins de $500,000	17.4	10.0 - 30.0	
- $500,000 ou plus	0.0	Moins de 1.0	
EXERCICE FINANCIER			
Espaces laissés en blanc, zéros ou données non numériques			

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

ALGORITHME POUR LA DÉTERMINATION DE RÈGLES OPTIMALES POUR L'APPARIEMENT D'ENREGISTREMENTS PROVENANT DE DEUX SOURCES

YASAR YESILCAY[1]

RÉSUMÉ

L'algorithme que nous présentons est une méthode par étapes de recherche des règles optimales pour l'appariement d'enregistrements provenant de deux sources. Ces règles sont optimales dans le sens que l'erreur d'appariement nette résultante ou biais d'appariement (B) et l'erreur d'appariement brute résultante (G) satisfont aux inégalités suivantes:

$$0 \leq G \leq g \text{ et } |B| \leq b$$

où b et g sont des niveaux maximaux préétablis en ce qui concerne les erreurs B et G respectivement et |B| correspond à la valeur absolue du biais.

1. INTRODUCTION

Dunn (1946) donne une définition de l'appariement des enregistrements dans un style poétique: (traduction libre)

> Toute personne venue au monde écrit son propre Livre de la vie. Ce livre débute à la naissance et se termine à la mort. Ses pages contiennent les enregistrements des principaux événements survenus pendant l'existence de la personne. L'appariement des enregistrements est le nom donné au processus qui consiste à assembler les pages du livre pour en faire un volume.

Si l'on interprète le terme "personne" de cette définition comme un élément d'une population, qu'il s'agisse d'une population d'évènements, de personnes ou de choses, la définition correspond alors au sens donné à l'appariement dans un grand nombre de domaines. Le livre dont il est question dans la définition peut n'avoir que deux pages (ou enregistrements) et le présent exposé traite précisément de ce cas, c'est-à-dire de l'appariement d'enregistrements provenant de deux sources différentes. La procédure que nous exposons ici peut évidemment être étendue (moyennant une modification appropriée de la définition de certains des concepts en cause) à l'appariement d'enregistrements provenant de sources plus nombreuses. L'appariement d'enregistrements provenant de deux sources ou plus est appliqué dans un grand nombre de domaines tels que les affaires, l'histoire, la démographie, la sociologie, l'éducation, les voyages et l'extraction des données, pour n'en citer que quelques-uns.

[1] Yasar Yesilcay, Université du Sultan Qaboos, College of Science, P.O. Box: 32486 AL-Khoudh, Muscat, Sultanate of Oman.

Si l'on peut attribuer un code d'identification unique et permanent à chacun des éléments de la population ou de l'échantillon étudiés et si ce code est reporté fidèlement sur chacun des enregistrements produits pour ces éléments, il n'est alors pas nécessaire de rechercher une règle d'appariement. Les enregistrements peuvent être appariés à partir de leur code, ceux qui portent le même code étant considérés comme corrrespondant au même élément et, par voie de conséquence, couplés (ou appariés); les enregistrements restants sont considérés comme des enregistrements non appariés. Pour être **idéale** (parfaite) une telle donnée d'identification doit présenter les caractéristiques suivantes: être unique, permanente, disponible et très différenciatrice (Nitzberg et Sardy, 1965). Or, dans la plupart des applications, les données d'identification utilisables ont une partie seulement de ces caractéristiques ou même n'en ont aucune.

L'utilisation de données d'identification imparfaites pour l'appariement des enregistrements peut engendrer trois types d'erreurs de couplage (ou d'appariement): enregistrement apparié par erreur, enregistrement non apparié par erreur ou enregistrement incorrectement apparié. Lorsqu'un enregistrement provenant d'une source n'a pas d'enregistrement correspondant dans une autre source mais est néanmoins apparié à un enregistrement de cette deuxième source, il s'agit d'un enregistrement apparié par erreur. De même, lorsqu'un enregistrement provenant d'une source a un enregistrement correspondant dans une autre source mais n'est pas apparié à un enregistrement de cette deuxième source, il s'agit d'un enregistrement non apparié par erreur. Enfin, lorsqu'un enregistrement provenant d'une source a un enregistrement correspondant dans une autre source mais est apparié à un autre enregistrement de cette deuxième source, il s'agit d'un enregistrement incorrectement apparié. (Pour une description et une analyse détaillées de ces erreurs, voir Marks, Seltzer et Krotki, 1974, ou Yesilcay, 1975.)

Les règles concernant les opérations d'appariement des enregistrements visent à réduire les erreurs de ce genre dans les cas où l'on doit utiliser des données d'identification imparfaites. Toutefois, c'est un fait établi que des règles trop "rigides", si elles permettent de réduire le nombre d'enregistrements appariés par erreur et d'enregistrements incorrectement appariés, entraînent un trop grand nombre d'enregistrements non appariés par erreur. À l'inverse, des règles trop "flexibles" donnent moins d'enregistrements non appariés par erreur mais un trop grand nombre d'enregistrements appariés par erreur et d'enregistrements incorrectement appariés. Il faut donc trouver un moyen terme entre ces deux extrêmes. L'équilibre recherché est obtenu si l'on a recours à un ensemble de règles "optimales", la définition d'optimal dépendant de l'utilisation prévue des résultats de l'appariement et de l'incidence possible des erreurs d'appariement sur cette utilisation.

Sunter et Fellegi (1967) proposent une approche pour la détermination de règles optimales selon laquelle on compare toutes les paires possibles d'enregistrements constituées à partir de deux sources différentes et on les classe comme enregistrements appariés, enregistrements non appariés ou enregistrements pouvant être appariés. On définit alors la règle optimale d'appariement comme l'ensemble de règles permettant de réduire au minimum la probabilité de classer des paires comme enregistrements pouvant être appariés et, par conséquent, d'augmenter au maximum la probabilité d'une classification positive des enregistrements pour certains taux d'erreurs. La théorie qui sous-tend cette approche a été révisée ultérieurement par Sunter (1968) et à nouveau par Fellegi et Sunter (1969).

Tepping (1955, 1960, 1968) a élaboré une méthode d'appariement des enregistrements pour vérifier l'acquittement des abonnements; il a par la suite (1969) adopté cette méthode comme système de double enregistrement. Ses règles optimales visent à réduire le coût des enregistrements appariés par erreur et des enregistrements incorrectement appariés et requièrent l'utilisation du coût de chaque unité considérée. On a fait valoir (Yesilcay, 1975) que l'imputation de ces coûts n'est pas réaliste dans un système de double enregistrement bien qu'elle ait été facilement faite dans le cas des abonnements. Par

ailleurs, Tepping traite les enregistrements incorrectement appariés comme des enregistrements appariés par erreur alors que, dans le cas d'un système de double enregistrement, les enregistrements incorrectement appariés n'ont aucune incidence sur l'estimation du nombre total d'événements ni sur la variance de cette estimation.

En ce qui a trait aux systèmes de double enregistrement, Marks, Seltzer et Krotki (1974) ont défini les règles optimales comme un ensemble de règles permettant d'avoir le même nombre d'enregistrements appariés par erreur et d'enregistrements non appariés par erreur, ce qui donne une erreur (ou un biais) d'appariement nette nulle (= B = nombre d'enregistrements appariés par erreur - nombre d'enregistrements non appariés par erreur) et une erreur d'appariement brute minimale (= G = nombre d'enregistrements appariés par erreur + nombre d'enregistrements non appariés par erreur). Bien que ces critères produisent une estimation non biaisée du nombre total d'événements, Yesilcay (1975) a montré que l'erreur d'appariement brute influe sur la variance de l'estimation et que, dans certains cas, une estimation biaisée peut être préférable en termes d'erreur quadratique moyenne. En conséquence, les critères d'optimalité sont modifiés et les règles d'appariement optimales sont définies comme un ensemble de règles qui satisfait aux inégalités suivantes:

$$0 \leq G \leq g \text{ et } |B| \leq b$$

où b et g sont des taux d'erreur préétablis et où B et G correspondent à l'erreur d'appariement nette et à l'erreur d'appariement brute précédemment définies. Il ne faut pas perdre de vue que les valeurs de B et G diffèrent en fonction de la source dans laquelle sont comptés les enregistrements appariés par erreur et les enregistrements non appariés par erreur. Il est suggéré que ces enregistrements soient comptés dans la source qui contient le moins d'enregistrements.

Un ensemble d'axiomes doit supporter l'algorithme présenté ici qui permet de déterminer les règles optimales telles que nous les avons définies précédemment. Ces axiomes sont décrits dans la section suivante.

2. AXIOMES

Le programme de recherche conçu pour déterminer la règle optimale pour l'appariement des enregistrements est basé sur les axiomes suivants.

1. On dispose d'enregistrements provenant de deux sources; pour chaque élément de la population ou de l'échantillon, il existe au plus un enregistrement dans chaque source.
2. Ces enregistrements contiennent des données sur quelques variables qui permettent de décider s'il existe ou non une paire d'enregistrements, tirés de chaque source, se rapportant à un même élément. Une donnée de ce type est appelée donnée d'appariement.
3. Les données d'appariement sont imparfaites, en ce sens qu'elles ne présentent pas une ou plusieurs des caractéristiques exposées à la section précédente.
4. Quand on compare les enregistrements provenant de deux sources, certaines erreurs d'appariement (ou de couplage) sont généralement inévitables et, même si elles sont indésirables, ces erreurs peuvent être tolérées.
5. Le fait d'accroître le nombre de variables en fonction desquelles les enregistrements sont comparés simultanément réduit le nombre d'enregistrements

appariés par erreur mais accroît le nombre d'enregistrements non appariés par erreur dans les deux sources.

6. Bien qu'il soit souhaitable d'avoir une "concordance exacte" en ce qui a trait à chaque variable pour réduire le nombre des enregistrements appariés par erreur et pour simplifier l'appariement, il est parfois préférable d'opter pour un accord relatif respectant certaines limites prescrites car cela permet de réduire le nombre d'enregistrements non appariés par erreur. Cependant, un niveau de tolérance trop "flexible" risque d'aboutir également à un trop grand nombre d'enregistrements appariés par erreur et d'enregistrements incorrectement appariés et n'est donc pas souhaitable; ainsi, pour chaque variable, il existe une **"concordance optimale par rapport au niveau de tolérance"** qui dépend elle-même des autres variables utilisées pour le couplage et de la concordance optimale par rapport au niveau de tolérance correspondant à celles-ci.

7. Pour déterminer les règles d'appariement, on dispose d'un échantillon d'enregistrements tiré de chaque système pour lequel on connaît le **statut d'appariement véritable** de chaque enregistrement. Ces enregistrements peuvent provenir d'un essai préliminaire ou des étapes précédentes de l'opération en cours. Il est essentiel que la qualité des données contenues dans ces enregistrements soit représentative de la qualité des enregistrements qui doivent être appariés au moyen des règles ainsi déterminées. C'est pourquoi il est préférable d'utiliser un échantillon tiré au hasard parmi les enregistrements visés par l'opération en cours et de tester périodiquement les règles obtenues lorsqu'il y a lieu de croire que les conditions de l'opération ont changé.

Il ne fait pas de doute que nombre de ces axiomes sont nécessaires à toute approche ou algorithme visant à déterminer des règles d'appariement (optimales ou non). La procédure utilisée pour déterminer le véritable statut d'appariement d'un échantillon d'enregistrements est expliquée dans la prochaine section.

3. STATUT D'APPARIEMENT VÉRITABLE D'UN ÉCHANTILLON D'ENREGISTREMENTS

Il est recommandé d'utiliser la procédure suggérée par Marks, Seltzer et Krotki (1974) et présentée en résumé ci-après pour déterminer le statut d'appariement véritable de chacun des enregistrements de l'échantillon.

Selon cette procédure, trois spécialistes ou plus, ou trois groupes de spécialistes ou plus, travaillant **indépendamment,** doivent classifier chacun des enregistrements comme étant appariés ou non appariés, en utilisant leurs propres règles d'appariement implicites. Parlant de leur utilisation de cette procédure dans les Philippines, Madigan et Wells (1974) précisent que:

> "À ce stade...., tout renseignement contenu dans les enregistrements, toute caractéristique culturelle connue, toute connaissance spécialisée des procédures d'interview et toute autre information ou connaissance de même nature doivent être utilisés..."

dans le cadre des règles implicites. Une fois que les trois spécialistes ont pris une décision concernant chacun des enregistrements, l'étape suivante consiste à comparer ces décisions. Si les trois spécialistes s'accordent pour dire que telle paire d'enregistrements constituent un appariement, on considère alors que tel est le statut d'appariement "véritable" de cette paire d'enregistrements. De la même façon, si les trois spécialistes s'accordent pour dire qu'à tel enregistrement d'une source donnée ne corrrespond aucun

enregistrement dans l'autre source, on considère que le statut d'appariement "véritable" de l'enregistrement est d'être non apparié. En cas de désaccord, tous les spécialistes sont invités à réexaminer leur décision. S'ils ne sont toujours pas d'accord, l'enregistrement est alors retourné sur le terrain avec une demande de vérification et de renseignements supplémentaires. On retourne également sur le terrain les enregistrements non appariés et, éventuellement, un échantillon d'enregistrements appariés pour connaître le détail des évènements conceptuels, temporels et géographiques hors du champ de l'enquête et pour obtenir des renseignements plus précis. Ensuite, les enregistrements sont à nouveau analysés et une décision est prise à partir des renseignements supplémentaires qui ont été obtenus, les enregistrements faisant partie du champ de l'enquête étant classifiés comme "véritablement" appariés ou "véritablement" non appariés.

Il est évident que cette procédure, qui correspond à la meilleure décision possible étant donné les conditions dans lesquelles on a déterminé le statut d'appariement de chaque enregistrement (sans erreur, espère-t-on), ne peut être appliquée à chaque opération d'appariement, particulièrement si l'opération doit être répétée un grand nombre de fois ou si le nombre d'enregistrements devant être appariés est très important. Le statut d'appariement véritable des enregistrements du petit échantillon est déterminé par la procédure présentée ci-dessus afin que les enregistrements puissent être utilisés pour déterminer le statut d'appariement de tous les enregistrements en cause immédiatement ou ultérieurement.

Nous pouvons maintenant fixer un ensemble de règles explicites qui donnent le statut d'appariement des enregistrements contenus dans l'échantillon d'une façon moins onéreuse, moins laborieuse à appliquer et moins exigeante en temps tout en satisfaisant les critères d'optimalité qui régissent de telles règles.

Marks, Seltzer et Krotki (1974) exposent en détail une telle procédure en adoptant comme critères d'optimalité une erreur d'appariement nette nulle et une erreur d'appariement brute minimale. Bien que de telles règles soient souhaitables dans certaines conditions, on a fait observer que leur application n'est pas toujours possible (Yesilcay, 1965). En outre, le nombre estimatif d'événements obtenu avec cette procédure, quoique non biaisé, peut avoir une variance due à G plus importante que dans le cas d'une estimation biaisée.

L'algorithme exposé ici recherche l'optimalité de façon quelque peu différente sur la base des inégalités $0 \leq G \leq g$ et $|B| \leq b$ où b et g sont des taux d'erreur établis en vue d'obtenir l'erreur quadratique moyenne minimale du nombre estimatif total d'événements dans un système de double enregistrement. Pour d'autres applications, ces valeurs peuvent être modifiées en fonction des objectifs particuliers à l'application en cause. En utilisant cet algorithme, il est toujours possible, dans la mesure où cela est faisable, d'atteindre l'optimalité définie par Marks, Seltzer et Krotki (1974) en posant $b = 0$.

La procédure qui détermine les règles d'appariement pour satisfaire aux critères précités est appelée la méthode d'appariement par étapes (MAÉ) et est expliquée dans la section suivante.

4. MÉTHODE PAR ÉTAPES SERVANT À DÉTERMINER LES RÈGLES D'APPARIEMENT OPTIMALES

Selon cette méthode, on élabore les règles d'appariement optimales en deux étapes ou plus. De cette manière, il n'est pas nécessaire d'envisager tous les seuils de tolérance possibles pour toutes les combinaisons de variables possibles en vue de respecter les limites (établies) d'acceptation des erreurs, le nombre de combinaisons possibles étant trop considérable pour être traité, même par un ordinateur, dans des délais raisonnables.

À la première étape, on identifie autant d'appariements véritables qu'il est possible de le faire et on les élimine de la série d'enregistrements avec un minimum d'enregistrements appariés par erreur. Pour ce faire, on apparie les enregistrements provenant de deux sources selon toutes les combinaisons possibles d'un petit nombre de variables (de 3 à 5) en respectant les niveaux de tolérance "exacts" ou le niveau de tolérance "rigide" établi. Parmi les combinaisons qui donnent les mêmes nombres minimaux d'enregistrements appariés par erreur, l'algorithme sélectionne la combinaison qui présente le plus grand nombre d'appariements véritables. À cette étape, le nombre d'enregistrements non appariés par erreur peut être important et le but de l'étape II est de réduire ce nombre à un niveau raisonnablement bas qui soit acceptable compte tenu des taux d'erreur admis.

À l'étape II, seuls les enregistrements qui ont été classifiés comme non appariés à l'étape précédente font l'objet d'une comparaison. Afin de réduire le coût et l'application de la recherche, à la seconde étape celle-ci est limitée à toutes les combinaisons possibles de deux variables. En outre, le nombre maximal de niveaux de tolérance "raisonnables" est limité à cinq pour chaque variable. Enfin, le nombre total de variables pouvant être utilisées dans le cadre de l'appariement est limité à vingt. Lorsque le nombre de variables disponibles dépasse vingt, il faut sélectionner les vingt plus différenciatrices. Marks, Seltzer et Krotki suggèrent d'utiliser l'erreur d'appariement brute comme substitut du pouvoir différenciateur d'une variable. Yesilcay (1975) propose d'adopter la fréquence modale de la variable comme substitut du pouvoir différenciateur, une fréquence modale faible étant l'indice d'un pouvoir différenciateur élevé.

Une caractéristique particulière de l'algorithme MAÉ est sa possibilité de contrôler la limite supérieure de l'erreur d'appariement brute. Le processus de recherche élimine à l'étape I les combinaisons qui ne semblent pas devoir répondre à ce critère. À l'étape II, chaque combinaison est à nouveau soumise à ce critère et on considère comme étant une possibilité sans intérêt toute combinaison dont la somme des enregistrements appariés par erreur aux deux étapes et des enregistrements non appariés par erreur à l'étape II est supérieure à g, limite de l'erreur d'appariement brute. En outre, si à l'étape I ce critère est satisfait en utilisant des combinaisons de trois variables, on s'abstient alors de rechercher les combinaisons de quatre ou cinq variables afin de réduire le coût et de simplifier le processus d'appariement. De la même façon, les combinaisons de cinq variables ne font l'objet d'une recherche que si le critère ne peut être satisfait avec les combinaisons de trois ou quatre variables. Le tableau 1 illustre le déroulement opérationnel des étapes I et II.

5. CONCLUSION

Une version légèrement différente du programme a été jugée satisfaisante après avoir été soumise à un essai, mené aux Philippines en utilisant les données contenues dans un système de double enregistrement (Madigan et Wells, 1974, et Yesilcay, 1975).

On peut se procurer auprès de l'auteur une version imprimée du sous-programme de recherche.

Tableau 1

Étapes opérationnelles de la méthode d'appariement par étapes (MAÉ)

ÉTAPE I

1. Établissez le statut d'appariement "véritable" d'un échantillon d'enregistrements.
2. Précisez le nombre de variables qu'il est possible d'utiliser pour l'appariement (NVAR $\leq$ 20), le code de la source où les erreurs doivent être comptées (ISMALL), l'erreur brute minimale, g, et la valeur absolue de l'erreur nette maximale, b.
3. Fixez le nombre de combinaisons de variables, NV = 3.
4. Procédez à l'appariement en fonction des niveaux de tolérance "exacts" pour toutes les combinaisons de variables NV.
5. Sélectionnez les combinaisons qui ont donné le nombre minimal d'enregistrements appariés par erreur et le nombre maximal d'appariements véritables.
6. Si le nombre minimal d'enregistrements appariés par erreur est $\leq$ g/4, passez à l'étape II. La combinaison de variables qui donne ce résultat doit être utilisée à l'étape I de l'opération d'appariement.
7. Si le nombre minimal d'enregistrements appariés par erreur est > g/4, posez NV = NV + 1 et passez à l'instruction 3.

   REPRENEZ ENSUITE LES INSTRUCTIONS 4 À 7 JUSQU'À CE QUE LE NOMBRE MINIMAL D'ENREGISTREMENTS APPARIÉS PAR ERREUR SOIT $\leq$ g/4, OU NV > 5 .

ÉTAPE II

1. Utilisez les enregistrements de l'étape précédente qui ont été déclarés comme non appariés.
2. Appariez les enregistrements pour toutes les **combinaisons** possibles **de deux variables** selon des niveaux de tolérance raisonnables.
3. Parmi toutes ces combinaisons, sélectionnez celle qui donne une erreur brute $\leq$ g et un biais absolu $\leq$ b . La combinaison de variables qui donne ce résultat et les niveaux de tolérance correspondants à ces variables doivent être utilisés à l'étape II de l'opération d'appariement.
4. Si aucune des combinaisons essayées précédemment ne satisfait les critères, reprenez l'étape II en appliquant les instructions aux enregistrements non appariés par erreur à l'étape II. Cela équivaut à l'étape III. Continuez au besoin avec les étapes IV, V, etc. jusqu'à ce que les résultats recherchés soient atteints.

BIBLIOGRAPHIE

Dunn, H.L. (1946). "Record Linkage", *American Journal of Public Health* 36, 14121416.

Fellegi, Ivan P., et Sunter, Alan B. (1969). "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64, 1183-1210.

Madigan, Francis C., et Wells, H. Bradley (1974). "Report on Matching Procedures of a Dual Records System in the Southern Phillipines", rapport inédit.

Marks, Eli S., Seltzer, W., et Krotki, Karol J. (1974). *Population Growth Estimation, A Handbook of Vital Statistics Measurement*, The Population Council, New York (1975).

Nitzberg, David M., et Sardy, Hyman (1965). "The Methodology of Computer Linkage of Health and Vital Records", *procès-verbal, Social Statistics Section, American Statistical Association*, 1963, 100-106.

Sunter, Alan B. (1969). "A Statistical Approach to Record Linkage", E.D. Acheson, rédacteur, *Record Linkage in Medecine, procès-verbal du symposium international, Oxford, juillet 1967*, London, E & S. Livingstone, Ltd.

Sunter, Alan B., et Fellegi, Ivan P. (1969). "An Optimal Theory of Record Linkage", *Bulletin of the International Statistical Institute, procès-verbal de la 36^{e} séance, Sydney, 1967*, vol. XLII, livre 2, 809-838.

Tepping, Benjamin J. (1955). *Study of Matching Techniques for Subscription Fulfillment*, Philadelphia, National Analyst Inc.

Tepping, Benjamin J. (1960). *Progress Report on the 1959 Matching Study*, Philadelphia, National Analyst Inc.

Tepping, Benjamin J. (1968). "A Model for Optimal Linkage of Records", *Journal of the American Statistical Association* 36, n^{o} 324, 1321-1332.

Tepping, Benjamin J. (1969). "The Application of Linkage Model to the Chandrasekar-Deming Technique for Estimating Vital Events", document rédigé à l'occasion du séminaire du Population Council sur les Optimum Procedures for Matching two Lists of Vital Events, New York, 17 avril 1969.

Yesilcay, Yasar (1975). *The Mean Square Error of the Estimate as a Criterion for the Assessment of Alternative Approaches to Matching in a Dual Record System*, Institute of Statistics, série mimeo, n^{o} 1035, The University of North Carolina, Chapel Hill.

SESSION X: DISCUSSION EN PANEL

Président: G.J. Brackstone, Statistique Canada

Symposium de Statistique Canada
sur les utilisations statistiques
des données administratives
novembre 1987

NOTES POUR UNE DISCUSSION EN PANEL SUR LES UTILISATIONS STATISTIQUES DES DONNÉES ADMINISTRATIVES

JOHN W. GRACE[1]

Mon rôle à titre de membre de ce panel revêt un caractère à la fois d'humilité et de découragement. En effet, je suis appelé à représenter 25 millions de Canadiens vivant une vie souvent désorganisée, parfois déréglée et fréquemment carrément monotone à exprimer tant leurs craintes que leur indifférence et à me faire leur porte-parole, par exemple celui du jeune homme divorcé qui détient un diplôme universitaire en sciences infirmières, réagit positivement à l'examen de tuberculose et a 1.9 enfant, un four micro-ondes et l'eau chaude.

(Pour ceux d'entre vous qui ne reconnaissent pas l'exemple que j'ai choisi, il s'agit d'un amalgame de plusieurs bases de données de Statistique Canada figurant dans le Répertoire des renseignements personnels.) Je suis sûr qu'il ne saurait pas ce que sont les données administratives, encore moins ce que vous en faites. Moi non plus d'ailleurs. Certes, lorsqu'on m'a demandé de vous adresser la parole, je n'étais même pas certain de la signification de "données administratives".

Je savais toutefois que la **Loi sur la protection des renseignements personnels** définit comme suit "fins administratives".

> "Destination de l'usage de renseignements personnels concernant un individu dans le cadre d'une décision le touchant directement."

Les droits d'accès aux renseignements personnels sont clairement définis. Quant aux données administratives, elle se situent dans une zone indéterminée.

Je ne veux pas semblé présomptueux en me faisant le porte-parole du jeune homme précité ni d'aucun des 25 millions de Canadiens. La plupart peuvent très bien parler en leur nom, et le font effectivement, lorsque par exemple des questions d'enquête les irritent. D'ailleurs, de plus en plus de Canadiens s'en trouvent offusqués. Or, bon nombre d'entre eux continuent de ne prêter aucune attention aux renseignements que Statistique Canada détient à leur sujet. Ce sont les Canadiens qui n'ont "rien à cacher", aveu qui ma toujours semblé dénoter des personnes plutôt ennuyeuses.

Toutefois, je parle effectivement au nom des Canadiens parce que le Parlement a voté la **Loi sur la protection des renseignements personnels,** qui établit les règles fondamentales régissant la collecte et utilisation de renseignements personnels par le gouvernement fédéral. En vertu de cette loi, il incombe au commissaire à la protection de la vie privée d'étudier les plaintes et de veiller à ce que le gouvernement respecte la Loi. D'ailleurs, je me suis suffisamment bien renseigné sur l'objet de votre réunion pour conclure que les enquêtes sur les données administratives, étant donné qu'elles sont menées indirectement, ne permenttent effectivement pas aux Canadiens d'exprimer leur point de vue parce qu'ils ne savent tout simplement pas à quelles fins peuvent servir les renseignements qu'ils fournissent.

---

[1] John W. Grace, Commissaire à la protection de la vie privée du Canada, 112 rue Kent, Ottawa, Ontario K1A 1H3

Laissez-moi inscrire mes commentaires dans ce contexte. Comme certains d'entre vous me l'ont peut-être déjà entendu dire au cours d'une conférence précédente, je reconnais que le Bureau est sensible à la question de la protection de la vie privée. Il l'a prouvé en créant un conseil de l'accès à l'information et de la protection des renseignements personnels, conseil chargé d'examiner les demandes de nature délicate ou contentieuse. Et je sais que les organismes statistiques des pays démocratiques de l'Ouest défendent également le principe de la protection de la vie privée. Guy Labossière et moi-même avons eu le privilège d'assister en juin à une conférence organisée par Statistique Suède à laquelle ont participé comme alliés et non comme ennemis des statisticiens et des protecteurs de données.

Cependant, le souci de la protection de la vie privée que partagent les démocraties occidentales n'a pas empêché les gouvernements de coupler et d'assortir leurs bases de données administratives afin d'obtenir, quelles que soient les raisons valables invoquées, suffisamment de profils détaillés sur les particuliers pour faire frémir le plus tolérant des protecteurs de données.

Les utilisations statistiques des données administratives sont visées par notre tendance soutenue vers la surveillance des données. En tant que citoyens, nous trouvons répugnante la surveillance électronique ou la surveillance des dossiers médicaux. Toutefois, la violation éventuelle de la vie privée que représentent le couplage et l'exploitation de diverses bases de données en vue de l'établissement du profil de particuliers ou de groupes de particuliers qui ne se doutent de rien fait pâlir le grand frère d'Orwell.

Évidemment, de nombreux Canadiens se sont déjà résignés à perdre la mainmise sur les renseignements qu'ils fournissent à leur sujet. Ils peuvent ne pas aimer les fins auxquelles servent les renseignements, si toutefois ils les connaissent, mais ils semblent dangereusement résignés. Je trouve cela malsain pour la société.

Les statisticiens et les chercheurs sont les bénéficiaires de cette résignation. Après tout, une enquête sur des documents administratifs obtient un taux de réponse de 100 pour cent, permet de recueillir des réponses claires, entraîne des coûts administratifs moins élevés et cause moins d'exaspération que d'autres enquêtes. Mais où est-ce que tout cela s'arrêtera? Quels aspects de ma vie est-il raisonnable que l'État connaisse dans l'intérêt de la recherche ou de la planification sociale? Je ne pense pas que ni les statisticiens ni les citoyens aient abordé suffisamment en profondeur cette question fondamentale.

Nous devrions peut-être délaisser pour un moment les possibilités futures et vos formidables nouvelles techniques pour examiner les enjeux.

À ma connaissance, le Canada n'a pas déployé d'efforts systématiques en matière de recherche. Aux États-Unis, les recherches effectués entre 1979 et 1982 sur la qualité des casiers judiciaires qu'établit le FBI ont permis de constater que 25.7 pour cent seulement de ces casiers étaient complets, exacts et sans ambiguïté.

Le FBI a aussi examiné 453 mandats délivrés à Mobile (Alabama) dans le cadre de son programme de 1984 en matière de contrôle de la qualité des données. Sur ce nombre, 338 faisaient état d'une grandeur de 7 pieds et 11 pouces, d'un poids de 499 livres et de cheveux de couleur "XXX". Comme le signalement joue un rôle tellement important dans l'application de la loi, ces constatations n'ont rien d'amusant. Je me sers de cet exemple de qualité inaceptable des données parce qu'il est emprunté à un sytème de classement documentaire dont la nature même devrait dicter un contrôle qualificatif plus rigoureux que la plupart des systèmes de classement de documents administratifs.

La qualité des dossiers personnels que tient le gouvernement deviendra pour moi un domaine d'intérêt spécial à mesure que je vérifierai si les institutions gouvenementales respectent les dispositions de la **Loi sur la protection des renseignements personnels.**

Le paragraphe 6 (2) de la Loi se lit comme suit.

> "Une institution fédérale est tenue de veiller, dans la mesure du possible, à ce que les renseignements personnels qu'elle utilise à des fins administratives soient à jour, exacts et complets."

Évidemment, la mesure du possible variera selon la catégorie du dossier et la fin à laquelle il servira. Toutefois, une chose est certaine: l'absence de mécanismes visant à assurer la qualité des données constitue une infraction à la Loi.

D'après nos observations, il est également clair que l'utilisation de documents administratifs à des fins statistiques supposent qu'on consent à se fier à des données qui sont parfois très douteuses. J'ai fait remarquer que Gerry Gates a dit avoir reçu du IRS des renseignements qui, à son avis, étaient "loin d'être parfaits .

Je m'inquiète aussi du fait que nous risquerions de perdre la capacité d'assurer le respect des droits d'accès aux renseignements personnels et de correction de ceux-ci si les données administratives nominatives étaient préférées en vue de la recherche statistique. En vertu de la **Loi sur la protection des renseignements personnels,** le particulier a le droit de voir les documents qui le concernent et de demander que des corrections leur soient apportés. L'exercice de ce droit présuppose que le particulier peut déterminer l'endroit où sont conservés ces documents et que le ministère peut trouver tous les exemplaires de ceux-ci si des corrections s'imposent.

Plus les ministères se partagent les renseignements personnels, plus il est difficile d'accomplir cette tâche. Il est parfois même impossible de la faire, et le droit à la protection de la vie privée s'en trouve diminué.

Même si vous ne partagez peut être pas tous mes soupçons, je pourrais peut-être vous demander de considérer au moins les questions qui suivent. Les réponses sont essentielles à la collecte de données de nature délicate et à la protection efficace des renseignements personnels, quels que soient vos lieux de résidence et de travail. Les principes se trouvent à la base de tous les codes internationaux de protection des renseignements personnels. Vous devriez pouvoir répondre "oui" à chacune des ces questions.

Questions 1. Tenez-vous un registre de toutes vos bases de données administratives contenant des renseignements personnels, et comment les utilisez-vous ces bases?

Question 2. Les personnes visées par vos bases de données administratives savent-elles à quelles fins servent les renseignements qui les concernent? Consentent-elles à cette utilisation?

Question 3. Avez-vous pris toutes les mesures possibles pour éviter la divulgation de renseignements personnels? Avez-vous considéré le codage de toutes vos données administratives de manière à empêcher leur identification au cas où elles tomberaient dans les mains de personnes mal intentionnées?

Question 4. Êtes-vous sûr de l'exactitude des données et pourriez-vous les corriger si la personne concernée prouvaient qu'elles sont erronées?

Question 5. Pouvez-vous fournir à la personne concernée l'accès aux données?

Enfin, avez-vous établi une politique officielle, cohérente et largement diffusée sur l'utilisation d'identificateurs personnels dans les données administratives?

Seules la connaissance et l'application répandues des principes de protection des données permettent aux organismes statistiques d'éviter aux citoyens le genre de surprise

que les Suédois ont reçu l'an dernier. J'estime qu'à titre de Nord-Américains ayant une opinion moins favorable de l'État, notre réaction serait nettement plus dur.

Voilà la perspective vue de l'autre côté de la barricade. J'ai l'intention de prendre position. Certains d'entre vous me trouveront peut-être obstructionniste, voir paranoïaque. Mais il est suffisamment connu que le Congrès des États-Unis est sur le point de voter un projet de loi visant à renverser la tendance au couplage des immenses bases de données administratives, tendance qui tournait en dérision la loi américaine sur la protection des renseignements personnels.

En vertu du projet de loi appelé "Computer Matching and Privacy Protection Act of 1986", les organismes qui s'échangent des données seront tenus de signer des ententes, aux termes desquelles ils devront établir le but, la justification et la base juridique de l'assortiment, écrire le programme d'assortiment en détail, informer les personnes dont les dossiers sont assortis, vérifier l'exactitude des nouveaux dossiers et en assurer la sécurité. Aux termes de ce projet de loi, on créerait aussi des conseils d'intégrité des données chargés de veiller sur les programmes d'assortiment informatisé de l'État.

Notre propre gouvernement a reconnu que le couplage des données constitue une menace réelle à la protection des renseignements personnels, que les citoyens ne devraient pas être mis à nu par la technologie et les méthodes statistiques. Il a établi, il y a un mois seulement, les règles du jeu, dont l'une stipule que les organismes devront m'informer 60 jours à l'avance de l'assoriment qu'ils comptent faire afin que je puisse recommander qu'il n'ait pas lieu ou dire au public qu'il aura effectivement lieu. Au fait, la décision du gouvernement de restreindre l'utilisation du numéro d'assurance sociale en tant qu'identificateur de l'administration fédérale constitue une autre initiative, qui peut même vous compliquer la vie.

Je ne veux pas que vous interprétiez aujourd'hui mes commentaires comme une interdiction de toute forme de couplage de données. La gestion de l'État est de plus en plus une grosse affaire. Et le Parlement est appelé à choisir parmi des produits concurentiels. Toutefois, je suis heureux qu'il ait reconnu que leur représentant, qui est aussi l'ombudsman en matière de protection des renseignements personnels, a un rôle à jouer.

En dernière analyse, les statisticiens décident de leur propre sort. Pour des valeurs plus importantes que la simple efficience, il faut continuer d'être conscient et respectueux des principes de protection des renseignements personnels pour assurer dans la mesure du possible l'accès à des sources directes et indirectes de données personnelles. Je crois que vous partagez ce cet engagement.

Statistique Canada Symposium
sur les utilisations statistiques
des données administratives
novembre 1987

REMARQUES FORMULÉES LORS D'UNE DISCUSSION EN PANEL

THOMAS B. JABINE[1]

Je fais ma présentation en deux parties. On peut dire que la première est météorologique, puisqu'elle se compose de quelques courtes remarques sur les changements climatiques aux États-Unis depuis 1970, qui ont influé sur les utilisations statistiques et autres des données administratives. En deuxième lieu, je présenterai avec un certain détail quelques résultats de questions d'enquête sur les attitudes du public face aux utilisations statistiques et autres des données administratives aux États-Unis. Ces résultats sont disponibles depuis peu et devraient aider à combler quelques-unes des lacunes que présente notre connaissance de ce que pense le public au sujet des genres d'activité que nous avons examinés au cours de ce symposium.

Le Census Bureau et les autres organismes statistiques américains "d'un certain âge" ont toujours profondément cru en la nécessité de protéger le caractère confidentiel des données sur les personnes, sur lesquelles leurs statistiques sont fondées, que ces données aient été collectées par voie d'enquête ou puisées à des sources administratives. Durant les années 1970, cependant, les questions de confidentialité et d'accès aux données ont pris une importance nouvelle pour tous les organismes statistiques et pour les dépositaires et autres utilisateurs de données administratives par suite d'un certain nombre d'événements marquants:

- La publication en 1973 du rapport du **Advisory Committee on Automated Personal Data Systems** au secrétaire à la Santé, à l'Éducation et au Bien-être des États-Unis. Ce rapport, intitulé **Records, Computers and the Rights of Citizens,** recommandait une série de dispositions législatives et administratives en vue d'établir des pratiques justes ayant trait aux données personnelles détenues par le gouvernement des États-Unis.
- L'adoption en 1974 du Privacy Act qui donnait force de loi à plusieurs recommandations du comité consultatif susmentionné. Cette loi est en plusieurs points semblable aux dispositions législatives adoptées plus récemment au Canada.
- Les études entreprises et les rapports publiés par la Privacy Protection Study Commission entre 1975 et 1977.
- L'adoption du Tax Reform Act de 1976 (à ne pas confondre avec le Tax Reform Act de 1986, plus importante encore). La loi de 1976 restreignait l'utilisation non fiscale des données du Internal Revenue Service (IRS) par d'autres agences à un nombre restreint de fins explicitement mentionnées dans la Loi.

Tous ces événements ont donné naissance à un nouveau milieu d'utilisation statistique des données administratives. Certaines utilisations importantes ont dû prendre fin: On n'a plus jugé possible, par exemple, aux termes des dispositions du Tax Reform Act de 1976 (Buckler and Smith, 1980), de rendre publiques les données microfilmées du Continuous

[1] Thomas B. Jabine, Consultant en Statistique, 3231 Nord-Ouest rue Worthington, Washington, D.C. 20015, États-Unis

Work History Sample de la Social Security Administration. Comme l'exigeait le Privacy Act, les organismes obtenant des renseignements des personnes à des fins administratives ou statistiques, ou les deux, ont commencé à mieux informer les gens des conditions dans lesquelles ont leur demandait de fournir des renseignements et de l'utilisation qu'on comptait en faire. Les organismes se sont sensibilisés graduellement à la possibilité de divulguer par inadvertance des données individuelles en publiant des données d'ensemble ou des micro-données à des fins statistiques. Ils ont donc révisé leurs politiques de diffusion des données et, d'une manière générale, diminué la quantité de détails que pouvaient comprendre les totalisations et les micro-données rendues publiques.

Ces tendances se sont maintenues dans les années 80, et de nouveaux éléments ont apparus. On s'est fortement opposé au recensement de la population et à certains types d'enquête dans plusieurs pays d'Europe occidentale, et il y a eu dans certains cas une chute du niveau et de la qualité des réponses. La mise au point de méthodes efficaces de couplage des enregistrements au moyen de puissants ordinateurs a accru la possibilité qu'un "agresseur" résolu identifie une ou plusieurs personnes dont les données faisaient partie d'un micro-fichier publié (Jabine et Scheuren, 1986).

Aussi bizarre que la chose puisse sembler en l'occurence, le recours au couplage des enregistrements à des fins de conformité a augmenté. Le Congrès des États-Unis a tenu le Internal Revenue Service de mettre les données fiscales à la disposition des organismes d'état pour les aider à identifier les bénéficiaires de prestations de bien-être social qui n'y ont pas droit et les personnes qui n'effectuent pas leurs paiements de pension alimentaire ou de remboursement d'emprunt. Les organismes statistiques prennent cependant bien soin d'entreprendre des activités qui pourront sembler moins menaçantes pour le public. Les normes régissant la publication de micro-fiches deviennent plus strictes et l'on hésite beaucoup à effectuer de nouveaux couplages d'enregistrements à des fins statistiques, même s'il y a des précédents.

Les statisticiens officiels justifient ces politiques en disant qu'elles sont essentielles à la protection de leur avoir le plus important, la confiance du public, car c'est elle qui porte la grande majorité des répondants à collaborer aux grands recensements et aux enquêtes. La plupart d'entre nous conviendrons que les attitudes et perceptions du public relatives à la collecte, au partage et à la diffusion des données déterminent dans une grande mesure ce qui peut être fait; il est donc important pour nous d'apprendre tout ce que nous pouvons au sujet de ces attitudes et de la façon dont elles évoluent avec le temps.

Nous avons eu l'occasion dernièrement de recueillir des données sur les attitudes du public au sujet du partage de statistiques entre les organismes, en insérant une série de questions à ce sujet dans une enquête sur les attitudes des contribuables, effectuée à l'été de 1987 par Louis Harris and Associates pour le Internal Revenue Service. Avant de vous en communiquer les résultats, j'ai deux observations à caractère général au sujet de la collecte des données sur les attitudes du public. D'abord, contrairement à certains autres types de données d'enquête examinées au cours de ce symposium, la qualité des données sur les attitudes ne peut pas être contrôlée par voie de comparaison avec des données administratives. Deuxièmement, les variables à observer sont définies dans une grande mesure par le libellé des questions, la manière dont les réponses possibles sont présentées aux répondants et le contexte global dans lequel les questions sont posées. C'est pourquoi vous trouverez dans le document distribué (et annexé) les résultats avec les sections pertinentes des questionnaires. Je ne vous parlerai que des réponses aux questions 97 à 99, qui touchent le partage entre agences des données.

La population visée par cette enquête se composait de personnes qui, normalement, font une déclaration d'impôt sur le revenu. Dans le cas des couples présentant une déclaration conjointe, seulement le plus connaissant des deux membres du couple a été interrogé. Le questionnaire a été présenté par voie d'entrevues personnelles. En

élaborant les questions sur le partage des données, nous avons commencé par supposer que bien peu de gens avaient vraiment réfléchi sur la question du partage de données entre agences, ou même en étaient bien informés. Nous avons donc abordé le sujet en trois étapes. Pour commencer, nous avons posé une série de questions (97, a à f) touchant certains facteurs sous-jacents pouvant déterminer l'attitude du répondant relativement au partage de données. Ensuite, nous avons posé une seule question (98a) pour connaître l'attitude du répondant quant au principe même du partage des données entre agences. Le répondant pouvait ensuite, en réponse à la question 98b, dire pourquoi il était de cet avis. Enfin, on demandait au répondant ce qu'il pensait du transfert de données du Internal Revenue Service à certaines agences, à des fins précisées (99 a à d). On donnait au répondant quatre exemples, dont deux à des fins statistiques (99 a et d) et deux à des fins administratives (99 b et c). Pour compenser l'effet de l'ordre dans lequel les questions étaient posées, ce dernier a varié de façon aléatoire entre les répondants.

Il ressort des résultats qu'une majorité appréciable des répondants voulaient savoir quelles agences possédaient des renseignements sur eux (97a) et pourquoi ces agences voulaient cette information (97b). Une majorité moins forte a convenu que le partage des données réduirait le nombre d'enquêtes à subir par le public (97d) et le coût, pour le gouvernement, de l'information dont il a besoin (97e). Pour ce qui est de la plus grande confiance inspirée par telle agence plutôt que par telle autre (97 c et f), les résultats ont été bien peu concluants et dépendaient de la manière dont la question était posée.

Les résultats de la question générale sur les attitudes face au partage des données (98a) concordent avec les résultats des questions précédentes, relatives aux facteurs pouvant influer sur les attitudes au sujet du partage des données. Compte non tenu des indécis et des non-réponses, à peu près la même proportion de répondants favorisaient et défavorisaient le partage des données (38 % contre 41 %), mais la proportion des répondants très opposés était sensiblement plus élevée que celle des répondants résolument en faveur (23 % contre 14 %).

Malgré ces opinions relativement négatives sur toute la question du partage des données, nous avons constaté que si la question portait sur le transfert de données du Internal Revenue Service à des agences précisées et à des fins précisées, que ces fins soient de statistique ou de conformité, le nombre de personnes favorisant le transfert était sensiblement plus élevé que celui des personnes qui s'y opposaient. Il n'en reste pas moins que, dans chaque cas, une personne sur six environ s'opposait fortement au partage des données.

L'étude de ces données, de même que d'autres variables de l'enquête, se poursuit. Il y a un autre résultat que je puis partager avec vous, qui n'étonnera personne, savoir que les gens qui ont eu des ennuis avec le Internal Revenue Service ou qui ont des opinions défavorables à son sujet sont plus aptes à s'opposer au partage des données.

Nous estimons avoir réussi à brosser un tableau assez clair des opinions actuelles sur le partage entre agences des données. Cependant, les opinions sur des sujets comme celui-ci peuvent changer rapidement, comme on a pu le constater à maintes reprises. Prenons pour exemple le projet de "carte australienne", dont M. Redfern nous a parlé. Par suite d'une campagne bien menée par les opposants à une carte d'identité nationale, l'appui du public pour une proposition en ce sens est passé en neuf mois de 65 % à 39 % (**Washington Post, 1987**).

En résumé, nos données montrent que l'opposition vigoureuse au partage des données, bien que minoritaire, est quand même plus répandue que l'appui vigoureux. En qualité de statisticien, je suis en faveur de l'expansion des utilisations statistiques des données administratives, mais les résultats de cette enquête nous conseillent la prudence. Nous devons continuer à nous renseigner sur l'opinion du public au sujet du partage entre agences des données et de questions semblables, mais je serai le premier à reconnaître qu'il y a de la place pour de l'amélioration dans les questions elles-mêmes.

Nota: Pour les mentions ci-dessus, voir l'étude de T. Jabine et F. Scheuren, intitulée "Statistical Uses of Administrative Records in the United States: Where Are We and Where Are We Going?"

ANNEXE

CERTAINS RÉSULTATS DE L'ENQUÊTE DE 1987 SUR LES ATTITUDES DES CONTRIBUABLES

Maintenant, nous voulons savoir ce qu'il faut faire, à votre avis, lorsque diverses agences du gouvernement désirent le même renseignement à votre sujet. Par exemple, le Internal Revenue Service, la Social Security et le Census Bureau réunissent tous des renseignements sur votre revenu. Ils s'en servent à des fins différentes: percevoir votre impôt, verser des prestations, suivre les tendances du revenu aux États-Unis. Certaines personnes croient que ces agences devraient partager cette information; d'autres estiment au contraire que chacune devrait vous interroger séparément.

(REMETTRE LA CARTE "R")

97. En vous servant de l'échelle indiquée sur cette carte, veuillez me dire votre opinion de chaque énoncé en choisissant une des quatre indications qui vont de "très favorable" à "très défavorable". Veuillez me dire la lettre à gauche de l'énoncé et le chiffre dans l'échelle.

	Très favorable	Favorable	Neutre, sans Opinion (Vol.)	Défavorable	Très défavorable	Incertain
a. Je veux savoir quelles agences possèdent des renseignements à mon sujet	(59(50-1	27-2	16-3	4-4	1-5	3-6
b. Je veux savoir pourquoi chaque agence veut ces renseignements	(60(53-1	29-2	12-3	3-4	1-5	3-6
c. J'ai confiance en certaines agences mais non en d'autres	(61(21-1	27-2	26-3	14-4	6-5	5-6
d. Le partage des données diminuerait pour le public le nombre de formules et de questions	(62(19-1	33-2	20-3	15-4	6-5	7-6
e. Le partage des données permettrait au gouvernements de réduire le coût d'obtention de l'information dont il a besoin	(63(22-1	35-2	17-3	14-4	5-5	7-6
f. Une agence m'inspire autant confiance qu'une autre. Elles font toutes parties du même gouvernement.	(64-13-1	26-2	20-3	23-4	13-5	5-6

n = 2,003

98a. Tout compte fait, quelle est votre opinion du partage de l'information par les diverses agences qui désirent le même renseignement à votre sujet?

Très favorable	(65(14-1
Relativement favorable	24-2
Indécis	20-3
Relativement défavorable	18-4
Très défavorable	23-5
Aucune réponse	1

98b. Pourquoi?

_____ (66-67)

_____ (68-69)

_____ (70-71)

99. Vous m'avez donné votre opinion générale au sujet du partage d'information par diverses agences. Maintenant, j'aimerais vous poser quelques questions précises au sujet du partage de ces renseignements par l'Internal Revenue Service avec d'autres agences. Veuillez me dire ce que vous pensez du partage d'information du Internal Revenue Service dans chaque cas. Êtes-vous très favorable, relativement favorable, relativement opposé ou très opposé au partage d'information du Internal Revenue Service avec (LIRE CHAQUE ARTICLE).

	Très favorable	Relativement favorable	Relativement opposé	Très opposé	Neutre (Vol.)	Incertain
VARIER L'ORDRE — COMMENCER À "X"						
( )a. Le Census Bureau, pour étudier les tendances démographiques	(72(27-1	34-2	11-3	16-4	6-5	6-6
( )b. Le Department of Justice, aux fins de grandes enquêtes criminelles (p. ex., la droque ou le crime organisé)	(73(37-1	28-2	10-3	16-4	4-5	6-6
( )c. Les gouvernements d'état, pour améliorer la perception des impôts d'état	(74(22-1	31-2	14-3	19-4	7-5	8-6
( )d. Le Commerce Department, pour étudier les tendances économiques	(75(23-1	33-2	12-3	16-4	7-5	9-6

INTERVIEWEUR: INDIQUER VOTRE PREMIÈRE QUESTION:

Question a	(76(27-1
Question b	22-2
Question c	24-3
Question d	17-4
Aucune indication	10

100a. Dans quelle mesure pensez-vous que le Internal Revenue Service protège le caractère confidentiel de l'information que vous lui donnez aux fins de votre impôt? Pensez-vous qu'il fait (LIRE ET NOTER)?

Un très bon travail	(77(11-1	
Un bon travail	31-2	
Un travail acceptable	23-3	
Un mauvais travail	12-4	
Incertain	23-5	782

n = 2,003 pour toutes les questions

100b. Pourquoi?

_____ (79-80)
_____ (0*-10-11)
_____ (12-13)

101. La fraude fiscale peut diminuer si certaines agences gouvernementales communiquent au Internal Revenue Service les renseignements qu'elles détiennent. Veuillez me dire, pour chacune des agences ci-dessous, votre opinion du partage de son information avec le Internal Revenue Service. Êtes-vous très favorable, relativement favorable, relativement opposé ou très opposé à ce que le Internal Revenue Service reçoive de l'information des organismes suivants (LIRE CHACUN):

	Très favorable	Relativement favorable	Neutre (Vol.)	Relativement opposé	Très opposé	Incertain
a. Department of Education	(14(19-1	26-2	16-3	14-4	18-5	7-6
b. Social Security	(15(23-1	31-2	12-3	13-4	16-5	5-6
c. Organismes de bien-être	(16(29-1	30-2	10-3	11-4	15-5	6-6
d. Veterans' Administration	(17(23-1	27-2	14-3	12-4	16-5	7-6

102a. Saviez-vous que le Internal Revenue Service perçoit des dettes impayées telles que les paiements de pension alimentaire, les remboursements de prêt aux étudiants ou aux anciens combattants et d'autres, en gardant une partie ou la totalité du remboursement d'impôt?

Oui	(18-42-1
Non	51-2
Incertain	7-3

102b. Êtes-vous favorable ou opposé à ce que le Internal Revenue Service aide d'autres agences gouvernementales à percevoir les dettes impayées de cette façon?

Très favorable	(19(28-1
Relativement favorable	25-2
Indécis	18-3
Relativement opposé	13-4
Très opposé	16-5
Aucune réponse	*

102c. Pourquoi?

_____ (20-21)
_____ (22-23)
_____ (24-25)

* moins de 0.5 %
n = 2,003 pour toutes les questions

Statistique Canada Symposium
sur les utilisations statistiques
des données administratives
novembre 1987

LES PROBLÈMES QUE SOULÈVE, POUR LA PROTECTION DES RENSEIGNEMENTS PERSONNELS, L'EXPLOITATION DES ENREGISTREMENTS ADMINISTRATIFS À DES FINS STATISTIQUES

GUY LABOSSIÈRE[1]

J'aimerais aborder ce matin les problèmes que soulève, pour la protection des renseignements personnels, l'exploitation des enregistrements administratifs à des fins statistiques, et ce du point de vue d'un gestionnaire responsable du contrôle de l'accès et de l'utilisation des enregistrements administratifs dans un organisme statistique.

Les discussions précédentes à ce symposium ont démontré assez clairement qu'au Canada comme ailleurs, les enregistrements administratifs sont de plus en plus une composante importante de nos systèmes statistiques. Nous savons qu'à l'heure actuelle on utilise les enregistrements administratifs directement pour produire des totalisations, indirectement pour établir des estimations, de même que pour remplacer les données d'enquête, pour construire et tenir à jour des bases de sondage et pour évaluer des données. Des exemples d'utilisation des enregistrements administratifs se retrouvent maintenant dans la plupart des programmes statistiques, et tout indique que l'exploitation de ces enregistrements ira en s'intensifiant et en s'élargissant puisqu'on pourra ainsi maintenir les programmes en dépit de la diminution des budgets réels, produire des renseignements allant davantage au fond des choses et alléger ou du moins ne pas augmenter le fardeau de réponse. En résumé, l'utilisation des enregistrements administratifs à des fins statistiques peut présenter des avantages importants, soit l'augmentation de la quantité et de la qualité des renseignements, l'accroissement de l'efficience, la réduction des coûts et la diminution du fardeau imposé aux répondants.

Il va s'en dire que la production de renseignements, en particulier de renseignements statistiques, suppose une collecte de données. Or, une collecte de données constitue nécessairement une intrusion dans la vie privée et est souvent mal perçue par le public à qui on demande de fournir ces données. Dans certains cas, ce ressentiment n'est pas lié directement aux renseignements demandés mais représente plutôt une réaction instinctive de l'enquêté devant ce qu'il perçoit être une érosion de son autonomie ou une violation de son droit à la vie privée. Dans l'absolu, le droit à la vie privée signifie que toute personne a le droit de contrôler l'accès à des renseignements personnels et de décider si ces renseignements peuvent ou non faire l'objet d'une diffusion et si oui, si cette diffusion doit être livre ou limitée. À l'extrême, toute collecte de renseignements est une intrusion dans la vie privée.

Toutefois, d'un point de vue plus réaliste, il est généralement admis que les renseignements nécessaires à l'établissement de politiques et à la prise de décisions constituent des besoins légitimes pour une société et que ces besoins priment sur un droit absolu à la vie privée.

Les lois, notamment la Loi canadienne sur la statistique qui **oblige** les répondants à fournir les renseignements demandés, sont en ce sens l'articulation de l'intérêt public. La Loi canadienne sur la protection des renseignements personnels, qui autorise une

[1] Guy Labossière, Statistique Canada, Parc Tunney, 26ième étage, Édifice R.H. Coats, Ottawa, Ontario. K1A 0T6.

institution fédérale à recueillir des renseignements personnels lorsque ceux-ci ont un lien direct avec ses programmes ou ses activités, représente elle aussi la preuve que les droits d'un particulier ne peuvent l'emporter sur la nécessité de satisfaire aux besoins du plus grand nombre.

Au Canada, contrairement à plusieurs pays de l'Europe de l'ouest où les projets de collecte de données doivent être approuvés par des organismes de protection des données, le pouvoir de créer de nouvelles banques d'informations est délégué aux chefs des institutions publiques, sous réserve d'un examen technique par un centre d'échange rattaché à Statistique Canada.

Par conséquent, un organisme qui décide de recueillir des données sur une nouvelle dimension d'une activité à caractère économique ou sociale, soit par voie d'enquête directe ou par l'exploitation de documents qu'une tierce partie à déjà recueillis à d'autres fins, doit s'en remettre à sa "conscience" pour juger du caractère personnel des renseigements en cause.

Pour survivre, un organisme statistique a tout intérêt à ne pas entreprendre d'activités de collecte qui risquent de compromettre la bonne volonté du public à son égard, même si ces activités n'enfreignent aucune loi sur la protection des renseignements personnels. Pour promouvoir la coopération et le climat de confiance indispensables, il importe au plus haut point de définir clairement certaines conditions et de les mettre au premier plan. Il s'agit notamment:

- de préciser en vertu de quelle loi les données sont recueillies;
- d'obtenir la coopération, voire l'engagement, du répondant en lui expliquant que les données sont recueillies dans l'intérêt public et serviront à telle ou telle fin;
- de s'engager à ne jamais divulguer à des tiers l'information fournie d'une manière qui puisse permettre d'établir un lien entre l'information diffusée et un fournisseur de données;
- de réagir promptement aux plaintes et aux critiques des répondants portant sur la présentation, le niveau de détail, le dédoublement ou le fardeau de réponse.

Dans le cas des enregistrements administratifs, la Loi sur la statistique autorise leur accès et la Loi sur la protection des renseignements personnels permet le transfert de renseignements personnels entre organismes à des fins de recherche ou à des fins statistiques, lorsque le but poursuivi nécessite des renseignements sous une forme identifiable.

Toutefois, lorsqu'on aborde l'acquisition de données de source administrative sous l'angle de la protection des renseignements personnels, on doit tenir compte du fait qu'une telle acquisition suppose l'obtention de renseignements auprès d'un autre organisme qui les a recueillis pour ses propres fins, et l'utilisation de ces renseignements à des fins secondaires à l'insu des personnes qui les ont fournis.

Pour montrer qu'il reconnaît cet aspect particulier de l'exploitation des enregistrements adminsitratifs, et pour dissiper dans une certaine mesure les craintes que des renseignements personnels soient divulgués, Statistique Canada a décidé de dépasser les exigences minimales de la loi en signant des ententes officielles avec les organismes de qui il obtient les enregistrements administratifs. Ces ententes

- préciseront la nature et la portée des renseignements demandés;
- attesteront officiellement que les renseignements demandés ne serviront qu'à des fins statistiques;
- indiqueront clairement les mesures de protection des renseignements et les restrictions sur leur utilisation, et désigneront des responsables pour contrôler les dates d'envoi et la méthodologie, l'utilisation et d'autres questions connexes.

Par ailleurs, il arrive que l'utilisation des données administratives à des fins statistiques ou à des fins de recherche nécessite le couplage de d'enregistrements administratifs à d'autres enregistrements administratifs ou à des données d'enquête portant sur le même répondant ou sur la même unité d'observation. Un tel regroupement de renseignements personnels de diverses sources dans de nouvelles banques de données et un tel couplage d'enregistrements qui à l'origine ont été constitués à des fins distinctes entraînent souvent des craintes, justifiées ou non, au chapitre de la protection de la vie privée. Statistique Canada, après avoir longuement étudié à l'interne l'opportunité de tels couplages et tenu de nombreuses consultations sur le sujet, a adopté une politique qui permet le couplage des enregistrements mais uniquement dans les cas où les avantages sur la plan de l'intérêt public et des **renseignementes statistiques** obtenus l'emportent clairement sur les risques d'intrusion dans la vie privée. Tous les projets de couplage d'enregistrements à Statistique Canada doivent répondre à toutes les conditions suivantes:

- le couplage doit être effectué à des fins statistiques ou de recherche et doit être conforme au mandat de Statistique Canada tel qu'il est énoncé dans la Loi sur la statistique; et
- les produits du couplage doivent être diffusés conformément aux garanties de confidentialité prévues dans la Loi sur la statistique et, le cas échéant, aux exigences applicables de la Loi sur la protection des renseignements personnes; et
- le coulage doit présenter des avantages très nets sur le plan des coûts ou du fardeau de réponse par rapport à d'autres solutions, ou être la seule option réalisable; et
- le couplage ne doit pas servir à des fins qui portent préjudice à des particuliers et doit présenter des avantages qui sont incontestablement dans l'intérêt public; et
- le couplage doit être de nature à ne pas compromettre les programmes futurs de Statistique Canada; et
- le couplage doit passer par un processus établi d'examen et d'approbation.

Le processus d'examen et d'approbation comprend la présentation des projets de couplage dûment documentés par des directeurs de division spécialisée de Statistique Canada à un comité officiel du Bureau. Les recommandations de ce comité sont transmises au statisticien en chef qui, s'il les appuie, les présentera au ministre pour approbation.

Ce mécanisme d'examen et d'approbation par de hatures instances prouve bien que l'on juge important que les projets de couplage soient bien documenté, soient étudiés cas par cas en fonction de critères pratiques et ostensibles dans le cadre d'un processus structuré, et que ce soit le ministre qui, au nom du grand public, décide en définitive si l'intérêt public justifie les risques d'intrusion dans la vie privée.

En résumé, même si Statistique Canada cherche à améliorer plusieurs aspects de son rôle de producteur de renseignements par le biais de l'exploitation d'enregistremnts administratifs, il reconnaît également la nature particulière des ces enregistremnts administratifs du point de vue de la protection de la vie privée et met en place des mesures qui vont au-delà de la stricte conformité aux exigences de la loi afin d'en arriver à ce qu'il juge être un juste équilibre entre la nécessité de protéger le droit à la vie privée et la nécessité de recueillir des données officielles, surtout lorsque la collecte de telles données est obligatoire. Comme vous avez pu le constater, la position actuelle du Bureau sur l'utilisation d'enregistrements adminsitratifs à des fins statistiques est empreinte de précautions et de prudence, même s'il n'existe pas à l'heure actuelle chez le public en général un sentiment hostile à l'endroit de cette utilisation. Dans un tel domaine où il faut concilier des priorités contradictoires, nous croyons qu'il est de toute première importance d'avoir une conduite irréprochable aux yeux du public et d'avoir des positions claires sur tout ce qui touche les besoins en information, les sources d'information, les buts poursuivis et les pratiques de divulgation afin de dissiper les idées fausses et peut-être de rassurer le grand public avant même qu'il ne s'inquiète.

Par ailleurs, il arrive que l'utilisation des données administratives à des fins statistiques ou à des fins de recherche nécessite le couplage de d'enregistrements administratifs à d'autres enregistrements administratifs ou à des données d'enquête portant sur le même répondant ou sur la même unité d'observation. Un tel regroupement de renseignements personnels de diverses sources dans de nouvelles banques de données et un tel couplage d'enregistrements qui à l'origine ont été constitués à des fins distinctes entraînent souvent des craintes, justifiées ou non, au chapitre de la protection de la vie privée. Statistique Canada, après avoir longuement étudié à l'interne l'opportunité de tels couplages et tenu de nombreuses consultations sur le sujet, a adopté une politique qui permet le couplage des enregistrements mais uniquement dans les cas où les avantages sur la plan de l'intérêt public et des **renseignementes statistiques** obtenus l'emportent clairement sur les risques d'intrusion dans la vie privée. Tous les projets de couplage d'enregistrements à Statistique Canada doivent répondre à toutes les conditions suivantes:

- le couplage doit être effectué à des fins statistiques ou de recherche et doit être conforme au mandat de Statistique Canada tel qu'il est énoncé dans la Loi sur la statistique; et
- les produits du couplage doivent être diffusés conformément aux garanties de confidentialité prévues dans la Loi sur la statistique et, le cas échéant, aux exigences applicables de la Loi sur la protection des renseignements personnes; et
- le coulage doit présenter des avantages très nets sur le plan des coûts ou du fardeau de réponse par rapport à d'autres solutions, ou être la seule option réalisable; et
- le couplage ne doit pas servir à des fins qui portent préjudice à des particuliers et doit présenter des avantages qui sont incontestablement dans l'intérêt public; et
- le couplage doit être de nature à ne pas compromettre les programmes futurs de Statistique Canada; et
- le couplage doit passer par un processus établi d'examen et d'approbation.

Le processus d'examen et d'approbation comprend la présentation des projets de couplage dûment documentés par des directeurs de division spécialisée de Statistique Canada à un comité officiel du Bureau. Les recommandations de ce comité sont transmises au statisticien en chef qui, s'il les appuie, les présentera au ministre pour approbation.

Ce mécanisme d'examen et d'approbation par de hatures instances prouve bien que l'on juge important que les projets de couplage soient bien documenté, soient étudiés cas par cas en fonction de critères pratiques et ostensibles dans le cadre d'un processus structuré, et que ce soit le ministre qui, au nom du grand public, décide en définitive si l'intérêt public justifie les risques d'intrusion dans la vie privée.

En résumé, même si Statistique Canada cherche à améliorer plusieurs aspects de son rôle de producteur de renseignements par le biais de l'exploitation d'enregistremnts administratifs, il reconnaît également la nature particulière des ces enregistremnts administratifs du point de vue de la protection de la vie privée et met en place des mesures qui vont au-delà de la stricte conformité aux exigences de la loi afin d'en arriver à ce qu'il juge être un juste équilibre entre la nécessité de protéger le droit à la vie privée et la nécessité de recueillir des données officielles, surtout lorsque la collecte de telles données est obligatoire. Comme vous avez pu le constater, la position actuelle du Bureau sur l'utilisation d'enregistrements adminsitratifs à des fins statistiques est empreinte de précautions et de prudence, même s'il n'existe pas à l'heure actuelle chez le public en général un sentiment hostile à l'endroit de cette utilisation. Dans un tel domaine où il faut concilier des priorités contradictoires, nous croyons qu'il est de toute première importance d'avoir une conduite irréprochable aux yeux du public et d'avoir des positions claires sur tout ce qui touche les besoins en information, les sources d'information, les buts poursuivis et les pratiques de divulgation afin de dissiper les idées fausses et peut-être de rassurer le grand public avant même qu'il ne s'inquiète.

Statistique Canada Symposium
sur les utilisations statistiques
des données administratives
novembre 1987

DÉBATS DE SPÉCIALISTES: CONSIDÉRATIONS RELATIVES À LA PROTECTION DE LA VIE PRIVÉE ET DES RENSEIGNEMENTS PERSONNELS DANS LE CONTEXTE DE L'UTILISATION DES DONNÉES ADMINISTRATIVES À DES FINS STATISTIQUES

J.M. LEYES[1]

En vertu de la **Loi sur la statistique,** Statistique Canada a le droit d'accès aux données administratives et la responsabilité générale de respecter et de protéger les renseignements personnels ou confidentiels contenus dans ses collections de données, y compris les données administratives.

Étant donné qu'il s'agit d'une responsabilité d'ordre général, Statistique Canada a élaboré des procédures et des méthodologies générales visant à protéger les données personnelles et confidentielles auxquelles il a accès. En outre, ce faisant, il a tenu compte des limites imposées par ses objectifs opérationnels.

De toute évidence, l'utilisation des données administratives comporte des avantages et des inconvénients. D'une part, ces données constituent des sources de statistiques. D'autre part, l'accès à ces données par les statisticiens est une violation manifeste de la vie privée des gens. Est-il possible que les procédures ou méthodologies générales utilisées par Statistique Canada pour protéger la vie privée des citoyens et la confidentialité des données laissent à désirer au plan des données administratives? En d'autres termes, faut-il considérer que l'accès à ces données dépasse les procédures et méthodologies générales visant à protéger la vie privée des gens et la confidentialité des données?

Avant la création du Bureau du Commissaire à la protection de la vie privée, Statistique Canada avait énormément de latitude pour définir les notions de vie privée et de confidentialité. Cet avant-midi, le Commissaire à la protection de la vie privée, M. Grace, a, dans son allocution, fait des observations et soulevé des questions qui dénotent que les politiques et les procédures de Statistique Canada, de même que sa connaissance des utilisations des données administratives, sont des points qui relèvent de la compétence du Commissaire. Ce qui signifie, en d'autres termes, que même si nous accomplissons notre travail de statisticiens avec toute l'honnêteté et tout l'altruisme voulus en ce qui a trait au respect de la vie privée et de la confidentialité des données, il y a maintenant quelqu'un qui nous surveille du coin de l'oeil.

Peut-être devrions-nous nous poser deux questions bien précises sur la façon dont nous traitons les données administratives, questions suggérées plutôt qu'abordées par M. Grace:

Statistique Canada a-t-il pris en considération l'adoption d'une politique officielle relative à l'utilisation, à la conservation et au chiffrage des identifiants personnels?
Statistique Canada a-t-il étudié la possibilité de chiffrer toutes les données administratives?

Il existe une autre question qu'il nous appartient, en qualité de statisticiens, d'aborder également:

Statistique Canada a-t-il envisagé de mener une enquête sur une base régulière ou de former des groupes de discussion pour définir les positions des Canadiens quant à l'utilisation des données administratives à des fins statistiques?

[1] John M. Leyes, Statistique Canada, Ottawa (Ontario) Canada.

Il a été particulièrement intéressant d'apprendre de M. Tom Jabine que le **Internal Revenue Service** obtient régulièrement des renseignements sur les attitudes des Américains à ce sujet. Comme nous le savons tous, l'**IRS** a le droit d'accès aux dossiers fiscaux en sa qualité d'agence de recouvrement des impôts aux États-Unis. Pour sa part, Statistique Canada n'est qu'un organisme statistique.

Autant que je sache, Statistique Canada n'a fait aucun effort formel ou systématique pour connaître l'opinion de la population en ce qui a trait à l'accès et à l'utilisation des données administratives à des fins statistiques. En dépit du fait que nous recueillons et diffusons des données sur bien des sujets, il n'existe pas de renseignements systématiques et cohérents relativement aux attitudes des Canadiens à l'égard des utilisations statistiques des données administratives.

En fait, j'estime que notre mode de fonctionnement est généralement basé sur la confiance:

- Nous sommes certains d'être en mesure de protéger la vie privée et la confidentialité des données des répondants.
- Nous prenons pour acquis que nos répondants savent que nous garantissons la confidentialité des données.

Lorsque nous recueillons des données directement, nous pouvons affirmer que nos répondants fournissent ces renseignements en toute connaissance de cause. Il en va tout autrement dans le cas des données administratives. Nous en tirons peut-être des renseignements très précis mais qui n'ont pas été fournis en toute liberté. Certains citoyens canadiens ont peut-être donné des renseignements en supposant que ceux qui sont d'ordre personnel ne serviraient à aucune autre fin, notamment à des fins statistiques.

Bien que nous ayons évidemment presque toute la latitude voulue pour exploiter le potentiel des données administratives à des fins statistiques, nous sommes tenus, parallèlement, de respecter les renseignements confidentiels fournis, grâce à des méthodes qui peuvent sembler inefficaces et trop chères. Tom Jabine a exprimé cette idée d'une façon très concise: la prudence est de mise, en dépit du potentiel statistique des données administratives.

En guise de conclusion, John Grace nous a lancé un défi: Statistique Canada doit obtenir des données de façon légale, c'est-à-dire que les organismes administratifs devraient préciser sur leurs formulaires que les données y figurant seront transmises à Statistique Canada à des fins statistiques.

Le point qui, à mes yeux, revêt un intérêt particulier est que l'observance stricte de la **Loi sur la protection des renseignements personnels** peut restreindre le droit actuel de Statistique Canada d'avoir accès aux données administratives en vertu de la **Loi sur la statistique** et même le priver de ce privilège. Car, si nous nous écartons de la norme relative au respect de la vie privée et de la confidentialité des données si chère aux Canadiens, Statistique Canada pourrait perdre le droit d'accès aux données administratives, et ce, malgré les documents soumis cet avant-midi et dans le contexte d'autres ateliers portant sur le fabuleux potentiel de ces données.

En guise de conclusion, on peut se demander si l'accès aux données administratives en vaut la peine à la lumière de l'éventualité que ce droit soit révoqué et des frais additionnels qu'entraînerait peut-être la protection de la vie privée et des renseignements personnels des Canadiens. Est-il possible qu'une importante partie du système statistique soit menacée du fait que ce dernier s'appuie sur les données administratives pour en tirer des statistiques?

MOT DE LA FIN

MOT DE LA FIN

GORDON BRACKSTONE

"Je pense que ce symposium traitant de l'utilisation statistique des données administratives a été une occasion unique de réunir ensemble plusieurs personnes de disciplines différentes et de divers domaines. Nous avons pris connaissance de l'utilisation de donnnées administratives dans différents secteurs tels la santé et l'éducation, les données socio-économiques, les données reliées aux entreprises et plus encore. Nous avons également entendu parler de méthodes et techniques qui utilisent des données administratives incluant l'appariement d'enregistrements, mais aussi de d'autres méthodes, et nous avons discuté des questions concernant la perception du public, ainsi que du caractère privé et confidentiel de cette utilisation. Je crois, et c'est de même pour plusieurs personnes avec lesquelles j'ai discuté, que cette conférence a donné l'opportunité aux participants de réagir, de prendre conscience de ce qui se fait dans d'autres domaines, et de se faire des contacts, ce que nous espérons sera productif dans le futur.

Filename: Journal-F Proof: 01 Date: 07/02/89 Disk: SM-19
Contact: Frank Mayda Operator: RM/sgl

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

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

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'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, 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 30.00$ par année au Canada, et de 35.00$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit, soit 16.00$ (É.-U.) (20.00$ Can.) est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.