Octobre 1990

SYMPOSIUM 90

Mesure et amélioration de la qualité des données

RECUEIL

Statistique Canada
Statistics Canada
Canada

SYMPOSIUM 90

Mesure et amélioration de la qualité des données

29 au 31 octobre 1990

Ottawa (Ontario) Canada

RECUEIL

Septembre 1991

Comité organisateur du Symposium 90

Mary March
Normand Laniel
Robert Lussier
Jeffrey Smith

Publication autorisée par le ministre responsable de Statistique Canada

PRÉFACE

Le Symposium 90 était le septième de la série des symposiums internationaux qui ont été tenus annuellement à Statistique Canada depuis 1984. Chaque année, le symposium porte sur un thème particulier. Le thème en 1990 était la qualité des données.

En 1990, plus de 350 personnes de plusieurs pays ont assité au symposium. Les participants se sont regroupés pendant 3 jours dans la salle de conférences Simon Goldberg à Ottawa pour écouter les experts en qualité des données provenent de nombreuses agences gouvernementales, des universités, et de l'industrie privée. Au cours du symposium, les participants ont écouté 30 communications invitées réparties dans 10 sessions. Ces sessions ont touché à un large évantail de questions sur la qualité incluant, par exemple, la mesure de la couverture dans les recensements, la mesure et l'amélioration de la qualité des dossiers administratifs, l'amélioration du traitement des données et l'estimation de la variance. Aussi, John Early (Juran Institute) a donné le discours-programme "La gestion de la qualité dans les programmes statistiques nationaux" lors du premier jour et le professeur Carl Särndal (Université de Montréal) a été le conférencier spécial invité avec "Méthodes pour estimer la précision des estimations d'enquête lorsqu'il y a eu imputation" pendant le dernier jour.

À part la traduction et la mise en page, le recueil du Symposium 90 est une copie conforme des articles tels que soumis par les auteurs. L'ordre de présentation des articles est le même que celui du Symposium.

Le sous-comité du Recueil du Symposium 90 désire souligner les contributions de plusieurs personnes qui ont aidé à préparer ce Recueil.

Naturellement, nous remercions les orateurs du Symposium 90 qui ont pris le temps de rédiger leur communication et de la soumettre pour fin d'inclusion dans ce livre. Les efforts de plusieurs autres personnes ont été d'importance vitale pour la publication de ce Recueil. Christine Larabie, Carmen Lacroix et Judy Clarke se sont occuppées du traitement des manuscrits d'une façon experte. La révision de la traduction a été faite par de nombreux méthodologistes et experts du contenu: B. Allard, J. Armstrong, S. Auger, Y. Beaucage, J.-M. Berthelot, R. Boyer, M. Bureau, J. Dumais, S. Giroux, H. Gough, S. Krawchuk, G. Laflamme, D. Lalande, E. Langlet, J. Mayda, S. Michaud, C. Morin, F. Pageau, G. Parent, G. Parsons, E. Rancourt, G. Reinhardt, G. St.-Louis, L. Swain, J. Tremblay, P. Whitridge.

Le huitième symposium annuel et international de Statistique Canada sera tenu à Ottawa les 12, 13 et 14 novembre 1991. Le titre sera "Questions spatiales liées aux statistiques".

Sous-comité du Recueil du Symposium 90
Normand Laniel et Jeffrey Smith
Août 1991

LA SÉRIE DES SYMPOSIUMS DE STATISTIQUE CANADA

1984 - L'analyse des données d'enquête

1985 - Les statistiques sur les petites régions

1986 - Les données manquantes dans les enquêtes

1987 - Les utilisations statistiques des données administratives

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

1989 - L'analyse des données dans le temps

1990 - Mesure et amélioration de la qualité des données

1991 - Questions spatiales liées aux statistiques

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

Pour commander des copies additionelles du recueil du Symposium 90: "Mesure et amélioration de la qualité des données", utilisez le bon de commande sur cette page. Un nombre limité des copies des receuils des symposiums 1987, 1988 et 1989 sont aussi disponibles. Pour commander, envoyez cette formule à l'addresse suivante:

RECUEIL DU SYMPOSIUM 90
STATISTIQUE CANADA
DIVISION DES OPÉRATIONS FINANCIÈRES
ÉDIFICE R.H. COATS, 6e ÉTAGE
PARC TUNNEY
OTTAWA (ONTARIO)
K1A 0T6
CANADA

Veuillez inclure le paiement avec votre commande (chèque ou mandat, en dollars canadiens ou l'équivalent, à l'ordre du "Receveur général du Canada - Recueil du Symposium 90").

RECUEIL DU SYMPOSIUM: NUMÉROS DISPONIBLES

1987 -	Les utilisations statistiques des données administratives - ANGLAIS	_____ @ $10 CHACUN
1987 -	Les utilisations statistiques des données administratives - FRANÇAIS	_____ @ $10 CHACUN
1987 -	ENSEMBLE DE 1 ANGLAIS ET 1 FRANÇAIS	_____ @ $12 L'ENSEMBLE
1988 -	Les répercussions de la technologie de pointe sur les enquêtes - BILINGUE	_____ @ $10 CHACUN
1989 -	L'analyse des données dans le temps - BILINGUE	_____ @ $20 CHACUN
1990 -	Mesure et amélioration de la qualité des données - ANGLAIS	_____ @ $20 CHACUN
1990 -	Mesure et amélioration de la qualité des données - FRANÇAIS	_____ @ $20 CHACUN
1990 -	ENSEMBLE DE 1 ANGLAIS ET 1 FRANÇAIS	_____ @ $35 L'ENSEMBLE
	S.V.P. AJOUTEZ $2 PAR LIVRE POUR LES FRAIS DE LIVRAISON	$ _____
	MONTANT TOTAL DE LA COMMANDE	$ _____

(les prix incluent la TPS; numéro d'enregistrement de la TPS: R121491807)

POINTS CULMINANTS DU RECUEIL 1989: Analyse d'enquêtes à passages répétés (W.A. Fuller), Caractéristiques et problèmes propres aux échantillons successifs (L. Kish), Modèle de série chronologique ajusté pour tenir compte des variations de qualité et servant à l'estimation des indices de prix du logement (D. Pfeffermann, L. Burke, S. Ben-Tuvia), Analyse des modèles ARMMI saisonniers au moyen des données d'enquête (D.A. Binder, J.P. Dick), Estimation des données régionales à l'aide de modèles qui combinent des séries chronologiques et des données transversales (G.H. Choudhry, J.N.K. Rao), Représentation cartographique de données agrégées (D.R. Brillinger), Analyse de séries chronologiques qualitatives en tableaux croisés (A.C Singh, G.R. Roberts), Autres approches d'analyse des éléments des séries chronologiques (W.R. Bell, M.G. Pugh), Ajustement pour tenir compte des déclarations tardives des cas de SIDA et estimation de la population infectée par le VIH aux États-Unis (I.B. MacNeill, Q.P. Duong, V.R. Jandhyala, L. Liu). Quelques méthodes statistiques d'analyse de données historiques personnelles de panel (J.D. Kalbfleisch, J.F. Lawless).

S.V.P. INCLURE VOTRE ADDRESSE POSTALE COMPLÈTE AVEC VOTRE COMMANDE !

NOM _____

ADDRESSE _____

_____

VILLE _____ PROV/ÉTAT _____ PAYS _____

CODE POSTAL _____ TÉLÉPHONE (_____)_____ FAX _____

S.V.P. Notez: Chaque participant au Symposium 90 qui n'est pas un employé de Statistique Canada recevra une copie gratuite du recueil du Symposium 90.

MESURE ET AMÉLIORATION DE LA QUALITÉ DES DONNÉES

TABLE DES MATIÈRES[1]

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

SESSION 7: Amélioration de la qualité des bases de sondage et de leur utilisation
Président: D. Pfeffermann, Hebrew University

SESSION 8: Amélioration du traitement des données et de l'estimation
Président: J. Coombs, Statistique Canada

SESSION 9: Nos produits statistiques sont-ils utilisables?
Président: R. Lussier, Statistique Canada

S.V.P. Notez: La langue originale de tous les articles du Symposium 90 était l'anglais sauf pour les suivants: Mot d'ouverture, I.P. Fellegi (bilingue); Session 4, Y. Leblond (français), Session 5, J.-M. Berthelot (français), Session 9, B.L. Khuong (français).

MOT D'OUVERTURE

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

MOT D'OUVERTURE

I.P. Fellegi[1]

Je vous souhaite la bienvenue au Symposium 90. Le symposium sur la méthodologie est maintenant un événement annuel à Statistique Canada. Depuis 1984, nous avons organisé sept symposiums, portant chacun sur un thème méthodologique différent. Au cours des années antérieures, nous avons traité des sujets suivants:

> l'analyse des données d'enquête;
> les statistiques sur les petites régions;
> les données manquantes dans les enquêtes;
> les utilisations statistiques des données administratives;
> les répercussions de la technologie de pointe sur les enquêtes; enfin, l'analyse des données dans le temps.

Le symposium de cette année porte sur "La mesure et l'amélioration de la qualité des données". En choisissant ce sujet, nous n'avons pas voulu limiter nos débats à l'examen de la qualité des données d'enquête, mais les faire porter sur la qualité de toutes les données recueillies, traitées, analysées et utilisées. Non seulement notre champ d'études comprend-il les données d'enquête ou de recensement, mais il englobe tant les données recueillies à des fins administratives que les observations faites dans le cadre d'une expérience ou les résultats servant à établir un diagnostic (p. ex. la tension artérielle, le rythme cardiaque ou les mesures obtenues au moyen de la machine que le mécanicien branche à votre automobile afin de déterminer l'origine de ses problèmes de fonctionnement).

L'intérêt manifesté mondialement à propos de la qualité témoigne de l'actualité brûlante du sujet. La qualité, de préférence à d'autres caractéristiques comme le prix ou la conception, est devenue un thème clé des slogans publicitaires, comme ceux de Ford "La qualité passe avant tout.", de CP Express "Notre équipe vous apporte la qualité.", ou encore de Toyota "La qualité. Qui pourrait en demander plus?".

Dans quelques jours, on annoncera le nom des gagnants, dans la catégorie qualité, des Prix Canada pour l'excellence en affaires. Cette récompense est maintenant une distinction prestigieuse et recherchée, au même titre que le "Deming Prize" au Japon et les prix "Baldridge" et "NASA Supplier" aux États-Unis.

De fait, l'assurance de la qualité de nos produits et services est un enjeu qui a pris une importance nationale. Lorsqu'il a proclamé le mois d'octobre Mois de la qualité au Canada, le premier ministre Brian Mulroney a déclaré qu'il incombait "à chaque Canadien et Canadienne de mettre l'épaule à la roue pour que le mot Canada devienne synonyme de qualité."

Logiquement, une question se pose: pourquoi ce soudain intérêt dans la qualité? Après la guerre, l'attention s'est portée plutôt sur la facilité de disposer des choses: la courte durée de vie des serviettes de papier, des couches, des objets de plastique, des contenants en verre et en fer blanc, et la durée de vie moyenne que représentent les nouveaux modèles annuels d'automobiles, les changements radicaux de la mode, les maisons de banlieue et les immeubles de bureau destinés à être amortis en 15 à 20 ans. De nos jours, la croissance rapide des revenus disponibles est terminée et on se préoccupe de plus en plus de l'environnement et du développement intégré. Ces facteurs ont donné lieu à une orientation entièrement nouvelle des préférences des consommateurs vers la qualité, la durabilité, la fiabilité et l'économie à long terme.

---

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

En outre, les Japonais ont prouvé que pour être concurrentiel un produit doit être de qualité. Ceci est dû en partie au fait, qui vient d'être mentionné, que les consommateurs insistent de plus en plus sur la qualité. Les Japonais se sont aussi rendus compte qu'il est plus économique, ne serait-ce que sur le plan du rendement, de concevoir un processus de production où la qualité est intégrée au produit que de se fier au contrôle de qualité de la dernière étape.

Afin d'intégrer la qualité au produit, il est essentiel de mesurer les erreurs qui se produisent dans le processus de production et de concevoir un système qui évite ces erreurs. La première de ces tâches est explicitement statistique, bien que, pour pouvoir mesurer les erreurs, le statisticien doit essayer de comprendre la marche du processus. Pour la deuxième tâche, qui est d'éviter les erreurs, les statisticiens doivent collaborer avec le personnel de production, les ingénieurs et les gestionnaires. Le rôle du statisticien n'est pas seulement de concevoir des graphiques sur le contrôle des processus, mais aussi d'aider à obtenir l'équilibre voulu entre les coûts et les erreurs durant les diverses étapes de traitement, c'est-à-dire ce qu'en statistique des enquêtes nous appelons un minimum d'erreurs pour un coût donné.

La question de qualité relève donc fondamentalement de la statistique, même si les statisticiens ne peuvent, seuls, la résoudre.

En tant que statisticiens officiels, nous nous efforçons depuis longtemps d'améliorer la qualité des données dans les deux sens esquissés ci-devant. C'est la volonté d'effectuer des enquêtes où les erreurs (dans le cas présent, les erreurs d'échantillonnage) seraient mesurables, et donc susceptibles d'être éliminées, qui a motivé les premières recherches dans le domaine de l'échantillonnage au hasard. À la suite des premières percées réussies dans ce domaine, nous avons entrepris le long et difficile processus consistant à améliorer les autres aspects de la qualité. Figurent au nombre des principales mesures prises à cet égard: les recherches sur les bases d'échantillonnage et sur les registres en vue d'améliorer la couverture; la mesure et le contrôle des erreurs de couverture; l'élaboration de modèles pour mesurer et expliquer les erreurs imputables à l'intervieweur et en réduire l'incidence; les recherches cognitives pour tenter de comprendre les réactions des répondants face aux questionnaires et aux méthodes de dénombrement, en vue de réduire le nombre d'erreurs de déclaration; l'élaboration de méthodes de contrôle et d'imputation sophistiquées pour réduire les biais dus à la réponse et à la non-réponse; l'élaboration de méthodes d'analyse de données d'enquête permettant d'évaluer des moments d'ordre deux même dans les plans de sondage complexes; et l'élaboration de diverses méthodes d'assurance de la qualité nécessaires pour assurer l'utilisabilité» du produit final.

Les raisons qui nous motivent, comme statisticiens de l'État, à travailler à améliorer et à maintenir la qualité de nos résultats vont au-delà du simple désir d'appliquer nos outils de travail à nos propres activités. Elles tiennent au fait que les informations dont la fiabilité est très douteuse sont pour l'essentiel inutilisables. Cependant, peu d'utilisateurs ont la possibilité de reproduire nos enquêtes ou d'évaluer d'une autre façon la qualité de nos résultats: ils doivent se fier, en dernière analyse, à la réputation de l'organisme qui publie les données. Ainsi, l'utilité de nos produits est directement fonction de notre réputation comme producteurs de données statistiques fiables. Cette réputation revêt donc une importance capitale. Cependant, elle est très difficile à soutenir: la quantité même de nos produits et la multiplicité des utilisateurs nous rendent particulièrement vulnérables aux "incidents de parcours" faisant les manchettes, sans compter les problèmes auxquels nous exposent les utilisateurs qui se servent des données à des fins pour lesquelles elles n'ont pas été conçues. C'est la conscience professionnelle dont nous ferons preuve face à ces problèmes, ou plus précisément en prévision de ceux-ci, qui déterminera si ces rares incidents viendront ternir notre réputation.

Notre capacité et notre empressement à donner une description détaillée de nos méthodes et à faire état de nos connaissances généralement incomplètes des limites de nos données constituent un aspect important de cette conscience professionnelle. De fait, ce sont ces raisons qui sous-tendent la Politique visant à informer les utilisateurs sur la qualité des données et la méthodologie de Statistique Canada, conformément à laquelle toutes les données publiées doivent être accompagnées d'une description des concepts et de la méthodologie utilisés pour les obtenir ainsi que d'une évaluation de leur qualité.

Nous avons abattu un travail énorme et nous pouvons être fiers de nos réussites.

Toutefois, je crois que nous serions tous d'accord pour reconnaître qu'il y a encore beaucoup de travail à faire en vue d'améliorer la qualité ou l'utilisabilité de nos données ainsi que de perfectionner nos méthodes d'évaluation et de déclaration de la qualité. Il est également nécessaire de poursuivre le travail de perfectionnement dans les domaines de la collecte des données, de l'utilisation des dossiers administratifs, du dépouillement et de l'évaluation des données, des bases d'échantillonnage ainsi que du contrôle qualitatif de la collecte et du dépouillement. Il y a aussi lieu de perfectionner nos méthodes de mesure des erreurs de couverture, des erreurs d'échantillonnage et des erreurs matérielles. De même, il faudrait accroître considérablement notre capacité à optimiser l'efficacité générale de nos systèmes d'enquêtes: nos connaissances dans ce domaine sont très rudimentaires. À cet égard, nous devons apprendre à mieux juxtaposer les données de diverses sources afin d'en accroître la valeur et à exploiter les possibilités d'amélioration de la qualité qu'offrent ces systèmes intégrés d'information.

En dernier lieu, il est tout aussi important que nous affinions nos techniques de gestion afin de pouvoir gérer adéquatement la planification de nos méthodes d'amélioration de la qualité et les efforts que nous déployons à cet égard. En cherchant la meilleure façon de gérer nos programmes d'amélioration de la qualité, nous constaterons sans nul doute que nous pouvons beaucoup apprendre du secteur industriel, et peut-être inversement. C'est pour cette raison que les rencontres comme celle-ci sont si importantes.

Les participants à la présente rencontre auront la chance d'assister à des allocutions prononcées par une brochette particulièrement riche et variée de conférenciers venus du Canada, de l'Australie, de la Suède et des États-Unis, parmi lesquels on compte des représentants du secteur public, de l'industrie et du milieu universitaire. À l'évidence, les organisateurs n'ont pas seulement su choisir un sujet à la mode, mais un sujet suscitant un intérêt manifeste et faisant actuellement l'objet de nombreux travaux de recherche. Je souhaite donc la bienvenue à tous les participants et j'espère sincèrement que vous pourrez tirer profit de ce symposium comme, j'en suis certain, nous, à Statistique Canada, saurons le faire.

Avant de terminer, je voudrais adresser des remerciements particuliers au Laboratoire de recherche en statistiques et probabilité de l'Université Carleton et de l'Université d'Ottawa, qui assure le parrainage de cet événement conjointement avec Statistique Canada. Leur assistance matérielle, leurs conseils ainsi que leur soutien moral et financier nous ont été d'une aide inestimable.

DISCOURS - PROGRAMME

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

LA GESTION DE LA QUALITÉ DANS LES PROGRAMMES STATISTIQUES NATIONAUX

J.F. Early[1]

POURQUOI LA QUALITÉ?

Pourquoi sommes-nous ici? Sans vouloir être présomptueux, j'aimerais présenter quatre des nombreux facteurs de motivation possibles. Premièrement, en tant que spécialistes, nous désirons tous assurer et améliorer l'exercice de nos disciplines. Deuxièmement, une pression considérable est exercée sur les organismes statistiques pour que ceux-ci améliorent leur rendement. Troisièmement, la demande pour des données plus nombreuses et de meilleure qualité est en croissance rapide. Enfin, on prend de plus en plus conscience à travers le monde que la qualité a été négligée dans tous les aspects de la société industrielle moderne.

Nous vivons derrière des digues de qualité qui nous protègent d'un déluge de désastres. La qualité nous protège de la fusion du coeur d'un réacteur nucléaire. La qualité nous protège de l'écrasement d'un avion gros porteur. La qualité nous protège d'une panne prolongée de pouvoir électrique sur de grandes étendues.

En ce qui a trait aux statistiques nationales, l'intégrité de nos digues de qualité nous protègent des décisions malavisées ou erronées concernant:

- la réduction (ou l'augmentation) des taux d'intérêt
- les augmentations (ou les diminutions) d'impôts
- la création (ou la suppression) de programmes publics coûteux pour "résoudre" les questions sociales de l'heure dans les domaines
  - -- de la santé
  - -- de l'instruction
  - -- du droit criminel
- l'opportunité de faire (ou de différer) des investissements
- l'embauchage, les licenciements et la formation
- l'opportunité de déménager, de changer d'emploi ou de retourner aux études.

Dans tous ces cas et dans de nombreux autres, de petites et de grandes catastrophes sont évitées grâce en partie à la qualité des données disponibles pour la prise de décision.

Aujourd'hui, je vais parler du rôle de la gestion relativement à la qualité. De façon générale, je crois que nous sommes beaucoup plus en mesure d'assurer la qualité par des moyens techniques que d'améliorer la qualité de façon significative par des pratiques de gestion. Une grande partie de ce que je vais vous dire ne vous semblera peut-être pas nouveau. Ce sont des choses que vous mettez peut-être déjà en pratique, particulièrement au niveau technique. Je vais vous décrire des méthodes de gestion qui doivent être appliquées de façon systématique et intégrale.

---

[1] John F. Early, Vice President, Research and Development, Juran Institute, Inc., 11, River Road, Wilton, CT 06897, U.S., (203) 834-1700.

COMMENT ENVISAGER LA QUALITÉ

Qu'est-ce que la qualité au juste? Une brève définition commune est "l'aptitude à l'usage", ce qui soulève immédiatement la question "À l'usage de qui?". Voici quelques-unes des catégories d'utilisateurs que nous devons prendre en considération.

Les utilisateurs qui traitent les données: ceux qui prennent nos données et les combinent ou les adaptent de différentes façons pour créer de nouveaux renseignements.

Les utilisateurs internes: ceux qui travaillent avec nous sont autant nos clients que les personnes de l'extérieur que nous pouvons considérer comme des clients.

La chaîne de distribution: ce sont les personnes qui forment le réseau de distribution. C'est un concept assez simple à saisir lorsque vous songez à la qualité en rapport avec des céréales pour le petit déjeuner ou des téléviseurs. Vous désirez satisfaire aux exigences de la chaîne de distribution en matière de manutention et de vente afin que celle-ci soit enthousiaste dans la vente de votre produit. Mais il existe aussi des chaînes de distribution pour les statistiques nationales: la presse, les universités, les sociétés professionnelles.

Les utilisateurs types: ceux qui utilisent les données pour des milliers d'applications. Il ne s'agit pas d'un groupe homogène d'utilisateurs, mais d'un ensemble varié et même divisé.

Le grand public: nos données touchent même ceux qui ne les consultent jamais; elles ont une incidence sur le revenu, la santé et la sécurité de l'ensemble de la population.

Dans les disciplines qualitatives, nous avons adopté la convention selon laquelle nous considérons comme les clients tous ceux qui sont visés par un produit. De nombreux organismes qui, traditionnellement, n'appliquent pas leur attention à leurs clients pourront juger ce terme plus ou moins heureux et en adopter un autre. Ce terme précis ne comporte assurément aucun pouvoir magique, mais le fait de considérer toutes les personnes visées par notre travail comme nos clients peut être une expérience très révélatrice.

Il est utile de faire la distinction entre les clients internes et externes afin de s'assurer de bien tenir compte des besoins à ces deux niveaux. Croyez-le ou non, bon nombre d'organismes agissant de bonne foi mais faisant preuve d'un manque de jugement n'ont pas réussi à améliorer la qualité de leurs produits parce qu'ils ont négligé l'un ou l'autre de ces deux groupes. Certains organismes concentrent tellement leur attention sur les besoins de leurs clients internes qu'ils perdent de vue le but premier de leur raison d'être; répondre aux besoins des clients externes. À l'inverse il en a d'autres qui oublient qu'on ne peut s'attendre à ce que nos employés (nos clients internes) servent les clients externes mieux qu'ils ne l'ont été eux-mêmes.

Parmi nos clients externes, nous devons garder en mémoire les besoins des utilisateurs suivants:

- Ceux qui analysent les données en vue de prendre diverses décisions de grande portée.
  - -- Les clients qui effectuent des analyses principalement en vue de l'élaboration de politiques publiques.
  - -- Les clients des milieux universitaires qui peuvent avoir un intérêt politique, mais qui ont en outre de nombreux autres intérêts sur le plan de la recherche en dehors de l'application immédiate.
  - -- Les clients du monde des affaires qui prennent des décisions touchant tous les secteurs de l'économie.
- Certains clients utilisent nos statistiques pour en créer de nouvelles. Par exemple, les estimations du produit national brut reposent sur de nombreuses autres statistiques économiques et la plupart des enquêtes fondées sur la population requièrent les valeurs de contrôles et les bases de sondage des recensements de la population et des enquêtes pertinentes.

- Les sociétés professionnelles s'intéressent non seulement aux estimations actuelles que nous produisons, mais aussi aux méthodes et à la documentation. Les domaines de spécialisation de ces clients comprennent non seulement les sciences statistiques appropriées, mais aussi la santé publique, l'application de la loi, les sciences de l'environnement, l'économie, etc.

- Je mentionne de nouveau le grand public pour souligner l'importante incidence que nos statistiques ont sur des millions de gens qui ne peuvent pas exprimer clairement leurs besoins dans le jargon technique ou politique que nous utilisons la plupart du temps.

Un organisme statistique d'envergure nationale fait face à des problèmes uniques et précis lorsqu'il répond aux besoins de ses clients et l'importance de ces problèmes ne doit pas être minimisée. On procède essentiellement de la même façon avec des clients internes dans un organisme statistique que dans tout autre organisme, mais des questions additionnelles sont soulevées dans le cas des clients externes. La plupart des organismes font face à de nombreux publics qui ont des demandes incompatibles, mais un organisme gouvernemental ne peut pas décider de façon unilatérale de répondre aux besoins d'une partie seulement d'entre eux. Et bien qu'il puisse être indiqué de produire d'autres ensembles de données, il demeure que certains types de données doivent constituer la seule source qui fait autorité. Il ne peut y avoir qu'un compte officiel du recensement. Le taux officiel de l'inflation pour l'ensemble des consommateurs doit correspondre exactement à un nombre. Sur le marché, une société peut segmenter sa clientèle et établir des distinctions entre ses groupes de clients, mais le gouvernement ne peut généralement pas se prévaloir d'une telle option.

Le politicologue David Easton a décrit le système politique comme étant la répartition absolue des valeurs. Les chiffres du recensement déterminent le nombre de personnes admissibles à voter aux élections et la répartition des ressources budgétaires nationales. Les estimations de l'inflation servent à contrôler la clause d'indexation pour les paiements de transfert. Les estimations de la pénétration du marché entraînent des restrictions commerciales de rétorsion.

C'est cette autorité monolithique d'une part et les valeurs importantes en jeu d'autre part qui font qu'il est plus difficile pour un organisme public que pour une société privé axée sur le marché de satisfaire le client.

Les représentants élus constituent une catégorie de clients qui sont particulièrement importants, mais avec qui il est spécialement difficile de faire affaire. Bien qu'ils représentent les besoins du reste de nos clients, on ne peut généralement pas s'attendre à ce qu'ils soient efficaces à cet égard. Cette observation ne doit pas être considérée comme un point de vue élitaire voulant que d'une certaine façon, "nous" connaissions mieux "qu'eux" les besoins des clients. Elle est simplement fondée sur le fait que bien que les représentants élus doivent élaborer et élaborent effectivement bon nombre des politiques générales qui définissent les besoins de nos clients, il y a beaucoup de choses qu'ils n'accompliront jamais à cet égard.

La qualité représente donc exactement ce que tous ces divers clients disent qu'elle est: ce qu'ils estiment convenable pour leur usage.

DEUX TYPES DE QUALITÉ

En fait, cette aptitude à l'usage peut être considérée à deux niveaux: les caractéristiques et l'absence de défauts. Il ne s'agit pas d'une distinction théorique. C'est une différenciation très importante parce qu'elle influe directement sur la façon de gérer la qualité. Les mesures prises pour réduire les défauts sont très différentes de celles qui sont appliquées pour optimiser la qualité des caractéristiques.

Là encore, je dois faire une mise en garde contre les gardiens bien intentionnés, mais mal avisés de la qualité qui semblent toujours poindre à l'horizon. Ceux-ci ont souvent été incapables de distinguer ces deux aspects. En conséquence, leurs solutions, qui font beaucoup d'effet, tendent cependant à être filandreuses, superficielles et décevantes à la longue pour quelqu'un qui recherche réellement un niveau supérieur de qualité.

Les caractéristiques d'un produit permettent de répondre aux besoins des clients. (Et par produit, j'entends tant les biens que les services parce qu'ils sont tous deux produits.) Les caractéristiques que nous offrons sont la

raison pour laquelle les gens s'adressent à nous. Ce sont ces caractéristiques qui répondent à leurs besoins. Sur le marché, la société qui offre la meilleure gamme de caractéristiques se taille une plus grande part du marché et (ou) est en mesure de demander un prix élevé. En ce qui concerne les caractéristiques, un produit d'une qualité supérieure est généralement plus cher. Il est plus dispendieux de voyager en première qu'en classe touriste. Une Jaguar coûte plus cher qu'une Honda.

Certaines des caractéristiques qualitatives des statistiques nationales sont présentées ci-après:

Fréquence: Des estimations mensuelles peuvent s'avérer plus utiles que des estimations mensuelles, mais elles sont également plus coûteuses.

Actualité: Toutes autres choses étant égales, les estimations publiées dans les 20 jours civils qui suivent la date de référence doivent être préférées à celles qui sont diffusées 40 jours plus tard. Toutefois, les estimations plus actuelles occasionnent peut-être une plus grande utilisation des ressources.

Toutefois, il y a ici un lien entre la qualité des caractéristiques et la qualité résultant d'une absence de défauts. Il est possible que nous puissions accélérer l'établissement des données, en partie en diminuant le nombre de défauts que comportent nos opérations de traitement, ce qui permettrait de réduire le temps et les ressources consacrés à corriger ces défauts.

Précision: L'erreur d'échantillonnage est fonction de la taille de l'échantillon et du plan de sondage. En général, plus l'erreur d'échantillonnage est faible, plus les coûts sont élevés.

Exactitude: Pour certains types de biais reliés au plan de sondage, les décisions ont trait à la qualité des caractéristiques. Des compromis explicites selon lesquels la racine carrée de la fluctuation est minimisée et un certain biais est toléré en vue d'obtenir une variance beaucoup moins forte sont des caractéristiques d'un type particulier de statistique. D'autres genres de biais qui résultent d'opérations inefficaces sont des défauts et doivent être traités comme tels.

Concept: Étant donné que nous sommes un organisme "statistique", il est possible que nous devenions accaparés par les points statistiques d'ordre technique et que nous négligions de déterminer en premier lieu si le concept qui doit être mesuré est bien le plus approprié. Par exemple, un indice des prix à la consommation doit mesurer les changements dans les prix à la consommation et ne doit pas inclure les coûts en capital.

Mesure du changement, du niveau, ou données longitudinales: Les clients ont-ils surtout besoin de mesures du changement d'un phénomène ou de mesures du niveau actuel? Ou veulent-ils pouvoir étudier les flux bruts à l'aide de données longitudinales? Ces choix influent considérablement sur nos méthodes de travail.

Documentation: La quantité de documentation fournie aux utilisateurs est une caractéristique, je dirais même une caractéristique très importante, mais elle doit être incluse dans le coût total de l'établissement des statistiques. L'élimination des défauts de la documentation est une autre question. Quel que soit le contenu de la documentation fournie, il doit être exact.

Analyse: Certains des débats concernant la politique nationale en matière de statistiques portent sur la quantité d'analyses à effectuer et par qui. Cette question devrait être étudiée dans le contexte de la satisfaction de nos clients et non de notre propre satisfaction.

Support/mode de livraison: Sous quelle forme le client veut-il ses données: rapports à présentation luxueuse? Immenses tableaux? Brefs sommaires? Sur disquettes? Sur babillard électronique? Demandez au client!

En éliminant les défauts de nos statistiques et de nos services, nous évitons de mécontenter nos clients. Et bien que la satisfaction et l'insatisfaction soient des antonymes sur le plan linguistique, ces concepts ne s'opposent pas directement au niveau opérationnel sur le marché ou dans le cadre des activités d'un organisme statistique.

Un client achète un billet d'avion en première classe parce qu'il veut manger un repas à quatre services dans des assiettes de porcelaine et bénéficier de plus d'espace. Ces exigences satisfont à ses besoins en confort durant ses déplacements. Mais si l'avion a deux heures de retard à cause de "problèmes mécaniques", aucun agrément ne pourra dissiper son insatisfaction fondamentale.

Reprenons l'exemple des automobiles. Nous avons dit que la différence de prix entre une Jaguar et une Honda correspondait à une différence dans la qualité des caractéristiques. Si vous recherchez des banquettes de cuir de première qualité, un tableau de bord recouvert de noyer véritable, un intérieur insonorisé et les regards admiratifs des plaisanciers du club nautique, vous pourrez satisfaire à ses "besoins" en achetant une Jaguar, mais vous serez fort mécontent si elle ne fonctionne pas!

Et maintenant le point clé au sujet de la qualité en tant qu'absence de défaut est qu'une plus grande qualité en ce sens est toujours moins dispendieuse! Si vous faites ce qu'il faut dès le départ, vous n'aurez pas à dépenser vos ressources pour faire des vérifications, apporter des modifications ou apaiser un client furieux (qu'il s'agisse d'un membre du public ou de quelqu'un de votre propre organisme). Les données incorrectes, les réponses fournies en retard, les interprétations inexactes et le traitement discourtois des demandes doivent tous être considérés comme des défauts.

COÛTS D'UNE QUALITÉ MÉDIOCRE

Arrêtez-vous un moment à toutes les activités entreprises par votre agence parce que quelque chose n'a pas été accompli correctement au départ. Pensez aussi à toutes les dépenses engagées parce que le travail comporte vraisemblablement des faiblesses que vous voulez régler avant que le client ne s'en rende compte. Ce sont là les coûts qu'occasionne une piètre qualité. Dans les entreprises manufacturières qui n'ont pas de processus de gestion intégré de la qualité totale émanant de la haute direction, ces coûts représentent 25% ou plus des ventes. Dans les banques, ils correspondent à 33% ou plus des recettes d'exploitation totales. Et dans un programme statistique, les résultats d'une enquête rapide ont révélé que les coûts attribuables à une mauvaise qualité représentent au moins 30% du budget.

Pas dans mon agence, vous empressez-vous de dire? C'est ce que tout le monde dit. Certains de ces coûts ne sont pas facilement visibles parce que le meurtre est commis dans un bureau, mais le corps est souvent découvert dans un autre bureau. Vous finissez habituellement par corriger mes erreurs.

Voici quelques exemples des coûts occasionnés par une qualité médiocre dans un organisme statistique, ces exemples étant répartis selon les trois catégories courantes des coûts imputables à une mauvaise qualité.

Inspection: Les coûts associés à la vérification des données, à la double saisie des données aux fins de la vérification et à tout genre de lecture d'épreuves. Ces coûts pourraient être réduits ou éliminés si nous étions persuadés de l'exactitude du travail initial.

Défauts internes: Les réexécutions informatiques requises à cause d'une mauvaise qualité sont très dispendieuses et prennent du temps.

Lorsque des ressources demeurent inutilisées parce que l'activité précédente présente un retard ou des inexactitudes, cela entraîne des coûts attribuables à la mauvaise qualité. Le personnel sur le terrain attend les échantillons à recueillir, le personnel des opérations informatiques attend les données et les responsables du plan de sondage attendent de connaître les variances des unités.

Les échantillons qui ont des taux de non-réponse élevés ou des rendements moins élevés que prévu pour d'autres occasionnent des coûts liés à la mauvaise qualité.

Défauts externes: La publication de données incorrectes signifie que les statistiques exactes doivent être établies, puis publiées. Il faut répondre aux questions des utilisateurs et donner suite à leurs plaintes.

Il fallait apporter en moyenne environ 10 corrections aux données publiées dans le cadre du programme de l'indice des prix à la consommation du début des années 80 aux États-Unis -- jamais à un élément de première importance au niveau national, presque toujours à des données régionales. Le coût de ces corrections et de l'assistance aux utilisateurs s'élevait à environ un quart de million de dollars. Sur une période de trois ans, la gestion systématique de la qualité a permis d'éliminer complètement les corrections à apporter aux données publiées.

La diffusion tardive des données entraîne habituellement des coûts additionnels -- temps supplémentaire, services de messageries, frais d'impression rapide -- et surtout, de nombreuses autres demandes spéciales des utilisateurs.

La gestion efficace de la qualité de nos caractéristiques peut accroître la satisfaction de nos clients et améliorer nos résultats par rapport à nos crédits budgétaires. Le fait de réduire les défauts contribuera à diminuer les coûts en éliminant le gaspillage d'efforts et le travail à reprendre et en minimisant les coûts entraînés par les plaintes des utilisateurs. Le fait de réduire les défauts contribue aussi à un service plus efficace. La diminution des coûts grâce à la réduction des défauts peut nous aider à mieux gérer les effets du resserrement fiscal et, sur une note plus positive, à élargir notre service aux clients. Nos clients sont plus satisfaits et nos budgets se portent mieux lorsque nous assurons une gestion efficace de la qualité.

UNE STRATÉGIE RELATIVE À LA QUALITÉ

La plupart des gestionnaires ont un aperçu assez juste des principes de base de la gestion financière.

Premièrement, il faut élaborer un plan financier -- un budget des ressources financières qui seront consacrées à des activités données.

Deuxièmement, on exerce une surveillance sur les activités financières à l'aide de ce plan. Des rapports sont établis pour indiquer le niveau des produits et des dépenses selon la catégorie appropriée et au besoin, des mesures sont prises pour rendre les résultats conformes au plan.

Enfin, des projets sont souvent mis en oeuvre pour améliorer les résultats financiers -- par exemple des réductions des coûts liés au classement par article précis.

La plupart des gestionnaires connaissent aussi au moins une des méthodes courantes de gestion de projet -- les méthodes du chemin critique, les graphiques PERT, etc. La gestion de projet comporte les trois mêmes étapes de base.

(1) Planifier le projet.

(2) Utiliser le plan pour contrôler le projet et assurer le respect des délais.

(3) Améliorer les résultats obtenus en créant des groupes de travail pour éliminer les entraves au bon fonctionnement ou pour déterminer les tâches qui peuvent être accomplies en parallèle.

Bref, la plupart des bons gestionnaires savent comment faire de la gestion financière (c'est-à-dire gérer le coût de ce qu'ils font) et comment gérer les activités d'un projet (le temps nécessaire pour les accomplir). Toutefois, la majorité d'entre eux n'ont pas appris la stratégie de la gestion de la qualité. Heureusement, cette stratégie comporte les trois mêmes processus de base: la planification, le contrôle et l'amélioration.

Pour faire la planification de la qualité, il faut d'abord déterminer nos clients et leurs besoins. Nous avons déjà mentionné quelques lignes directrices établies à cette fin dans le cadre d'un programme statistique.

Ces besoins doivent être précisés dans les termes du client, ne vous attendez pas à ce que celui-ci exprime ses besoins comme nous le ferions. Nous devons alors explicitement traduire ces besoins en notre langage.

Nous devons ensuite élaborer les caractéristiques du produit qui répondent aux besoins du client. Par exemple, pour satisfaire à la demande d'un client qui désire avoir des estimations à la fois du niveau et du taux de changement dans la population active, nous pourrions déterminer que la racine carrée de la fluctuation (RCF) des estimations du niveau et que la racine carrée de la fluctuation des estimations du changement sont des caractéristiques essentielles, puis établir des buts précis pour chacune d'elles.

Nous devons ensuite élaborer un processus qui nous permettra d'atteindre ces buts. Des plans de sondage et des estimateurs précis sont établis pour nous permettre d'atteindre nos objectifs relatifs à la RCF du produit. La principale différence entre cette mesure et l'approche traditionnelle pour le plan d'enquête est que celle-ci est intégrée dans le cadre global de planification de la qualité. À une extrémité, le plan est explicitement lié aux besoins des clients. À l'autre extrémité, on détermine que le plan permet d'obtenir le produit avant de le remettre aux opérations.

Selon l'approche traditionnelle, nous avons tendance à escamoter les deux extrémités de ce processus; les besoins des clients sont soit donnés comme probable, soit établis à partir d'un petit groupe de clients. À l'autre extrémité, nous prenons rarement soin de démontrer qu'un processus permet d'obtenir les résultats voulus avant de le soumettre aux opérations. Et tout au long du processus, nous ignorons souvent qu'il serait bon de rechercher une planification qui tient compte de l'avis de tous les participants plutôt que de s'en remettre à la sacro-sainte opinion des "experts".

Pour assurer le contrôle qualitatif, nous devons déterminer les sujets devant faire l'objet du contrôle, c'est-à-dire ce que nous voulons contrôler, et les unités de mesure appropriées. Ces choix sont très importants. Les sujets qui font l'objet du contrôle doivent être des variables qui sont importantes sur le plan de la qualité du processus. En fait, leur sélection fait aussi partie du processus de planification de la qualité. Il existe deux règles pour la sélection des sujets à contrôler. Premièrement, chaque sujet doit être relié directement par le processus aux besoins essentiels des clients. Deuxièmement, il doit être possible de répondre à chaque besoin essentiel d'un client à l'aide d'une caractéristique du produit, qui doit être obtenue au moyen des caractéristiques appropriées du processus, qui constituent à leur tour un sujet de contrôle.

Il est possible de contrôler rigoureusement un paramètre d'importance secondaire sans améliorer grandement la qualité du produit aux yeux du client. Il peut être mentalement et mécaniquement facile de contrôler les erreurs typographiques de la correspondance à l'intention de la direction. C'est pourquoi nous nous appliquons à cette tâche avec ardeur. La question plus importante du contenu de ces messages à nos clients peut toutefois être plus difficile à traiter et, par conséquent, a tendance à être ignorée. Les cadres supérieurs doivent s'assurer que les points importants sont bien maîtrisés.

Les mesures ci-après relèvent du contrôle de qualité type: 1) établir une norme; 2) évaluer les résultats réels; 3) interpréter la différence entre les résultats réels et la norme; et 4) prendre les mesures nécessaires pour ramener le processus à l'intérieur des limites des caractéristiques d'exploitation établies.

L'amélioration de la qualité est le processus qui vise à rendre les choses meilleures. Il s'agit d'abord de déterminer les problèmes relatifs à la qualité les plus urgents et pertinents auxquels l'organisme fait face. Une équipe de projet est ensuite formée pour étudier chaque problème.

Ces équipes doivent mener et (ou) diriger le travail en vue de diagnostiquer les causes du problème observé et élaborer les solutions appropriées qui s'appliquent à ces causes. L'efficacité de ces solutions doit être démontrée, la résistance culturelle au changement doit être vaincue et des contrôles doivent ensuite être établis pour consolider les gains réalisés.

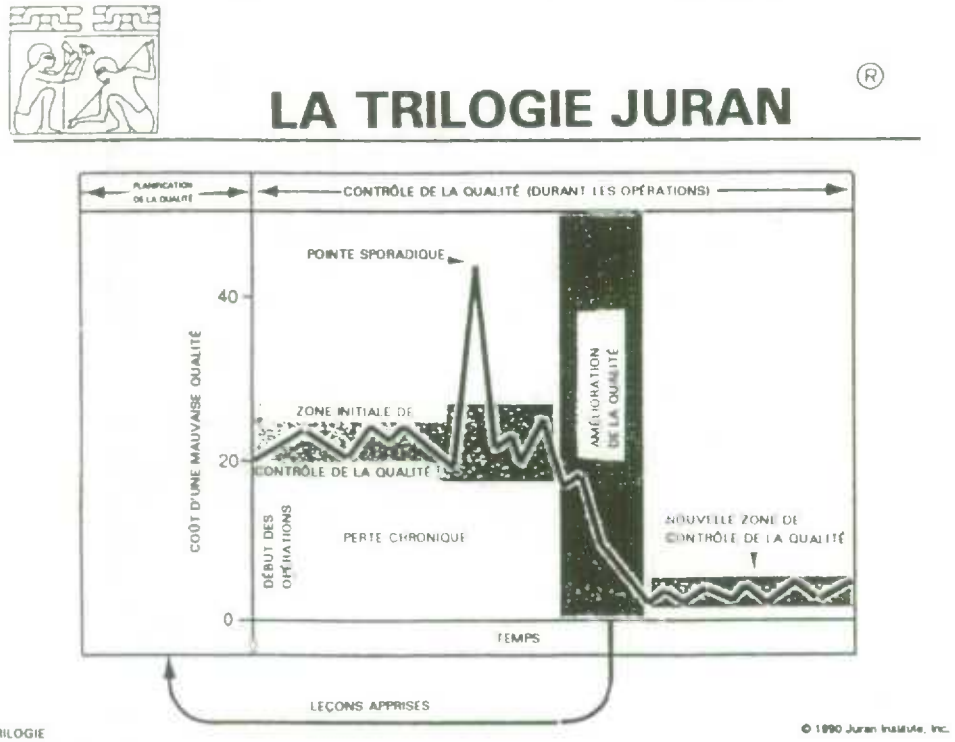
La plupart d'entre nous connaissons la marche à suivre pour diagnostiquer les causes avant d'appliquer un remède. Nous l'appliquons constamment dans le cadre de nos activités techniques. La différence ici est (1) qu'il faut insister pour que la même rigueur soit appliquée aux processus de gestion et aux processus opérationnels, et (2) que les personnes intéressées doivent participer au processus.

De nombreux processus reliés aux opérations de bureau présentent des caractéristiques d'activités de production et constituent des cibles faciles et évidentes pour des projets d'amélioration. Bien entendu, l'amélioration de ces processus a du sens. Je voudrais toutefois recommander fortement que les projets d'amélioration soient également rapidement appliqués aux processus relatifs aux activités des spécialistes et aux processus de conception et ce, pour deux raisons importantes. Premièrement, le personnel de bureau ne doit pas avoir l'impression que l'amélioration de la qualité nous offre simplement une occasion de les harceler. Deuxièmement et plus important encore, les processus relatifs aux activités des spécialistes et les processus techniques seront souvent ceux où les améliorations réalisées seront les plus substantielles.

L'expérience est claire à cet égard. Au moins 80% des problèmes de qualité sont attribuables à la conception du processus. Selon mon expérience, il est également clair que dans les organismes statistiques, la conception des processus pour spécialistes -- statisticiens, économistes, comptables, analystes fonctionnels, etc. -- nécessite autant d'améliorations sur le plan de la qualité et peut être traitée tout aussi efficacement que les opérations plus répétitives. Il faut parfois faire preuve de plus d'imagination et d'habileté en communications interpersonnelles, mais les résultats en valent la peine.

La meilleure façon de commencer votre gestion de la qualité consiste vraiment à appliquer ce processus d'amélioration de la qualité. Les améliorations doivent être réalisées un projet à la fois de sorte qu'il est facile d'amorcer l'opération, et certains résultats peuvent être obtenus au tout début. Pour accélérer l'amélioration de la qualité, il suffit d'augmenter le nombre de projets d'amélioration en cours à un moment donné. Vous constaterez éventuellement que dans certains cas, les améliorations réalisées seront plus substantielles si vous procédez à la nouvelle planification qualitative d'un processus. Vous serez alors en mesure de vous attaquer sérieusement à la planification de la qualité pour de nouvelles activités.

La trilogie Juran des processus de gestion de la qualité



Cette figure illustre les liens entre les éléments importants de la gestion de la qualité que nous avons nommée la trilogie Juran (Juran Trilogy). L'échelle verticale correspond au coût d'une mauvaise qualité, c'est-à-dire au

coût associé à la vérification et à la correction du travail qui n'a pas été effectué correctement la première fois. La figure montre que pour le processus donné en exemple, ce coût représentait initialement environ 20% du coût total.

Ce coût initial de 20% attribuable à une piètre qualité résulte de la planification du processus, la planification étant l'activité à la gauche de l'ordonnée. Lorsqu'une pointe sporadique fait grimper le coût de la mauvaise qualité à 40%, le contrôle de la qualité renvoie rapidement le processus à la zone de contrôle établie. Le contrôle de la qualité ne peut toutefois pas améliorer le processus. Seule une activité d'amélioration de la qualité qui permet de diagnostiquer les causes du coût élevé de la mauvaise qualité et d'élaborer les solutions appropriées peut corriger le processus de façon fondamentale. Grâce aux efforts d'une équipe de projet chargée de l'amélioration de la qualité, une nouvelle zone de contrôle de la qualité est établie aux environs de 3%.

Lorsque nous entreprenons des activités visant à améliorer la qualité, nous apprenons habituellement des leçons importantes concernant le fonctionnement de nos processus et les causes de la mauvaise qualité. Si nous évaluons systématiquement ces leçons, celles-ci forment une base de données précieuse qui peut servir à la planification de la qualité de tout nouveau processus.

LA QUALITÉ ET LA HAUTE DIRECTION

L'état de la qualité des statistiques nationales a fait couler une quantité incroyable d'encre. À ce stade, ma position peut être considérée quelque peu éthnocentrique, mais je dirais que la plupart de ces discussions peuvent être réparties en quatre catégories:

1. Le problème budgétaire. Dépenser plus d'argent.

2. Le problème structurel. Ce point comporte plusieurs variations, qui reflètent habituellement le rôle du narrateur dans une troupe d'acteurs. Certaines comprennent:

   - Une plus grande centralisation de la direction et du contrôle
   - Plus d'autonomie
   - Un accroissement du rôle de la fonction de recherche à l'intérieur de l'organisme et réduction de l'importance des questions relatives aux politiques à court terme.
   - Des liens plus étroits avec les besoins en matière de politiques.

3. Le problème des données.

   Nous avons besoin de plus de données sur les secteurs internationaux. Plutôt, nous avons besoin de plus de micro-données sur les petites régions.

   Nous devons cesser de détruire la continuité chronologique en changeant les méthodes et les classifications. Celles-ci sont complètement dépassées et doivent être modifiées sans plus tarder.

   Nous avons besoin d'indicateurs macro-économiques plus actuels et précis. C'est un point de vue tout à fait irréfléchi, tout le monde sait que nous avons besoin de bases de données longitudinales.

4. Le problème technique.

   Les défauts du recensement décennal américain disparaîtraient si seulement le Census Bureau adoptait les méthodes qui, de l'avis de tout statisticien, sont les meilleures.

   Nous avons besoin d'un fichier commun central sur l'univers des établissements.

Nous avons besoin d'un système d'ITAO.

Au cours des années, je me suis trouvé plongé dans bon nombre de ces débats et d'autres discussions semblables. Et certaines de ces questions sont d'une importance considérable.

Mais aucune d'elles ne sera d'une grande utilité pour l'amélioration de la qualité des statistiques nationales à moins que les cadres supérieurs des organismes statistiques nationaux ne transforment profondément leur mode de gestion. La qualité dépend du leadership de la haute direction à cet égard.

Les cadres supérieurs doivent remplir certaines fonctions précises pour assurer une amélioration substantielle des données. Nous allons jeter un coup d'oeil sur celles-ci dans un moment, mais il importe d'abord de comprendre pourquoi la qualité est une responsabilité qui incombe à la haute direction.

L'ORGANISATION ET LA QUALITÉ

Voyons de plus près le cas du National Statistics Agency for the Federation of Eastern Antartica - ou de tout autre endroit où j'ai travaillé ou que j'ai visité. Cette structure est l'aboutissement de deux siècles et plus d'organisation industrielle. Elle présente deux caractéristiques principales. Le travail acheminé suit un mouvement horizontal de va-et-vient, mais l'autorité et une bonne partie des communications suivent un schéma vertical ou ascendant et descendant. On ne saurait s'étonner qu'il y ait un problème de qualité. Un élément presque toujours caractéristique de tout problème de qualité est que le problème se manifeste dans les espaces entre les diverses unités figurant dans l'organigramme.

Chaque unité organisationnelle est responsable du travail qu'elle accomplit. Les normes de qualité du travail sont souvent fixées par l'unité elle-même. Lorsqu'il existe des normes imposées de l'extérieur, il s'agit le plus souvent de normes "professionnelles" correspondant à la discipline dans laquelle l'unité organisationnelle travaille. Il est rare que ces normes reflètent les besoins de l'unité suivante dans le processus organisationnel et il est encore plus rare qu'elles soient établies en fonction des besoins de l'utilisateur final. La raison en est que personne n'est responsable des zones grises entre les unités organisationnelles. Personne, à l'exception de la haute direction. C'est pourquoi la qualité est d'abord et avant tout la responsabilité de la haute direction.

Afin de régler les problèmes de qualité les plus cruciaux, qui sont presque toujours de portée transversale, la haute direction doit faire office de conseil qualitatif afin d'orienter les efforts. Ses membres doivent identifier les problèmes de qualité les plus cruciaux et nommer des équipes transversales chargées de trouver une solution à chaque problème. Si les dirigeants des principales fonctions organisationnelles ne participent pas tous au conseil qualitatif, leurs employés n'auront pas le soutien requis pour faire partie d'une équipe de projet constituée en vue de trouver des solutions aux problèmes identifiés.

L'expérience nous a révélé que ce n'est qu'avec une approche projet par projet visant à améliorer la qualité que l'on peut espérer faire des progrès réels.

Voici quelques types de projet d'amélioration de la qualité qui ont été couronnés de succès dans les programmes statistiques.

- Réduction du nombre d'erreurs dans les données publiées de 10 par année à zéro. Cela a fait le bonheur des utilisateurs et a permis d'économiser un quart de million de dollars.

- Réduction de 80% du temps nécessaire pour fournir au personnel sur le terrain des clarifications sur les procédures. Cela a contribué à une plus grande satisfaction du personnel et a réduit la nécessité d'effectuer des corrections par la suite.

- Élimination des paiements d'intérêts et de pénalité au titre des avantages sociaux dans le cas des employés ayant quitté leur emploi. Cette "simple question administrative" a drainé des dizaines de milliers de dollars du budget des programmes.

- Réduction du nombre de réexécutions informatiques et des coûts inhérents de plus de 100 000$

- Réduction du temps consacré par le personnel sur le terrain à vérifier les données recueillies et réduction du taux d'erreur grâce à l'adoption de mesures d'auto-vérification et d'instruments peu propices aux erreurs.

- L'effet cumulatif de quelques 2½ années de projets sur le programme de l'indice des prix à la consommation des États-Unis s'est traduit par l'élimination des erreurs dans les données publiées, il a permis d'avancer de 2 jours la diffusion des données (10% plus vite) et de fournir des données accusant 20% moins d'erreur d'échantillonnage, le tout en utilisant 7% moins de ressources.

La planification de la qualité demande également que des équipes transversales soient constituées par les conseils de la haute direction. Il est arrivé trop souvent que la planification ait été reléguée à un groupe "d'experts" - les spécialistes, les statisticiens, le personnel du bureau de planification de programme, etc. Tous ont généralement échoué lamentablement dans leur tâche de planification de la qualité.

Essentiellement et principalement, ils n'ont pas été formés à la planification des multiples paramètres qualitatifs. Ils ne sont pas en mesure d'apprécier le rôle des utilisateurs (internes aussi bien qu'externes) ni la nécessité de déterminer l'efficacité du processus. Deuxièmement, ils ont pour mandat de planifier dans l'absolu, ce qui ne peut être fait avec succès que par une équipe spéciale de composition transversale.

Voici quelques exemples de projets de planification de la qualité pour un organisme statistique:

- Remplacer un système informatique trop ancien.

- Élaborer un processus d'examen et de rajustement professionnel des données de base.

- Élaborer un processus qui servira à sélectionner des échantillons multiples et différents pour des enquêtes similaires.

- Réviser une série statistique existante.

- Introduire une nouvelle série statistique.

- Élaborer un système d'ITAO (interview téléphonique assisté par ordinateur).

- Fournir un panneau d'affichage électronique pour les données courantes.

Le contrôle de la qualité est souvent perçu comme une activité technique plutôt qu'une activité de gestion. Il est certain que le contrôle de la qualité comporte beaucoup d'aspects techniques essentiels, mais une des principales différences entre les organismes qui se contentent de parler de qualité et celles qui font véritablement quelque chose sous ce rapport est que ces dernières considèrent en pratique le contrôle de qualité comme une activité de gestion.

Il y a deux principes de gestion fondamentaux. Nous avons déjà parlé du premier - à savoir que tous les besoins importants des clients doivent être reliés à un sujet de contrôle et que tous les sujets de contrôle doivent être reliés à un besoin essentiel des clients.

Le deuxième principe de gestion important pour le contrôle de la qualité est le concept d'état de l'auto-contrôle. Il ne s'agit pas d'un état psychologique, mais d'un état de gestion à l'intérieur d'un processus qui se fonde sur trois conditions:

1. L'employé sait exactement quel résultat il doit obtenir.

2. L'employé sait avec précision quel est son rendement.

3. L'employé a la possibilité de régler le processus, en d'autres mots:

   -- le processus est en mesure de conduire au résultat attendu,
   -- l'employé a les méthodes, les connaissances et les compétences voulues pour régler le processus,
   -- l'employé a l'autorité nécessaire pour régler le processus.

TÂCHES DES GESTIONNAIRES SUPÉRIEURS

En plus de participer activement au conseil qualitatif afin d'identifier et de sélectionner des équipes de projet chargées d'apporter des améliorations qualitatives précises ou d'élaborer de nouveaux produits ou des produits de remplacement au moyen de la planification de la qualité, les gestionnaires supérieurs ont quelques autres tâches spécifiques qu'ils sont les seuls à pouvoir accomplir.

Ils doivent approuver les objectifs stratégiques de qualité qui font partie intégrante des plans stratégiques de l'organisme. Ces objectifs peuvent inclure notamment les éléments suivants:

- Réduire la fréquence des corrections publiées, selon des quantités définies.
- Réduit l'écart entre la période de référence et la diffusion des données.
- Réduire les coûts liés à une piètre qualité, selon des montants définis.
- Accroître le taux de réponse aux enquêtes, en fonction d'un taux cible défini.
- Améliorer les cotes de satisfaction données par les utilisateurs à nos services.
- Devenir la source privilégiée pour des types particuliers de données, pour certains groupes d'utilisateurs.
- Accroître la portée et la précision de la couverture que la presse accorde à nos données.

Toutefois, ces objectifs stratégiques ne feront que demeurer des vieux pieux si les gestionnaires supérieurs ne prennent pas des mesures précises pour "révéler" ces objectifs aux diverses unités opérationnelles et aux équipes transversales requises.

Même cette répartition des tâches précises visant à atteindre ces objectifs sera sans effet si elle n'est pas suivie d'un examen pertinent des progrès réalisés et d'une reconnaissance du travail bien fait. L'un des moyens les plus efficaces dont les gestionnaires supérieurs disposent pour avoir un effet directeur consiste à donner l'exemple - c'est-à-dire à participer eux-mêmes au projet.

Et, bien sûr, nous savons tous que les gestionnaires supérieurs doivent fournir les ressources nécessaires. Ils doivent libérer les personnes pour qu'elles puissent se consacrer à leur objectif de qualité. Ils doivent fournir des accomodateurs aux nouvelles équipes transversales qui étudient de nouveaux moyens de fonctionner. Ils peuvent avoir à donner à ces équipes le soutien diagnostique nécessaire pour découvrir la cause des problèmes - du temps machine, des avis spécialisés, etc.

Comme il s'agit de révolutionner la façon de fonctionner, tout le monde, y compris le chef de l'organisme, devra recevoir une formation. Les objectifs, les mesures prises et les progrès réalisés devront être communiqués de façon efficace dans toute l'organisation.

Les progrès réalisés doivent être analysés soigneusement par la direction et il faut souligner la contribution de ceux qui sont des moteurs du changement. Étant donné que la qualité est maintenant une priorité absolue, il devra également y avoir des changements au niveau des systèmes d'évaluation au mérite ainsi que des façons par lesquelles les employés sont évalués et récompensés.

CONCLUSION

À part quelques exceptions, la plupart des communications que vous entendrez à la présente conférence porteront sur des méthodes opérationnelles et techniques qui pourraient contribuer à améliorer nos systèmes statistiques nationaux. Une fois ou deux durant la conférence, après avoir entendu une idée particulièrement intéressante, prenez le temps de vous interroger sur les aspects de la gestion de la qualité du sujet traité. Par exemple, vous pourriez vous poser les questions suivantes:

- Quels besoins du client sont visés par la proposition et quelle est l'importance de ces besoins? Comment l'avez-vous déterminée?

- Comment comptez-vous démontrer l'efficacité du processus avant sa mise en application?

- Comment abordez-vous la question des besoins des clients internes qui exploitent, appuient et utilisent les résultats de la proposition?

- Si la proposition vise à corriger un certain défaut qui affecte actuellement la qualité:

  -- S'agit-il d'un défaut d'importance capitale? Le cas échéant, comment le savez-vous?

  -- Le défaut a-t-il été clairement défini?

  -- La cause principale de ce défaut a-t-elle été déterminée au moyen de données et de la participation de tous les secteurs de l'organisme touchés par le processus?

  -- L'idée discutée est-elle censée remédier à cette cause principale? La cause sera-t-elle éliminée ou du moins deviendra-t-elle moins importante?

  -- Ce remède est-il le meilleur possible tant pour l'organisme statistique que pour les clients dans leur ensemble?

  -- Dans quelle mesure y aurait-il de la résistance à ce changement et que proposez-vous pour régler cette question?

  -- Comment saurez-vous si le remède a été efficace et de quelle façon continuerez-vous à vous assurer de son efficacité?

- La proposition est-elle en accord avec le concept d'auto-contrôle?

- La haute direction a-t-elle établi une infrastructure et un système de soutien pour nous aider à répondre à ces questions d'une façon qui s'avère avantageuse tant pour l'organisme que pour ses clients?

- Les cadres supérieurs fournissent-ils les ressources et le leadership en matière de qualité qui nous permettront de faire les changements nécessaires?

Si vous vous posez ce genre de questions, c'est que vous commencez à voir les dimensions de gestion de la qualité. Si vous obtenez des réponses satisfaisantes à la plupart d'entre elles, votre organisme est alors en bonne voie de réaliser son objectif de qualité totale.

SESSION 1

Le défi de réduire les ressources et d'améliorer la qualité des données

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

APPLICATION DE L'APPROCHE GLOBALE DE LA CONCEPTION D'ENQUÊTE POUR DÉTERMINER DES STRATÉGIES DE RÉPARTITION DES RESSOURCES POUR L'ENQUÊTE SUR L'UTILISATION DES VÉHICULES AUTOMOBILES

S. Linacre[1]

RÉSUMÉ

La collecte de données statistiques comporte plusieurs sources d'erreur, et il conviendrait d'adopter des stratégies et de prévoir des ressources qui permettraient, pour un coût fixé, de réduire au minimum l'erreur globale qu'introduisent dans une enquête ces sources d'erreur. Cette communication considère le cas d'une enquête où il y a une assez forte "erreur de réponse" et donne un aperçu de stratégies de collecte qui pourraient être adoptées et des avantages qu'elles offriraient du point de vue de la réduction de l'erreur globale introduite dans l'enquête. L'exposé montre que l'approche globale de la conception d'enquête suppose des ressources qui rendent possible ces stratégies de collecte.

MOTS CLÉS: Approche globale de la conception d'enquête; erreur de déclaration; enquête avec journal.

1. INTRODUCTION

La collecte de statistiques comporte des risques d'erreurs de sources diverses. La tâche du responsable de la conception d'une enquête consiste à réduire au minimum, pour un coût déterminé et compte tenu des exigences précises fixées au sujet des produits de l'enquête, l'erreur totale introduite dans les résultats. D'un autre point de vue, le responsable peut considérer qu'il a pour tâche de concevoir une enquête à un coût minimal en vue de fournir les produits demandés sans dépasser l'erreur totale maximale permise. Dans un cas comme dans l'autre, l'erreur totale comprend non seulement l'erreur d'échantillonnage, mais aussi toutes les autres sources d'erreurs qui entraînent un écart entre l'estimation et le concept que l'utilisateur veut mesurer.

Pour atteindre un tel objectif, le concepteur doit tenir compte des aspects suivants: l'objet de l'enquête, la façon dont les données seront utilisées, les types d'erreur qui peuvent influer sur ce genre d'utilisations, les sources de ces erreurs et la façon dont ces dernières peuvent être corrigées dans le contexte opérationnel et les méthodes les plus économiques susceptibles de réduire les erreurs. En faisant cela, le responsable adopte une "approche globale de la conception d'enquête". En revanche, le fait de tenir compte de façon indépendante d'erreurs comme l'erreur de couverture ou l'erreur d'échantillonnage entraîne généralement une utilisation inefficace des ressources affectées au plan de sondage dans son ensemble.

Comme dans le cas de tout processus dont on tente de gérer la qualité, le concepteur d'une enquête à caractère permanent doit suivre un cycle précis d'évaluation et d'amélioration. Tout d'abord, le processus doit être examiné dans son ensemble et il faut se servir des résultats de discussions, de comptes rendus, d'enquêtes pilotes et d'études d'évaluation pour cerner les principales sources d'erreurs. Ensuite, il faut entreprendre des travaux pour obtenir des données quantitatives sur ces sources d'erreurs et élaborer des méthodes permettant de les réduire. Enfin, d'autres études d'évaluation doivent être réalisées pour obtenir des données quantitatives sur l'efficacité de ces méthodes, et le cycle reprend à nouveau.

---

[1] S. Linacre, Australian Bureau of Statistics, C.P. 10, Belconnen, ACT, 2616, Australia.

Le présent document a pour objet de faire ressortir pour toute enquête donnée l'importance d'une approche de conception axée sur l'ensemble du processus de collecte plutôt que sur une série de sous-procédés discontinus.

2. ENQUÊTE SUR L'UTILISATION DES VÉHICULES AUTOMOBILES DESCRIPTION GÉNÉRALE

L'enquête sur l'utilisation des véhicules automobiles (l'EUVH) du Australian Bureau of Statistics (ABS) fournit des données sur les caractéristiques et sur l'utilisation qui est faite des véhicules automobiles privés et commerciaux en Australie. Des données sont notamment recueillies au sujet des variables suivantes:

- type et caractéristiques du véhicule
- distance parcourue
- but de l'utilisation
- secteur d'utilisation du véhicule
- marchandises transportées
- nombre estimé de tonnes transportées au kilomètre
- quantité de combustible utilisé
- caractéristiques du chauffeur
- voyageurs transportés (dans le cas des autobus).

Cette enquête a été menée à tous les trois ans depuis 1976, la plus récente ayant eu lieu en 1988. Les principales utilisations des données sont:

- servir de base au calcul de répartition des fonds du Commonwealth entre les gouvernements des États pour le financement des routes;
- aider à définir les modalités de recouvrement des coûts dans les États, en particulier pour ce qui est des frais d'immatriculation et des taxes;
- faciliter le calcul des taux de risques d'accidents de la route;
- produire des estimations de l'utilisation des installations de transport routier par l'industrie;
- fournir des données repères pour des études de la politique en matière de transport routier, surtout relativement à l'utilisation des véhicules lourds et des habitudes de consommation de combustible;
- aider à contrôler la pollution de l'air et la pollution par le bruit; et
- faciliter le contrôle du transport des matières dangereuses.

L'enquête s'applique à tous les véhicules immatriculés autres que les roulottes, les semi-remorques et les tracteurs. La base de sondage est établie à partir des divers fichiers d'immatriculation des États et des territoires, lesquels sont stratifiés selon le type de véhicule, la masse et l'année de fabrication. Pour l'enquête, on a recours à un échantillonnage stratifié simple à un degré et l'unité d'échantillonnage est le véhicule automobile. L'enquête permet aussi de recueillir des données sur les types de véhicules et des estimations sont produites à partir des déclarations faites à ce sujet par les répondants. La distribution d'échantillonnage est considérée comme une distribution à plusieurs variables, les niveaux d'exactitude sont précisés par type de véhicule par état et les niveaux prévus d'erreur de stratification sont pris en compte dans cette distribution. Un questionnaire est envoyé par la poste aux propriétaires de certains véhicules (environ 68,000 au total) qui doivent fournir de mémoire des données portant sur les douze derniers mois. La période de référence va du mois d'octobre de l'année précédant l'envoi du questionnaire au moins de septembre de l'année en cours.

Du point de vue de la qualité, les principaux domaines de préoccupation pour cette enquête sont: la difficulté pour les répondants de se rappeler les données; une stratification inefficace ayant pour cause la piètre qualité des données complémentaires contenues dans les bases de sondage du registre; et les problèmes de traitement attribuables en grande partie aux difficultés qu'éprouvent les répondants à fournir l'information demandée.

3. SOURCES D'ERREURS

Les principales sources d'erreurs associées à la collecte des données sont, dans le cas de l'EUVH, les suivantes:

Erreur conceptuelle. Les éléments d'information recueillis dans le cadre de l'EUVH sont relativement simples et il n'y a pas de problèmes au niveau des concepts utilisés. Les données sont utilisées par les responsables du transport dans les États comme mesures repères de l'utilisation des routes et les principaux points d'intérêt sont la distance parcourue et les charges transportées. La signification de ces éléments d'information ne pose aucun problème pour les responsables du transport et ils sont également assez bien compris par les répondants. Par conséquent, les points d'information nécessaires à l'EUVH sont relativement simples et aucun des problèmes de définition, concepts, etc.,fréquents dans le cas de nombreuses enquêtes-ménages, n'est associé à l'enquête.

Définition de la population. La population de l'EUVH est aussi relativement bien définie et la liste de ses unités est fournie dans les fichiers d'immatriculation des véhicules. Cependant, selon l'utilisation qui en est faite pour les besoins de l'EUVH, ces fichiers ne fournissent qu'une vue d'ensemble des activités de toute une année. Il y a donc le problème des véhicules qui cessent d'être immatriculés au cours de l'année ou qui changent de propriétaire. Ce ne sont pas tant les questions de sous-dénombrement ou de surdénombrement qui importent, mais plutôt le problème de la qualité des renseignements complémentaires qu'on trouve dans ces fichiers et qui sont utilisés dans l'élaboration du plan de sondage.

Les détails sur les types de véhicules contenus dans les fichiers sont souvent erronés et donnent lieu à des erreurs de stratification des véhicules. Comme les estimations sont produites par type de véhicule, ces dernières doivent plutôt être fondées sur les déclarations relatives aux types de véhicules, sources d'erreurs-types nettement plus importantes que si les renseignements contenus dans la base étaient plus précis. On estime qu'avec des données de très haute qualité sur les types de véhicules, la taille des échantillons pourrait être réduite d'environ 20 à 30% et l'importance des erreurs-types associées aux estimations ne serait pas plus grande qu'elle ne l'est présentement.

Erreur d'échantillonnage. Comme c'est toujours le cas, l'erreur d'échantillonnage dont sont entachées les estimations de l'EUVH pourrait être moins importante si on augmentait la taille de l'échantillon. Nous pourrions obtenir le même effet encore une fois grâce à l'amélioration de la qualité des renseignements complémentaires contenus dans les fichiers d'immatriculation et du même coup de l'efficience de la stratification des véhicules.

Erreur de déclaration. Une des grandes préoccupations liées à l'enquête est précisément l'importance de l'erreur de déclaration associée à un certain nombre des questions de l'EUVH. En effet, un grand nombre de ces questions font appel à la mémoire des répondants concernant des données pour les douze derniers mois; c'est le cas des questions sur la distance parcourue, sur la région du déplacement, sur les niveaux de consommation de carburant ainsi que sur les passagers et les marchandises transportés. Selon les résultats d'enquêtes pilotes, ces données n'ont pas toujours été consignées par écrit par les répondants et certaines estimations sont nécessaires.

Erreur de dépouillement. Comme l'EUVH n'est menée qu'à tous les trois ans, il faut former du nouveau personnel affecté au dépouillement à chacun des cycles de collecte. Lors de l'enquête de 1988, 15 des 57 agents du centre de dépouillement étaient des employés temporaires, le reste du personnel venant d'autres divisions ou services du ABS ou de la fonction publique. Très peu de ces employés connaissaient les concepts de l'EUVH avant le dépouillement et ils n'avaient suivi qu'une semaine de formation. Le dépouillement exige que de nombreux changements soient apportés aux données déclarées par les répondants et que l'on ait largement recours à l'imputation des données manquantes pour les répondants qui n'ont pas été en mesure de fournir les chiffres demandés. Les risques d'erreurs de dépouillement sont par conséquent très grandes. Dans le cas de l'EUVH de 1988, menée à Perth, des problèmes de communication se sont également posés en raison de l'éloignement.

Erreur due à la non-réponse. Un taux de non-réponse de 15% est associé à l'enquête, mais il serait plus élevé dans le cas de l'imputation pour non-réponse aux questions. Au moment de la production des estimations finales, les personnes n'ayant pas du tout répondu à l'enquête font l'objet d'une imputation qui correspond à la moyenne de leur strate, laquelle est définie selon le type de véhicule, le nombre d'années d'utilisation et l'État. Nous n'avons pas encore entrepris une analyse du biais attribuable à la non-réponse.

En résumé, dans le cas de l'enquête sur l'utilisation des véhicules automobiles, les principales sources d'erreurs sont: le manque de précision des renseignements complémentaires dont on se sert pour l'élaboration du plan de sondage, des problèmes d'erreurs de déclaration liées au fait que les répondants doivent faire appel à leur mémoire sur une période de douze mois et le problème des erreurs de dépouillement et autres problèmes de traitement causées par le manque d'expérience du personnel et l'éloignement du centre de dépouillement. Les effets des problèmes liés à la base de sondage sont assez faciles à mesurer. Un certain nombre d'études ont été réalisées pour aider à mieux faire comprendre l'importance et les conséquences relatives des erreurs de déclaration et celles-ci sont décrites brièvement dans la section qui suit.

4. ÉTUDES ENTREPRISES POUR MESURER L'IMPORTANCE DES PROBLÈMES LIÉS AUX ERREURS DE DÉCLARATION

4.1 Enquête pilote

Les questionnaires de l'EUVH de 1988 ont fait l'objet de nombreux essais. Dans le cadre de cette enquête pilote, les répondants devaient remplir le questionnaire et une interview de suivi a ensuite été menée. L'interview portait sur la méthode utilisée par le répondant pour l'aider à remplir le questionnaire. Pour certaines des questions, les répondants devaient en plus indiquer s'ils jugeaient leur réponse très précise, assez précise, approximative ou encore s'ils n'avaient aucune idée de sa précision. Le tableau 1 ci-dessous donne un aperçu des résultats du test.

Comme le montrent les résultats du test, il est clair que les répondants éprouvent certains problèmes à se rappeler des données se rapportant aux douze derniers mois. Évidemment, le problème est plus ou moins grave selon l'interprétation que l'on donne du terme "assez précise". Il convient tout particulièrement de remarquer le faible pourcentage de répondants qui avaient accès à des dossiers.

4.2 Analyse des opérations de traitement et de dépouillement

De nombreux renseignements utiles à la gestion ont pu être tirés des opérations de traitement des données de l'EUVH. La collecte et l'analyse d'une telle information avaient deux objectifs. Premièrement, on cherchait à déterminer l'efficacité des opérations qui pour la première fois avaient été centralisées, dans un centre de dépouillement précis, et confiées à du personnel devant être familier à la fois avec l'entrée des données et leur vérification manuelle. Deuxièmement, compte tenu en particulier des problèmes connus liés à la déclaration, on voulait évaluer l'effet de ces problèmes sur la réponse partielle ainsi que sur les taux de vérification et d'imputation. Les résultats pouvaient être communiqués au besoin aux personnes concernées afin que les changements appropriés puissent être apportés sur le plan des méthodes et de la conception des questionnaires.

Tableau 1: Pourcentage de participants à l'enquête pilote ayant classé leur réponse comme très précise ou assez précise; pourcentage de répondants ayant trouvé leurs réponses dans des dossiers

	Distance parcourue	Objet du déplacement[1]	Proportion de répondants s'étant déplacés à l'intérieur des régions définies[2]
Véhicules privés			
Très précise	42%	36%	17%
Assez précise	36%	48%	70%
Utilisation de dossiers	30%	12%	20%
Véhicules commerciaux			
Très précise	36%	62%	41%
Assez précise	57%	31%	47%
Utilisation de dossiers	32%	30%	15%
Autobus			
Très précise	62%	69%	58%
Assez précise	35%	18%	37%
Utilisation de dossiers	42%	21%	38%

1 En ce qui concerne l'objet du déplacement, les catégories étaient: voyage d'affaires - entreprise, professionnel ou secteur public; déplacement entre le domicile et le lieu de travail; raisons personnelles; et autres.

2 Les régions étaient définies de la manière suivante: principale ville et environs; autre à l'intérieur de l'État; autre État ou Territoire (à préciser).

Parmi les renseignements tirés de ces opérations, certains sont particulièrement intéressants. En effet, au cours du traitement des 67,000 questionnaires, les répondants ont adressé quelque 45,000 appels téléphoniques au centre de dépouillement et les employés du centre ont fait 50,000 autres appels, malgré le fait qu'il s'agit d'une enquête où la collecte se fait par la poste avec rappels également envoyés par la poste. On pense que le volume élevé de contacts avec les répondants pour cette enquête serait attribuable en grande partie au problème de disponibilité des données, dans le cas des demandes de renseignements provenant tant des répondants que du centre de dépouillement. Le taux de réponse global et final de l'enquête s'établissait à 85.0%.

Pour analyser l'importance de la vérification et de la reprise de contact avec les répondants ainsi que les effets du traitement sur les données finales, on a sélectionné un sous-échantillon de 551 véhicules. Pour chaque véhicule sélectionné, on devait inscrire la valeur originale déclarée, tout changement apporté par la suite à cette valeur et la justification d'un tel changement (contact avec le répondant, accès à des manuels de véhicules, etc.). En résumé, selon les résultats fondés sur cet échantillon, il a été estimé qu'une modification quelconque a été apportée à 91% des questionnaires, pour une moyenne de 6.1 changements par questionnaire vérifié et de 5.6 changements par questionnaire retourné. Les changements ont été faits après contact auprès des répondants dans 38% des cas, d'après des renseignements contenus dans des manuels de référence dans 10% des cas et d'après d'autres informations (habituellement sur les questionnaires) dans 52% des cas. Sur les changements apportés, 66% découlaient d'une réponse initiale laissée en blanc, les autres 34% étant des changements à une valeur déclarée. Bien qu'une modification à une réponse laissée en blanc à l'origine entraîne inévitablement une hausse de l'estimation pour cette enquête, il est intéressant de constater l'effet de pondération global des changements apportés à des questions non laissées en blanc étant donné que presque toutes ces questions avaient une valeur négative.

Si nous supposons que la non-réponse initiale à une question signifie que le répondant ne disposait pas des données nécessaires, certaines questions peuvent être particulièrement sensibles à ce genre de problème. Pour les questions comme celles portant sur la distance parcourue selon l'objet et la région du déplacement, entre 15 et 30% de l'estimation provenait de valeurs corrigées ou imputées. En revanche, le pourcentage de la variable "distance totale parcourue" calculée à partir de questionnaires vérifiés (de façon générale les questions avaient à l'origine été laissées en blanc) est négligeable, soit 3%.

4.3 Enquête avec odomètre (EO)

S'il est vrai que l'enquête pilote comme les résultats des vérifications effectuées en cours de traitement faisaient ressortir un taux relativement élevé d'erreurs de réponse, il n'avait pas clairement été démontré si de telles erreurs pouvaient entraîner une plus grande variabilité des estimations ou encore un biais important.

Tableau 2

	Voitures	Camions rigides	Camions articulés	Total
$D = \frac{EUVH^1 - EO^1}{EO^1}$	14.0%	10.3%	5.8%	13.15%
IC de 95% pour D (intervalle de confiance)	(8.9%, 19.1%)	(3.9%, 16.6%)	(1.3%, 10.4%)	(9.0%, 18.0%)
CV pour l'EUVH (coefficient de variation)	.96	1.63	.82	
CV pour l'EO	.77	1.06	.92	
Fraction de l'échantillon de l'EUVH nécessaire pour que l'EO donne la même ETR (erreur type relative)	.64	.42	1.26	.79
Échantillon de l'EUVH de 1988	12896	20469	20014	53379
ETR pour l'EUVH	1%	2%	1%	1%

[1] EUVH représente l'estimation de la distance totale parcourue fondée sur les données de l'EUVH. EO représente l'estimation correspondante fondée sur les résultats de l'enquête avec odomètre.

L'enquête avec odomètre a été conçue dans le but de fournir une mesure quantitative de l'effet de l'erreur de déclaration dans le cas d'une variable importante, la distance totale parcourue. L'enquête pilote et l'analyse des questionnaires traités révèlent que les composantes de la variable "distance parcourue" sont davantage sujettes aux erreurs de déclaration qu'en ce qui concerne la distance parcourue en soi, mais l'enquête avec odomètre a pour avantage de fournir une mesure précise de la distance totale parcourue, de façon assez simple.

La méthodologie de l'enquête pilote avec odomètre reposait sur la sélection d'un échantillon de véhicules dans un État (Nouvelles-Galles du Sud), en parallèle avec l'échantillon choisi pour l'EUVH. Les répondants de

l'échantillon parallèle, plutôt que d'avoir à se rappeler les données pour l'ensemble de la période, ont été contactés au début d'une autre période correspondante afin qu'ils fassent une première lecture de leur odomètre, laquelle a été suivie d'une seconde lecture à la fin de la période de référence de douze mois. Les résultats de l'enquête avec odomètre et de l'enquête mémoire ont été comparés.

Le tableau 2 fait nettement ressortir deux choses. Premièrement, il y a un écart positif marqué et constant entre la distance mesurée selon l'EUVH et celle fondée sur l'enquête avec odomètre. La valeur la plus élevée est observée dans le cas des voitures et la moins élevée pour les camions articulés. (Dans ce dernier cas, il est possible que la valeur moins élevée soit attribuable à une meilleure tenue de dossiers pour les propriétaires de ce groupe, mais après examen des résultats, on s'est rendu compte que le nombre de ces véhicules est proportionnellement plus élevé dans la catégorie des 10,000 kilomètres et moins de l'EUVH, comparativement à l'enquête avec odomètre, les exploitants ayant pu vouloir échapper ainsi à la réglementation assez stricte en vigueur dans l'industrie concernant la distance parcourue et les heures de déplacement.) Dans tous les cas, le biais à la hausse est très important par rapport à l'erreur type fixée pour l'enquête. Deuxièmement, les chiffres témoignent d'une variance de réponse significative. La variabilité des données de l'EUVH est beaucoup plus grande que celle des données de l'enquête avec odomètre, sauf dans le cas des camions articulés. La conclusion qu'on peut en tirer est qu'une estimation de la distance parcourue peut être obtenue avec un échantillon beaucoup plus petit lorsque l'enquête s'appuie sur l'utilisation d'un odomètre, l'erreur type relative restant la même que pour l'EUVH. Le tableau 1 indiquait également la réduction de la taille de l'échantillon pour chaque catégorie de véhicules.

Les fonctions de distribution de probabilité empirique des distances parcourues de l'EUVH et de l'enquête avec odomètre ont fait l'objet d'une comparaison (voir les figures 1, 2 et 3), laquelle montre que les résultats de l'enquête avec odomètre donnent une courbe beaucoup plus lisse et que les valeurs sont beaucoup plus concentrées que dans le cas de l'EUVH.

De toute évidence, les répondants de l'EUVH ont largement eu recours à l'arrondissement puisque les données présentent des pointes à tous les 5,000 kilomètres environ: 34% des données déclarées de l'EUVH correspondent à des unités exactes de 5,000 milles ou kilomètres, comparativement à 0.44% des données de l'enquête avec odomètre! Un autre signe de l'absence de fiabilité de certaines des données accessibles aux répondants.

Figure 1. Fonction de distribution cumulative: voitures, distance parcourue

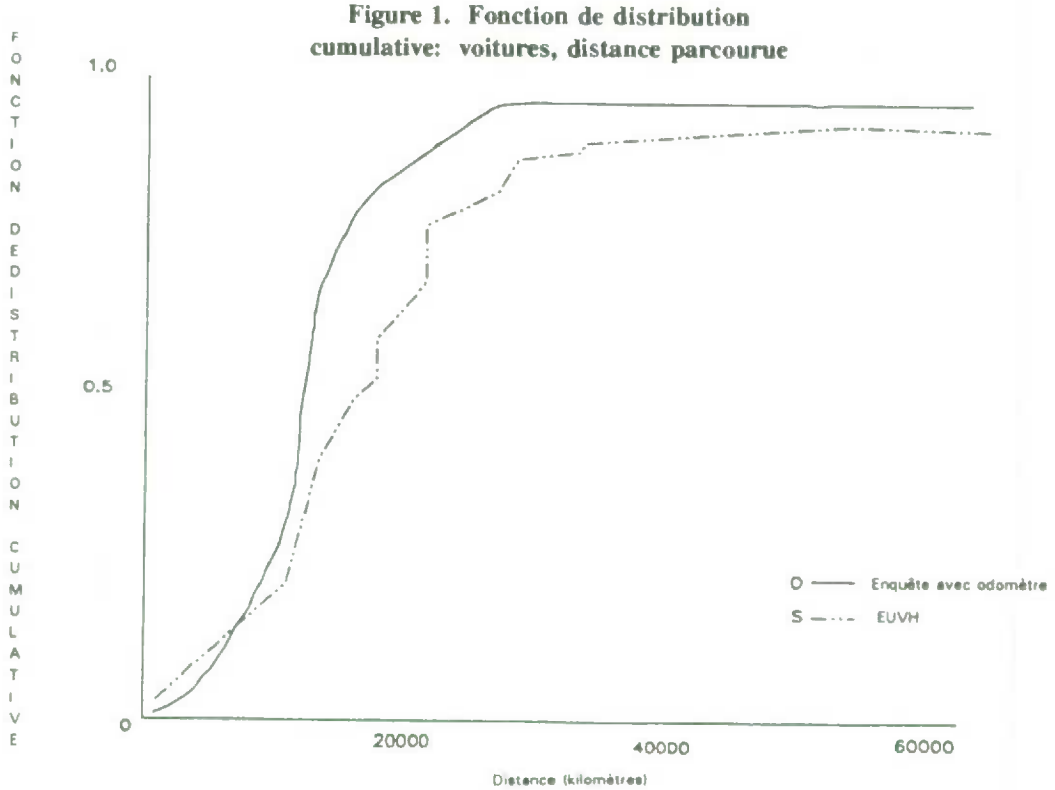
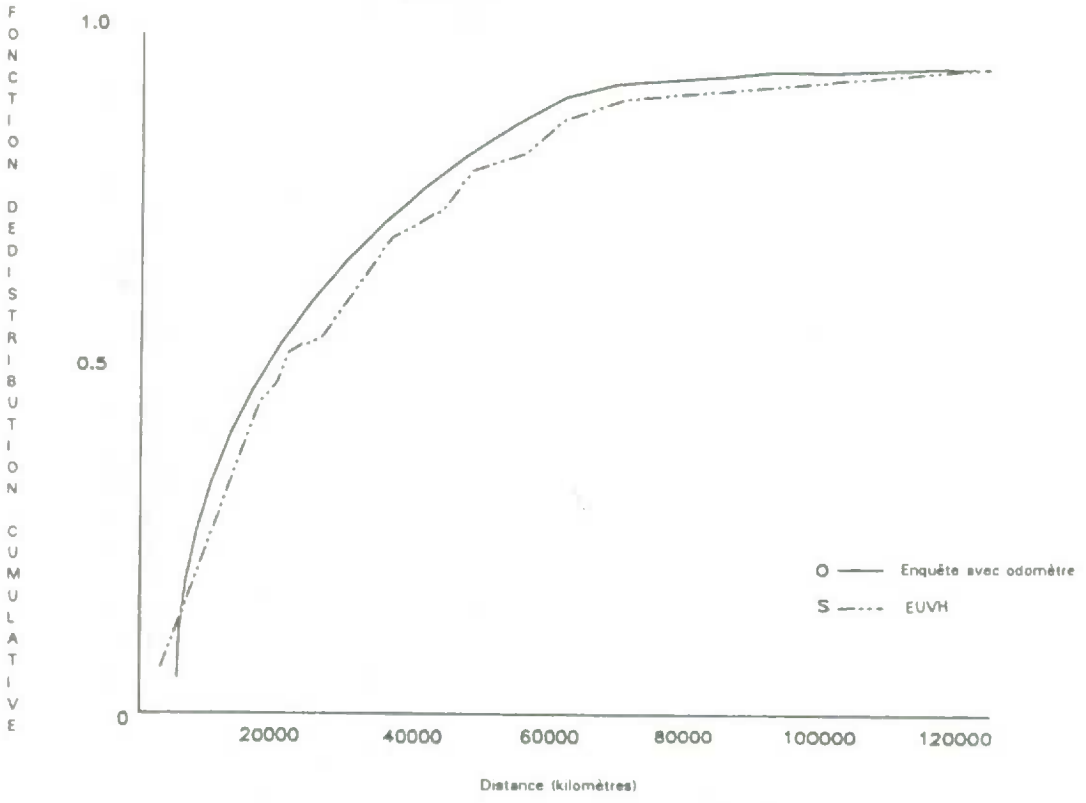
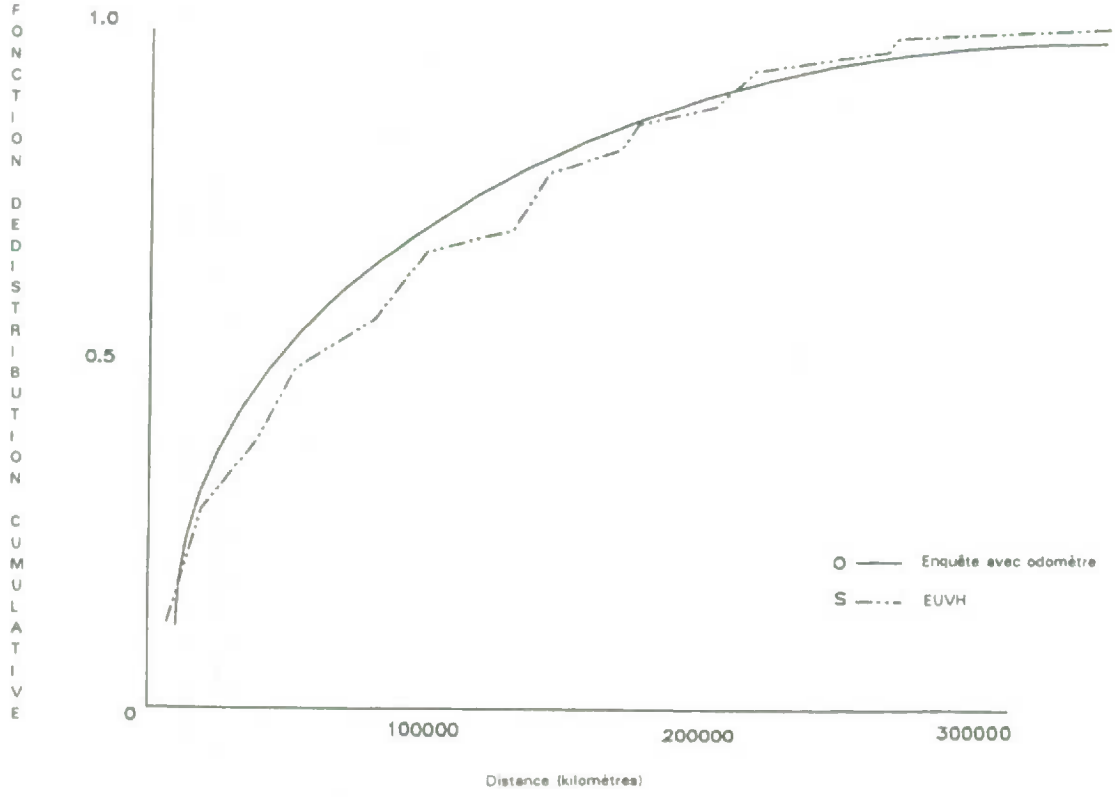


Figure 2. Fonction de distribution cumulative: camions rigides, distance parcourue



3. Fonction de distribution cumulative: camions articulés, distance parcourue



Un autre point digne d'intérêt est la lourde queue de la courbe de distribution des distances parcourues par les voitures et les camions rigides de l'EUVH, indiquant que la moitié supérieure des répondants tendent à fournir de mémoire une surestimation de la distance parcourue, tandis que la moitié inférieure des réponses cadrent relativement bien avec celles des répondants de l'enquête avec odomètre.

En résumé, l'enquête avec odomètre fait clairement ressortir le biais positif important des estimations de l'EUVH ainsi que le plus haut niveau de variabilité. Le biais est beaucoup plus élevé que l'erreur d'échantillonnage de l'EUVH et, de fait, il est plus important que l'erreur quadratique moyenne totale pour l'enquête.

Soulignons qu'en plus de l'enquête avec odomètre décrite ci-dessus, une enquête similaire avait été menée en même temps que l'EUVH de 1985. La première enquête complémentaire était de taille beaucoup plus petite et n'a pas permis de déceler un biais significatif, mais un écart positif entre les estimations de l'EUVH et de l'enquête avec odomètre avait été mesuré.

5. AUTRES MÉTHODES POSSIBLES

Bien que l'enquête avec odomètre est un outil d'évaluation très utile pour l'EUVH, elle ne constitue pas en soi une méthode de rechange étant donné qu'elle ne peut fournir de réponses à certaines questions clés comme la ventilation des déplacements selon l'objet et la région, le nombre de voyageurs transportés, le volume de marchandises, etc. Le journal serait une méthode appropriée pour ces besoins. On pourrait notamment envisager d'avoir recours à une enquête fondée sur la tenue d'un journal d'un mois avec un échantillon réparti également sur toute l'année.

On ne connaît pas avec certitude l'effet qu'aurait le choix d'une telle méthode sur le fardeau de la réponse et les coûts de collecte. D'une part, le fardeau des répondants augmenterait puisque ces derniers auraient à faire un certain nombre d'inscriptions dans le journal pendant le mois au lieu de remplir un seul questionnaire de mémoire. En outre, pour maintenir le même niveau d'erreurs types relatives, il faudrait augmenter la taille de l'échantillon étant donné que chaque questionnaire ne couvrirait qu'un seul mois et non pas l'ensemble de la période de référence de douze mois. (L'augmentation de l'échantillon dépendrait du coefficient de variation des estimations mensuelles comparativement aux estimations annuelles, de la corrélation d'un mois à un autre entre les valeurs des distances parcourues et des facteurs d'échantillonnage actuels.)

Par contre, l'enquête avec odomètre a montré que le fait de recueillir des chiffres plus proches des vraies distances parcourues a non seulement pour effet de réduire le biais de manière substantielle, mais aussi de faire réaliser quelques gains sur le plan de la variance. La taille de l'échantillon requise pour atteindre une erreur quadratique moyenne relative égale à celle de l'EUVH actuelle, en supposant que la méthode axée sur le journal entraînerait une variance de réponse ainsi qu'un biais négligeables, ne représenterait donc qu'une fraction de la taille de l'échantillon nécessaire à l'EUVH. De plus, l'enquête avec journal permettrait sans doute d'améliorer les réponses à d'autres questions du questionnaire, dans une mesure encore plus grande que pour la question de la distance totale parcourue. Enfin, avec un journal, les répondants pourraient trouver moins difficile de fournir des données directement accessibles que d'avoir à se rappeler des données se rapportant aux douze derniers mois. Il faudrait pousser davantage l'étude de ces possibilités par des discussions avec les utilisateurs et des essais.

Les économies de coûts et autres avantages de la méthode axée sur la tenue d'un journal, comparativement à ceux de l'enquête mémoire, sont énoncés brièvement dans les paragraphes qui suivent.

1. **Conception de l'enquête.** L'enquête mémoire sur l'utilisation des véhicules automobiles fonctionne déjà et les systèmes informatiques et mécanismes de traitement nécessaires sont en place. L'adoption d'une méthode axée sur un journal exigerait l'élaboration des méthodes et des systèmes s'y rapportant. Cela représente un obstacle majeur au changement.

2. **Création de la base de sondage et sélection de l'échantillon.** Dans un plan de sondage qui prévoit la répartition de l'échantillon dans le temps, la création de la base de sondage et la sélection de l'échantillon pourrait se faire au début de l'année ou périodiquement, par exemple chaque trimestre. Comparativement

à l'EUVH actuelle, une telle façon de procéder permettrait d'inclure les véhicules ayant cessé de figurer dans les fichiers au cours de l'année, mais ne pourrait tenir compte des nouveaux cas d'immatriculation en cours d'année. La seconde option serait plus coûteuse puisqu'elle exige trois séries additionnelles d'opérations pour chacun des huit fichiers d'État et territoire, plus l'appariement des unités sélectionnées en vue d'éviter tout chevauchement. Cependant, elle assurerait une très bonne couverture des véhicules pendant toute l'année, les véhicules ayant cessé de figurer dans les fichiers en cours d'année pouvant être pris en compte, alors que ce n'est pas le cas présentement avec l'EUVH.

3. **Expédition et traitement.** En vertu de la méthode axée sur le journal, les opérations d'expédition et de traitement seraient réparties tout au long de l'année. Cela comporte à la fois des avantages et des désavantages. Une plus petite équipe d'employés permanents pourrait se voir confier une plus grande part de la charge de travail, plus étendue, réduisant le recrutement et la formation. Pour l'EUVH de 1988, les opérations de traitement ont été réparties sur une période de cinq mois et un personnel formé de 57 personnes a été nécessaire. Si le traitement est réparti sur quatorze mois, le nombre d'employés nécessaires diminuerait d'au moins de moitié, avec une baisse correspondante des coûts de recrutement et de formation. On peut également s'attendre à ce qu'un personnel plus expérimenté fasse un travail de plus grande qualité.

   De plus, compte tenu de la plus grande disponibilité des données, les répondants devraient éprouver moins de difficulté à remplir le journal et les activités de suivi pourraient diminuer de manière significative. En revanche, étant donné que la tenue d'un journal représente un fardeau plus lourd pour les répondants, les activités de suivi pourraient rester tout aussi importantes dans le but de maintenir des taux de réponse acceptables. Sans des essais en bonne et due forme de l'enquête avec journal, il ne nous est pas possible de prévoir dans quelle mesure ces facteurs pourraient se faire contrepoids.

   Une réduction de la taille de l'échantillon faite en fonction du même niveau d'erreur quadratique moyenne que pour l'enquête de 1988 entraînerait d'autres économies sur le plan de l'expédition et du traitement.

4. **Actualité des données pour les utilisateurs.** Une enquête fondée sur l'utilisation d'un journal serait menée tout au long de la période de référence plutôt que l'année suivante; on pourrait donc s'attendre à une amélioration sur le plan de l'actualité des résultats.

5. **Mesure de la qualité des données.** En supposant qu'une enquête avec journal entraînerait une réduction importante du biais, la fiabilité des estimations deviendrait relativement visible par l'intermédiaire des ETR. Avec un échantillon restreint, au niveau des petites cases, certaines ETR seraient assez élevées. Les utilisateurs pourraient réagir davantage qu'aujourd'hui car le niveau actuel d'EQM est élevé mais peu connu dans l'ensemble (les résultats de l'enquête avec odomètre ont été publiés dans une note technique, mais les ETR n'avaient pas fait l'objet d'un ajustement par rapport aux EQM).

6. CONCLUSION

L'orientation future de l'EUVH consiste à tirer profit des connaissances d'une équipe multidisciplinaire afin de réunir en un tout les divers aspects de la conception d'une enquête et d'élaborer une méthode qui assure aux utilisateurs la stratégie globale la plus efficace. Compte tenu de l'effet relativement peu important de l'erreur d'échantillonnage sur l'erreur totale, la taille des échantillons sera réduite et les travaux méthodologiques ne seront plus axés sur la réduction de l'erreur d'échantillonnage, mais plutôt sur la réduction des erreurs de réponse et de non-réponse. Enfin, il y aura lieu d'informer davantage certains utilisateurs sur la notion d'erreur totale de l'enquête. Pour éviter que l'attention ne se fixe que sur les erreurs d'échantillonnage, nous devrons essayer de trouver des moyens de fournir plus régulièrement et pour un plus grand nombre d'estimations une mesure correspondante de l'erreur totale.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

MISE À L'ESSAI D'UN PLAN DE PRIMES DE RENDEMENT À STATISTIQUE CANADA

J.-F. Gosselin[1]

RÉSUMÉ

Un plan de primes de rendement comportant le partage des gains avec les employés participants a été mis sur pied à titre expérimental dans trois petites unités chargées des aspects opérationnels de programmes statistiques. Cette communication montre qu'il y a eu des gains de productivité significatifs tandis que les objectifs de qualité et de respect des délais ont été maintenus ou dépassés. Les réactions des employés ont été très positives, et il y a lieu de penser que le plan, quand on le compare aux groupes de contrôle, réduit le stress en milieu de travail.

MOTS CLÉS: Primes; productivité; qualité.

1. INTRODUCTION ET RENSEIGNEMENTS GÉNÉRAUX

Une grande réorganisation a eu lieu à Statistique Canada vers le milieu des années 80 afin de centraliser les opérations statistiques qui se déroulaient auparavant dans les divisions spécialisées. Le principal objectif consistait à réaliser des gains d'efficience grâce à une utilisation plus judicieuse et plus rationnelle des ressources opérationnelles. Nous nous sommes rendu compte dès le début que pour maximiser les gains de cette vaste entreprise, nous devions non seulement faire une meilleure utilisation des ressources, mais aussi nous engager à long terme à repenser la conception et la gestion des opérations.

Il y a eu plusieurs initiatives visant à automatiser divers processus, à appliquer avec succès de nouvelles techniques ou méthodes telles que l'interview téléphonique assistée par ordinateur (ITAO) et le codage automatisé, etc. Nous discuterons ici d'une expérience que nous avons tentée avec succès et qui est complètement différente de nos méthodes habituelles de gestion des opérations dans la fonction publique.

Dès le début, il a été décidé d'élaborer et de mettre en place des mécanismes de rétroaction permettant de surveiller la productivité et la qualité à cause de la nécessité d'atteindre des niveaux de rendement plus élevés. Par conséquent, des mesures ont été prises pour commencer l'établissement de normes de travail adaptées aux besoins ainsi que l'élaboration de méthodes de contrôle de la qualité statistique pour les principales applications. Bien qu'un accroissement de la productivité ait été observé par suite de l'adoption de ces mesures, il est rapidement devenu évident qu'on ne pourrait réaliser d'améliorations importantes et durables sans le plein appui des employés et l'utilisation de primes de rendement. C'est alors que nous avons pensé à créer un plan de partage des gains.

Le but premier des plans de partage des gains consiste à partager avec les employés les gains résultant d'un accroissement de la productivité au-dessus de certains niveaux déterminés. De tels plans sont très courants dans le secteur privé et les avantages qu'ils comportent sont généralement des bonis. Pour notre part, nous avons décidé d'offrir des congés. La caractéristique distinctive de notre expérience vient du fait que notre plan a été appliqué dans le contexte de la fonction publique.

---

[1] J.-F. Gosselin, Directeur, Division de la recherche et du développement des opérations, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6.

L'application initiale du plan a été limitée dans l'enquête des permis de bâtir. L'idée des congés a plu énormément à ces employés qui ont réussi à faire passer leurs niveaux de productivité d'environ 60%, ce qui n'est pas un niveau inhabituel dans des secteurs non soumis à des normes de travail, à des niveaux bien au-delà de 100%.

À partir des résultats de cette étude pilote de portée très restreinte qui ne visait que cinq employés, il a été décidé de passer à un essai de grande envergure dans deux autres Unités: celles de l'enquête sur les voyages internationaux (EVI) et de l'enquête sur l'emploi, la rémunération et les heures de travail (EERH). Ces Unités comptaient respectivement 15 et 25 employés, ce qui portait à 45 le nombre d'employés des trois Unités participant à l'expérience.

Le tableau 1 illustre les gains de productivité réalisés.

Tableau 1: Plan de primes de rendement - résultats sommaires

	UNITÉS DE PRODUCTION		
	Permis de bâtir[1]	Voyages internationaux	EERH (enquête sur l'emploi, la rémunération et les heures de travail)
Employés	3 - 4	15	25
Productivité (initiale)	58%	69%	75%
Productivité (24/11/89)	108%	118%	95%
Total des jours de congé gagnés	62	357	101
Jours/employé	18	24	4

[1] Le plan des permis de bâtir a été interrompu à la fin de février 1989 à cause de la régionalisation de l'enquête.

Selon l'Unité, la productivité a grimpé de 60%-75% à 95%-118%, ce qui représente une augmentation considérable, avant que le plan soit interrompu. Encore selon l'Unité, le nombre de congés gagnés par les employés variait entre 4 et 24 pour la période de l'essai (environ 16 mois), soit une moyenne de 12 jours par employé. Comme on peut le voir, les congés constituaient un facteur de motivation très puissant.

Alors si l'expérience s'est avérée un tel succès, pourquoi, vous demanderez-vous, a-t-on mis fin à ce plan? La raison est très simple: la Cour fédérale a jugé qu'il contrevenait à la Loi sur les relations de travail dans la fonction publique. La présente communication ne vise pas à traiter les aspects juridiques de cette question. Elle a plutôt pour objet de se pencher sur les points suivants:

- Quelle a été l'incidence du plan sur la productivité?
- L'accroissement de la productivité a-t-elle eu un effet négatif sur la qualité et le respect des délais?
- Le plan a-t-il occasionné plus de stress?

Avant d'aborder ces sujets, les principales caractéristiques du plan sont décrites ci-dessous.

2. CARACTÉRISTIQUES DU PLAN

Une des caractéristiques fondamentales du programme de primes de rendement mis sur pied à Statistique Canada est qu'il s'agit d'un plan **collectif**. C'est là une différence fondamentale entre ce programme et les plans de rémunération à la pièce que l'on retrouve dans le secteur privé et selon lesquels la rémunération d'un employé est fonction de sa productivité individuelle.

Dans le cadre de notre plan, les congés gagnés dépendent de la productivité d'une unité entière de sorte que l'accent est mis sur le travail d'équipe, c'est-à-dire sur la façon d'atteindre les objectifs, y compris les objectifs de qualité et de respect des délais, en mettant à contribution chaque employé de la façon la plus efficace possible. La cohésion et la communication à l'intérieur du groupe deviennent des facteurs décisifs. La culture organisationnelle doit être modifiée; en effet, un employé qui auparavant faisait ce qu'on lui disait de faire est maintenant responsable en partie de l'amélioration des moyens utilisés pour obtenir les résultats voulus. Dans de telles situations, des séances de promotion du travail d'équipe ont eu lieu.

Une autre des principales caractéristiques du plan est qu'il est fondé sur la **participation volontaire**. La motivation et, conséquemment, l'appui des employés sont à la base de cet essai qui devrait exclure toute forme de contrainte. Cet aspect va de pair avec **l'emploi garanti**. Personne n'est intéressé à se placer dans une situation qui lui ferait perdre son emploi, donc il va sans dire que c'est un point essentiel. **Des primes sous forme de congés** sont gagnés lorsque la productivité de l'Unité commence à dépasser 80% de la norme. Initialement, la cible était 100%, mais cet objectif aurait découragé les employés. Il a été décidé de baisser le seuil et d'accorder une proportion de plus en plus élevée des gains réalisés, cette proportion pouvant aller jusqu'à 50% pour un niveau de productivité excédant 100%.

Afin de maintenir un bon moral et de créer un climat propice à l'amélioration continue par l'entremise de la participation des employés, le plan ne prévoyait **aucun changement dans les normes de travail**, sauf dans les cas où la nature du travail avait été modifiée, par exemple à la suite du remaniement d'une enquête ou d'un investissement important de la part de la direction. Enfin, afin de promouvoir une communication sans entraves et de permettre aux suggestions des employés d'être évaluées, un **comité à deux niveaux** a été créé avec la pleine participation des employés.

3. ÉVALUATION

3.1 Analyse de la productivité

Les normes de travail forment la pierre angulaire du programme de primes de rendement. Elles présentent un élément important d'objectivité au moment de l'établissement des objectifs de production et constituent une base qui permet de surveiller les niveaux de rendement des employés et d'en informer ces derniers chaque semaine.

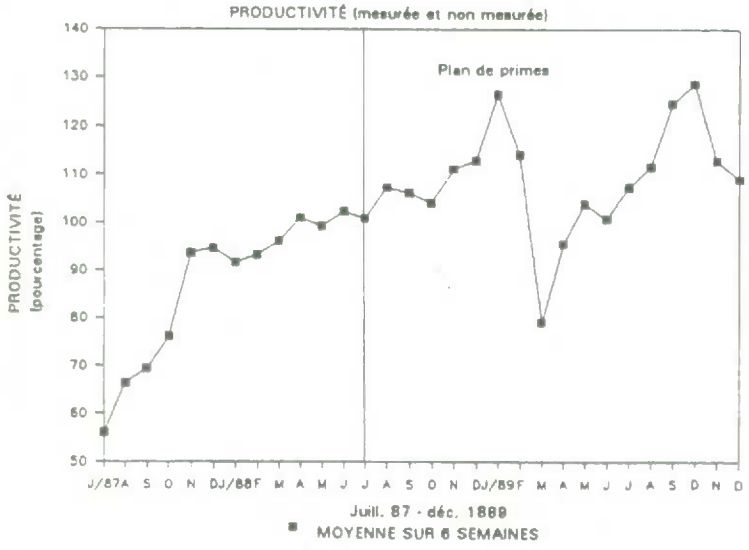
Les niveaux de productivité des trois groupes participants ont été analysés pour l'année précédente et pour les 16 mois suivant la mise sur pied officielle du plan le 1er juillet 1988. Les principales observations qui découlent de cette analyse sont présentées ci-après.

a) Dans le cas de l'Unité des permis de bâtir (PB), la productivité moyenne s'établissait à 58% avant que toute forme de prime de rendement ne soit offerte. En l'espace de quelques mois, la productivité de l'Unité variait entre 115 et 130%. Durant la période d'essai, elle a été systématiquement supérieure à 125% sauf pour la période précédant immédiatement la fin du plan par suite de la régionalisation qui était alors en suspens.

b) Durant la période d'essai, la productivité de l'Unité de l'EVI (tableau 2) s'est régulièrement maintenue au-delà du seuil donnant droit aux primes. La plupart des mois, elle a été supérieure à 100%, le maximum atteint s'établissant à près de 130% et le minimum, à environ 80%, ce dernier taux étant attribuable aux mois de faible volume. Si la direction avait prévu ce ralentissement, des tâches additionnelles auraient pu être confiées à l'Unité et la productivité aurait alors vraisemblablement pu être maintenue. Cet exemple fait ressortir l'importance d'une planification attentive et de l'établissement d'un calendrier des activités dans le cadre de ce genre de plan.

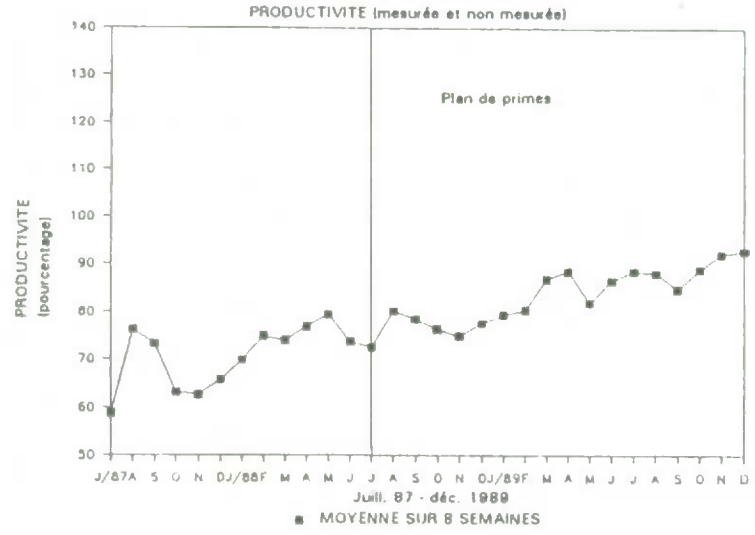
En ce qui a trait à la période précédant l'application du plan, nous avons observé une tendance graduelle à la hausse de la productivité qui a atteint 90%. Cette augmentation a résulté à la fois d'une meilleure organisation du travail, de l'application de normes et de la tenue des séances officielles de promotion du travail d'équipe. Toutefois, le personnel de l'Unité de l'EVI savait qu'il était possible qu'un plan soit appliqué dans leur Unité et travaillait dans l'expectative que cette possibilité se réalise. Ainsi, durant la période de février à juin 1988, la productivité a augmenté de 90% à plus de 100%. Nous sommes d'avis qu'il aurait été très difficile de dépasser et de maintenir des niveaux bien au-dessus de 80% sans la motivation offerte par un facteur tangible comme des congés.

Tableau 2. Voyages Internationaux



c) Pour l'Unité de l'EERH (tableau 3), nous avons constaté une augmentation graduelle et régulière de la productivité qui est passée de son niveau le plus faible de 60% à environ 90% à la fin de la période d'essai. Là encore, ces hausses sont notables, bien qu'elles ne soient pas aussi marquées que celles qui ont été observées pour l'Unité de l'EVI. Cette différence peut être due au fait qu'il s'agissait de la dernière Unité pour laquelle on établissait des normes. On a remarqué une tendance évidente à la hausse depuis l'application du plan et cette tendance aurait pu se poursuivre si le plan n'avait pas été interrompu, étant donné que les employés ne faisaient que commencer à profiter des fruits de leurs efforts.

Tableau 3. EERH



d) Les données ci-dessous représentent les congés gagnés par les employés des Unités participant à l'essai.

	Congés gagnés		
	P.B.	**E.V.I.**	**EERH**
Journées totales	62	357	101
Jours/employé	18	24	4
Période	Juillet 88 Février 89	Juillet 88 Novembre 90	Juillet 88 Novembre 90

e) Dans l'ensemble, il y a eu un accroissement important de la productivité dans toutes les Unités. Si les taux de productivité étaient demeurés à leurs niveaux initiaux, il aurait fallu augmenter la taille du groupe d'environ 9 A.-P. pour accomplir le travail effectué par les employés participants depuis la mise sur pied du plan.

3.2 Qualité et respect des délais

L'accroissement de la productivité a-t-elle eu une incidence sur la qualité du travail des employés et sur leur aptitude à respecter les délais?

Les données recueillies laissent entendre qu'en fait, les objectifs de qualité et de respect des délais ont été atteints et, dans certains cas, dépassés.

a) Le nombre de documents de l'EERH non encore traités à la date limite (tableau 4) est bien inférieur au nombre correspondant pour l'année précédente. En fait, le programme de primes ainsi que les changements relatifs aux procédures de régionalisation ont contribué à réduire pratiquement à zéro le nombre de documents non traités à la date limite d'inclusion dans l'enquête depuis janvier 1989.

b) Dans le cas de l'Unité de l'EVI, les niveaux qualitatifs des divers produits observés pour la période d'essai et l'année précédente n'ont pas varié de façon notable, comme on peut le voir d'après les activités de codage (tableau 5) ou d'introduction au clavier (tableau 6) pour lesquelles la qualité moyenne après contrôle (QMAC) est demeurée bien inférieure de la limite de la qualité moyenne après contrôle (LQMAC).

c) Pour l'EERH, la période d'essai correspondait approximativement à la deuxième année de la mise en pratique de cette nouvelle application de CQ. Il est très intéressant de noter (tableau 7) que la qualité s'est améliorée régulièrement malgré l'accroissement de la productivité. Nous croyons que ces résultats sont attribuables à un système de rétroaction efficace dans le cadre duquel des réunions sont convoquées à intervalles réguliers afin d'examiner avec le personnel la nature des erreurs et les mesures à prendre pour améliorer la qualité. Le niveau de qualité devrait se stabiliser une fois que le plan sera parvenu à maturité.

Tableau 4. EERH - Respect des Delais

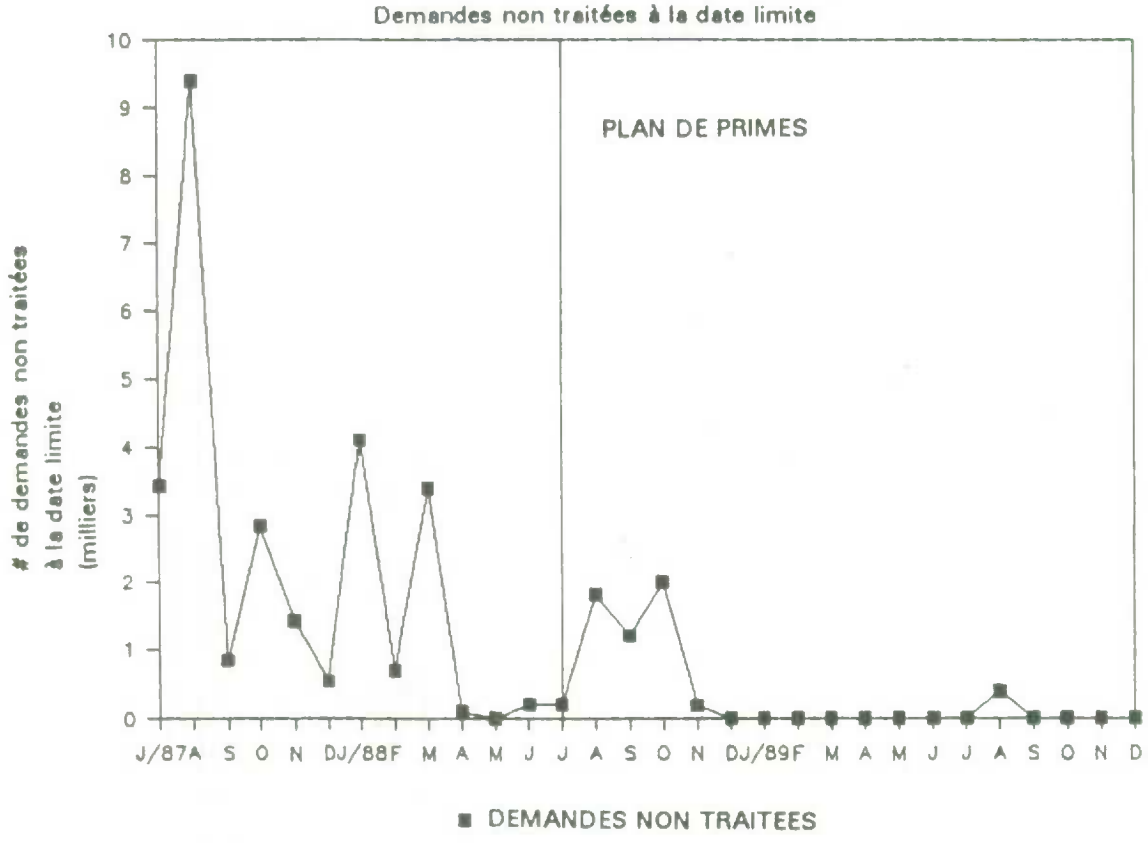


Tableau 5. EVI - Qualite (Codage)

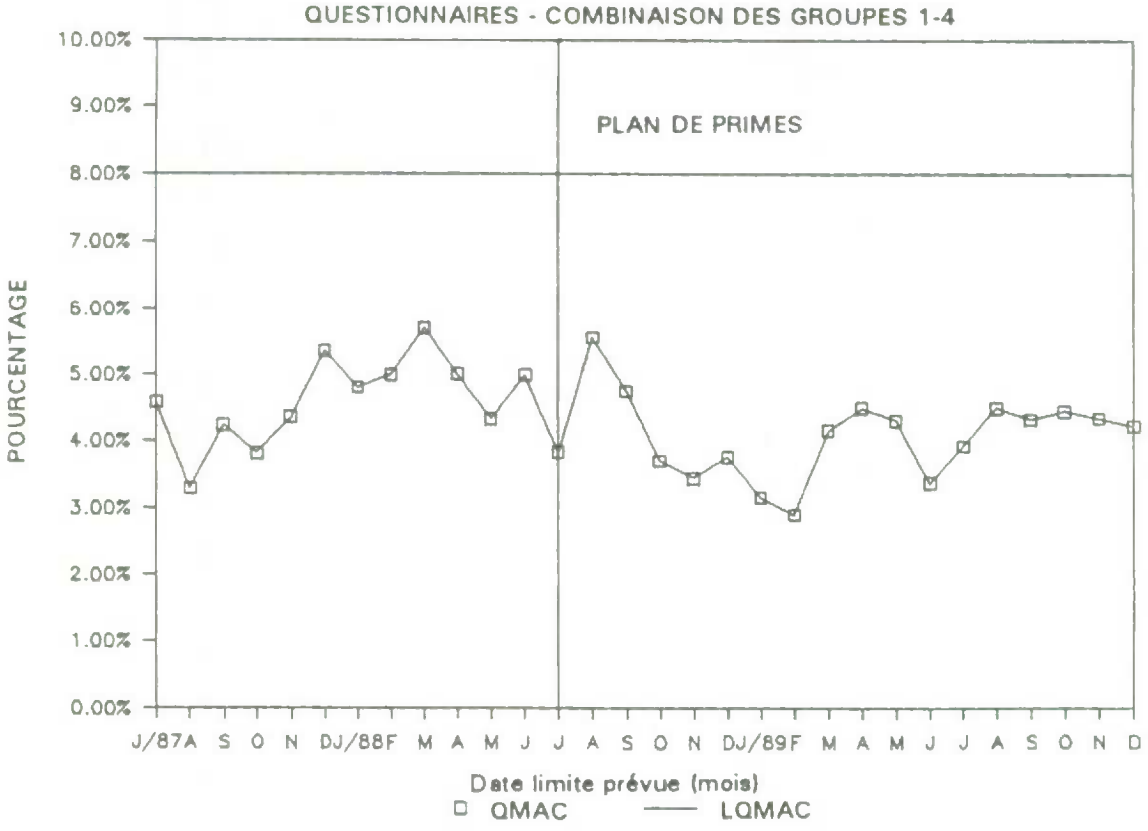


Tableau 6. EVI - Qualite (Introduction au Clavier)

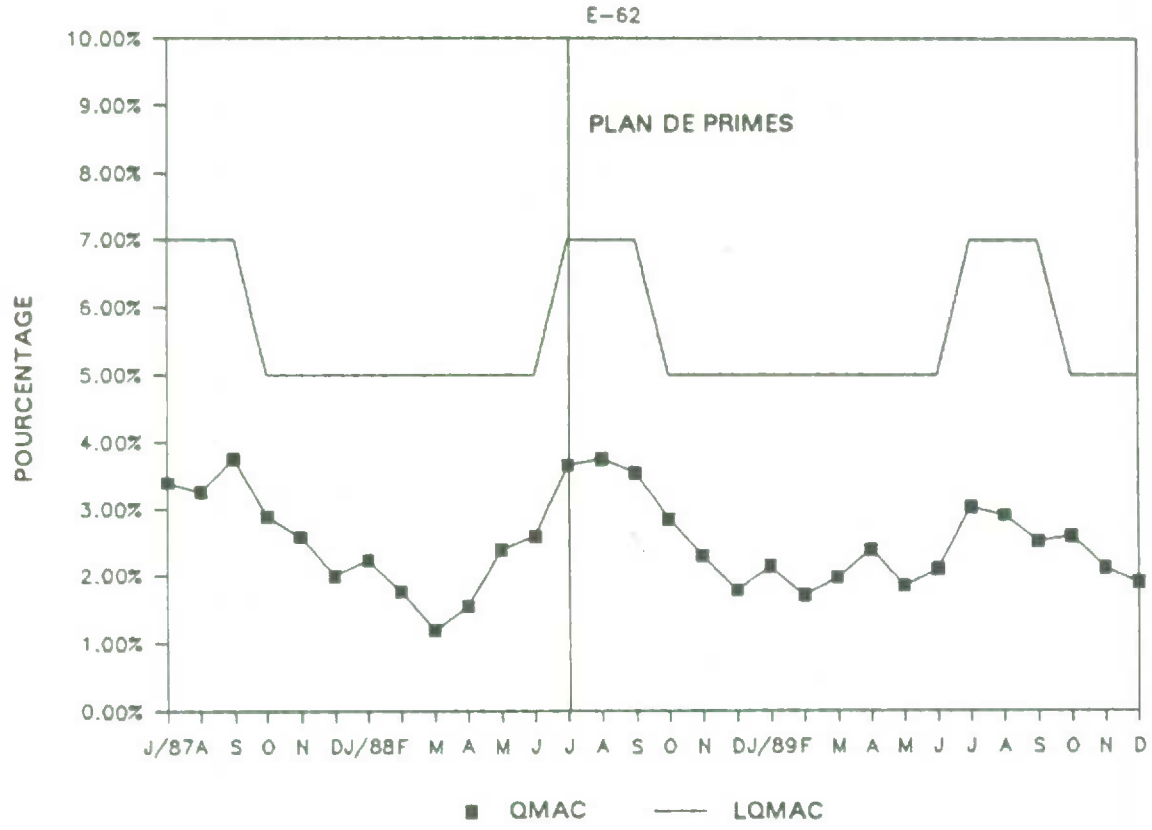
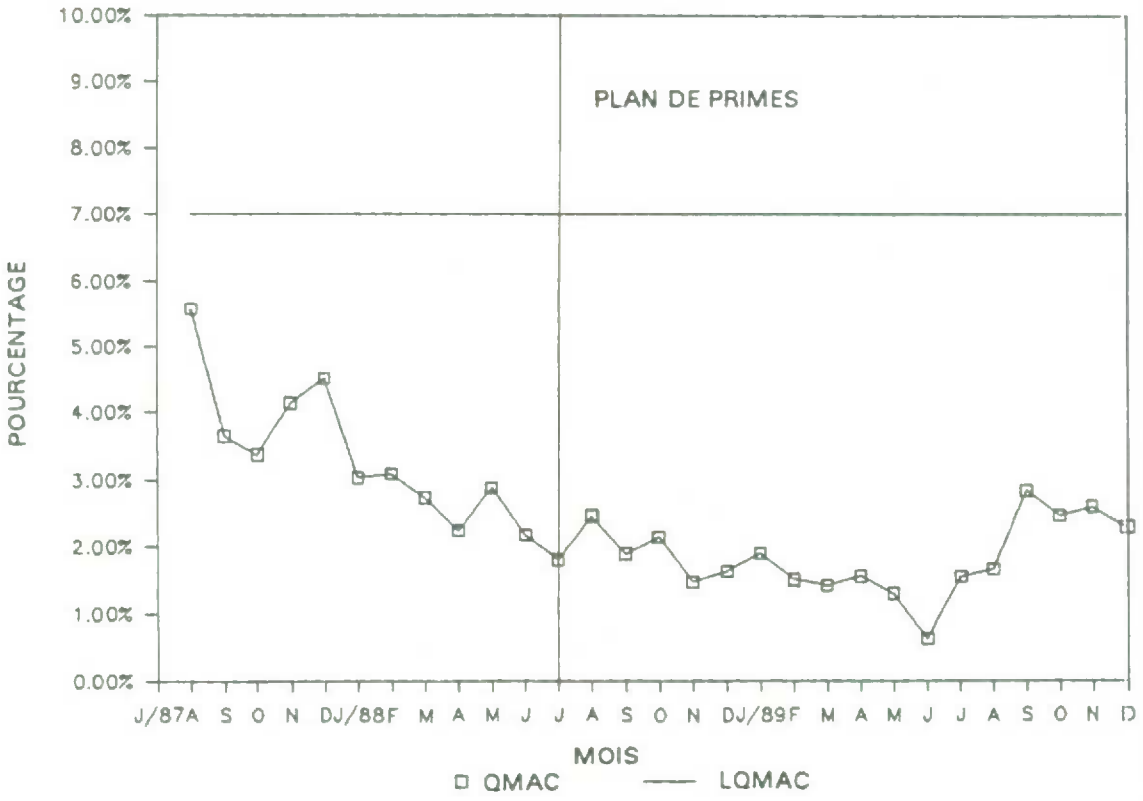


Tableau 7. EERH - Qualite (Controle)



3.3 Réaction des employés

La réaction des employés à l'égard de leur participation au plan a été évaluée de plusieurs façons:

a) Selon les résultats d'un questionnaire interne distribué au début de juin par la Section du développement organisationnel de la Division du personnel de Statistique Canada, 94% des employés participants souhaitaient que le plan soit maintenu. Le sondage a aussi révélé un taux élevé de satisfaction à l'égard de l'emploi: 48% de ceux qui ont répondu ont déclaré un degré de satisfaction plus élevé que l'année précédente. Seulement 3 répondants (9%) ont indiqué que leur emploi leur procurait moins de satisfaction que l'année précédente. Le sondage a donné lieu à bon nombre de commentaires et de suggestions.

b) Le professeur William Jones de l'Université Carleton a été chargé d'examiner les niveaux de stress chez les employés participants par rapport à ceux d'autres employés de Statistique Canada de catégories et de niveaux équivalents. Voici un résumé de ses conclusions:

"Le stress total relié au travail est plus faible, et non plus élevé, chez les employés participant au plan de primes de rendement. ...

Les employés participant au plan de primes de rendement ont connu considérablement moins de conflits de rôles et ont manifesté un degré plus élevé de satisfaction à l'égard de leur emploi et un plus grand sens d'appartenance au Bureau. Il n'y avait absolument aucune indication que la mise en oeuvre du plan avait entraîné plus de stress ou de tension chez les employés ou entre eux."

3.4 Facteurs de succès

Les éléments qui suivent sont considérés comme les principaux facteurs ayant assuré le succès de l'expérience:

a) Tout commence par une idée. Et il arrive souvent que les bonnes idées soient rejetées trop vite avant d'avoir pu être explorées ou améliorées.
b) La direction devait être prête à prendre certains risques. Des changements radicaux peuvent rarement être faits sans cette condition de base.
c) Les employés avaient l'occasion de participer du début à la fin de l'expérience.
d) Il y avait des avantages tangibles tant pour les employés que pour la direction et une incitation à travailler en équipe pour assurer le succès de l'essai.
e) Les employés et la direction à tous les niveaux ont fait preuve de beaucoup de franchise, d'esprit d'entraide et de flexibilité. Peut-être le fait qu'il s'agissait d'un essai sur une petite échelle dans un milieu de travail contrôlé a contribué à créer ce climat favorable.
f) La sécurité d'emploi a été un facteur clé de la réussite de l'essai. En somme, personne n'est disposé à se mettre dans une situation qui lui fait perdre son emploi.

Tout cela peut être résumé en trois conditions essentielles: CONSULTATION, CONCERTATION et PARTAGE à tous les niveaux, c.-à-d. avantages, travail, responsabilité, décisions, améliorations.

4. CONCLUSION

La mise à l'essai d'un plan de primes de rendement à Statistique Canada a démontré qu'une telle approche pourrait être utilisée avec grand succès dans un milieu opérationnel pour améliorer la productivité tout en maintenant ou en dépassant les objectifs de qualité et de respect des délais. Des études ont aussi indiqué que les avantages importants comprennent une hausse marquée du degré de satisfaction à l'égard de l'emploi, un plus grand sens d'appartenance au Bureau et une réduction des niveaux de stress.

Il ne fait aucun doute que la mise en oeuvre de programmes de primes de rendement de ce genre modifierait de façon très importante les activités de gestion de la fonction publique canadienne si elle était étendue à grande

échelle. Par ailleurs, ces programmes ne pourraient être appliqués que dans les services où le travail peut être mesuré et la qualité peut être contrôlée.

Il faudra vraisemblablement jeter un regard neuf sur certains aspects que nous considérons actuellement comme admis. Par exemple, comment doit-on classifier les emplois et évaluer les employés lorsque le travail est réparti entre les employés d'une unité entière?

Une recommandation a été faite dans le cadre de Fonction publique 2000 en vue de modifier la Loi actuelle de façon à permettre l'adoption de tels plans. Il est à espérer qu'il en résultera des occasions dont les employés et la direction pourront profiter.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

UN SYSTÈME POUR MESURER LA QUALITÉ DES ENQUÊTES PÉRIODIQUES*

R.D. Tortora[1]

RÉSUMÉ

Le présent article décrit à grands traits un système permettant de mesurer la qualité des enquêtes périodiques. L'approche proposée consiste à élaborer une matrice de mesures de la qualité tout au long du déroulement de l'enquête. Les éléments de cette matrice sont définis à partir de l'interaction de trois facteurs: les critères de qualité, les sources d'erreurs et les étapes de l'enquête. L'analyse de l'évolution de ces éléments dans le temps peut ensuite servir à améliorer la qualité de l'enquête.

MOTS CLÉS: Enquêtes périodiques; qualité; amélioration de la qualité.

1. INTRODUCTION

Le présent article propose un système de mesure de la qualité des enquêtes périodiques. Le gestionnaire d'enquête qui dispose d'un tel outil de gestion est à même d'ordonner ses priorités en matière d'amélioration de l'enquête. L'outil de gestion que constitue le système susmentionné est axé sur l'amélioration de la qualité par le biais du déroulement de l'enquête plutôt que par la réduction au minimum de l'erreur relative par coût unitaire.

Les écrits spécialisés accordent une place privilégiée à la qualité des enquêtes statistiques. Aussi, les statisticiens ont-ils eu tendance à écarter une définition plus générale de la qualité en faveur des seules notions de précision ou de justesse. Il y a profusion d'articles consacrés à l'amélioration de ce critère de qualité d'une enquête que l'on appelle précision. En revanche, d'autres articles de nature plus générale abordent la notion de qualité sous l'angle de la justesse. Hansen, Hurwitz et Bershad (1961) sont les auteurs d'un article qui fait autorité à cet égard. Pour leur part, Platek et Singh (1981) établissent un lien entre le rendement d'une enquête et ses mécanismes de contrôle en essayant de cerner les principales sources d'erreurs et en étudiant les mesures de contrôle pouvant assurer un rendement donné. Brooks et Bailar (1978) décrivent la structure des erreurs d'une estimation et d'une enquête données, à savoir l'emploi mesuré au moyen du recensement de la population actuelle. Spencer et Mulry (1990) ont élaboré un modèle de l'erreur totale aux fins de l'enquête postcensitaire. D'autres encore se sont penchés sur plusieurs critères de qualité. Groves (1989) établit un rapport entre les erreurs et les coûts. Jabine (1990) élabore un profil de qualité pour l'enquête sur le revenu et la participation aux programmes. Des organismes statistiques se sont dotés de lignes directrices concernant la qualité [Statistique Canada (1987)] et de mesures du rendement des enquêtes [Energy Information Administration (1989)] afin d'évaluer les performances de leurs produits. Ces deux derniers documents sont de nature très générale et touchent à la gamme complète des activités d'enquête, quoique, dans ces deux cas également, les questions de justesse occupent une large place. Aucun de ces articles ou produits ne favorise une approche globale de la qualité largement répandue actuellement dans les affaires et dans l'industrie, afin d'améliorer la qualité des enquêtes. Dans le présent article, certaines des idées proposées par Juran (1964) et Deming (1986) sont mises à profit afin de favoriser l'amélioration soutenue des enquêtes périodiques.

---

* Le présent article n'exprime que les opinions de l'auteur.

[1] R.D. Tortora, Chief, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233. U.S.A.

2. DONNÉES DE BASE

Il existe plusieurs raisons importantes d'élaborer un système de mesure de la qualité des enquêtes. D'abord, fort d'un système, le gestionnaire d'enquête est à même d'évaluer la qualité globale de l'enquête et de repérer les lacunes que présentent les méthodes de mesure de la qualité mises en oeuvre. Deuxièmement, non seulement le système permet-il au gestionnaire de connaître la qualité de l'enquête à un moment donné, mais il lui permet aussi d'interpréter les données disponibles pour mesurer l'évolution de la qualité de l'enquête dans le temps. En dernier lieu, grâce au système, le gestionnaire a une vue d'ensemble de la méthode d'enquête, à partir de la conception initiale et de l'échantillonnage jusqu'à la publication et à la révision des estimations. Le gestionnaire est ainsi mieux en mesure de traiter des questions ayant trait au rendement de l'enquête sur une période donnée, à l'utilisation ou à l'affectation des ressources et à l'anticipation des problèmes éventuels, y compris la justification des futurs besoins en ressources.

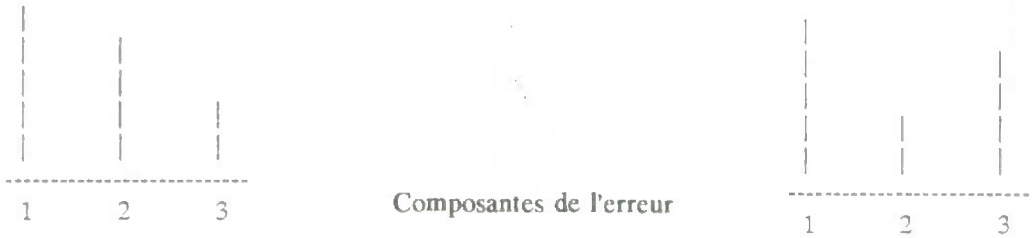
Le contenu du présent article repose sur cinq hypothèses. On suppose d'abord qu'il ne sera pas toujours facile d'obtenir les ressources financières nécessaires à l'amélioration de la qualité des enquêtes. Il se peut même que le contraire soit vrai et que les ressources dont on dispose s'amenuisent. Le gestionnaire d'enquête devra donc envisager une baisse éventuelle de la qualité de l'enquête, à défaut de méthodes qui lui permettent de tirer un meilleur parti de ses ressources (qui, par ailleurs, vont en diminuant). Deuxièmement, on suppose que le gestionnaire dispose de nombreuses données lui permettant de mesurer la qualité de l'enquête. Ces données, comme l'erreur d'échantillonnage, le taux de non-réponse, les éléments de coût, etc., si elles sont recueillies périodiquement, peuvent constituer des indices de la qualité de l'enquête dont le gestionnaire peut déduire de nombreux renseignements au sujet du rendement de celle-ci. On suppose, en troisième lieu, que la qualité ne se résume pas uniquement à la justesse aux yeux du gestionnaire. Elle englobe d'autres notions telles le respect des délais, la pertinence et les ressources. Le respect des délais comprend non seulement les délais de publication des résultats mais également les délais d'exécution des différentes étapes de l'enquête. Il va sans dire que la qualité englobe également la qualité des ressources, il n'est point besoin d'insister sur ce fait. Le dernier critère de qualité est la pertinence. À cet égard, il s'agit non seulement de tenir compte des utilisateurs des données et de la pertinence de l'enquête pour ces derniers, mais également de la pertinence des produits ou résultats à chaque étape de l'enquête. Par exemple, le gestionnaire s'intéresse à la pertinence du questionnaire non seulement par rapport aux données qu'il permettra de procurer aux utilisateurs, mais également par rapport à l'utilisation qu'en feront les interviewers et les répondants. Aux yeux du gestionnaire, tant les premiers que les seconds sont des clients touchés par le déroulement de l'enquête. En quatrième lieu, on invoque l'un des principes fondamentaux mis de l'avant par les maîtres à penser de l'amélioration de la qualité, à savoir qu'une des caractéristiques fondamentales des systèmes qui tournent rondement est l'existence de nombreuses données servant à en décrire le rendement. Dans ces circonstances, l'amélioration de la qualité peut découler d'une analyse solide de données existantes plutôt que d'études spéciales nécessitant la collecte de données de base sur les performances du système et épuisant le peu de ressources dont on dispose. En dernier lieu, sans nier la valeur de la mesure directe de la qualité, on suppose néanmoins que des renseignements tels l'erreur systématique de non-réponse, des indices de qualité prélevés périodiquement, les taux de non-réponse, les caractéristiques des non-répondants, etc., peuvent se révéler tout aussi utiles. À remarquer que les hypothèses énumérées ci-dessus commandent beaucoup d'analyses de données. Pour être utile, tout système doit ordonner ses données d'une façon qui encourage des analyses et des comparaisons aisées et utiles, surtout sous forme de graphiques chronologiques et d'autres représentations graphiques.

Avant d'aborder le modèle décrivant le système de mesure de la qualité des enquêtes périodiques, il est utile d'envisager certaines questions comme la suivante: Comment le gestionnaire d'enquête saura-t-il que la qualité d'une enquête s'est améliorée? Bien sûr, la question ne se pose pas si les crédits supplémentaires alloués à l'enquête servent à assurer un meilleur respect des délais ou à accroître la justesse ou la pertinence des données. Il est toutefois plus difficile de répondre à cette question lorsque les ressources consacrées à l'enquête demeurent inchangées. Quand y a-t-il eu amélioration de la qualité? Si le gestionnaire est en mesure d'améliorer au moins un des critères de qualité et si les autres critères demeurent constants, on peut alors dire qu'il y a eu amélioration de la qualité de l'enquête.

Il semble que l'analyse de Pareto constitue un outil de grande valeur pour le gestionnaire qui envisage l'amélioration de la qualité. Juran (1964) a recourt à ce concept pour aider un organisme à se concentrer sur les principaux problèmes avec lesquels il est aux prises. Si trois des vingt-cinq produits que fabrique une

entreprise engendrent soixante-dix pour cent de ses revenus, alors, suivant l'analyse de Pareto, l'entreprise devrait se consacrer à l'amélioration de la qualité de ces trois produits de base. Lorsqu'il s'agit d'enquêtes, il peut ne pas être évident qu'une étape ou qu'un produit soit plus important qu'un autre. Il semblerait plutôt que la qualité de l'enquête comme telle soit fonction de la qualité de la mise en oeuvre de l'ensemble des étapes de l'enquête. Il est néanmoins possible de modifier l'analyse de Pareto afin qu'elle puisse être utile au gestionnaire d'enquête. La figure 1 illustre l'utilisation d'une telle analyse pour évaluer deux critères de qualité, soit les coûts et la justesse. La figure 1a présente une estimation du coût de chacune de trois composantes de l'erreur. Ces composantes sont présentées arbitrairement dans l'ordre décroissant des coûts. La figure 1b présente l'estimation de la contribution de chacune de ces composantes à l'erreur totale, les trois composantes étant présentées dans la même séquence qu'à la figure 1a.

Figure 1. Analyse de Pareto: Coûts en fonction de l'erreur



1a. Coût de trois composantes de l'erreur

1b. Contribution des mêmes trois composantes à l'erreur totale

Étant donné que l'erreur dont sont entachées ces trois composantes ne suit pas la même courbe descendante, le gestionnaire d'enquête a l'occasion d'entreprendre un projet d'amélioration de la qualité de l'enquête. Il peut par exemple transférer les coûts de la composante 2 à la composante 3 en vue de diminuer la contribution de cette dernière à l'erreur totale.

3. LE SYSTÈME

Cette section présente les grandes lignes d'un système de mesure de la qualité des enquêtes périodiques. L'approche consiste à étudier l'interaction de trois facteurs, les critères de qualité, la structure des erreurs d'une enquête et les étapes de l'enquête, puis à élaborer une matrice de la qualité données à chaque cycle d'enquête. Chaque élément de la matrice correspond à une donnée simple qui permet de mesurer la qualité de l'enquête, et chacun de ces éléments est déterminé à partir de l'interaction des critères de qualité, de la structure des erreurs et des étapes de l'enquête. Certains critères, comme le taux de non-réponse, seront faciles à obtenir à partir de l'enquête, tandis que d'autres peuvent se révéler plus difficiles à obtenir et nécessiter la réalisation d'études spéciales. Ces derniers ne pourront être obtenus à chaque cycle d'enquête, mais à mesure que la matrice évolue, on peut planifier la réalisation périodique d'études spéciales ou encore réaliser de telles études au moment où il manque des données particulières sur la qualité de l'enquête. Une limite de tolérance est un autre renseignement important qui peut être associé à chacune des cellules; elle aura été préétablie et aidera à juger si le déroulement de l'enquête est acceptable. L'erreur d'échantillonnage relative préétablie visant une estimation particulière obtenue à l'aide de l'enquête constitue un exemple de limite de tolérance. Il n'existe aucune matrice de mesures de la qualité qui soit universelle: elles différeront selon l'enquête et le gestionnaire d'enquête.

Les critères de qualité sont la pertinence, la justesse, le respect des délais et les ressources. Étant donné que le présent article prône une approche globale de la qualité, on y insiste sur 1) la mesure des autres critères de qualité et 2) l'importance des clients internes. Ainsi, tandis qu'il sera éventuellement très difficile de mesurer la pertinence du point de vue de l'utilisateur des données (notamment en ce qui concerne les enquêtes

gouvernementales), il est néanmoins possible de l'apprécier du point de vue du client interne en se demandant, par exemple, si les manuels de formation de l'intervieweur sont faciles à utiliser, si les répondants comprennent bien les questions, si les questions sont plus faciles à repérer et à quantifier de façon utile. Il est relativement facile de déterminer si les délais de publication du produit final d'une enquête, soit les données publiées, ont été respectés. Il sera de même possible d'établir des mesures des délais d'exécution ayant une incidence directe sur la qualité de l'enquête en tenant compte du client interne à chaque étape. L'arrivée dans les bureaux régionaux du matériel didactique servant à la formation des intervieweurs à un moment qui permet au formateur de bien préparer les séances de formation constitue un exemple de l'importance du respect des délais en regard du client interne. Lorsqu'on établit des mesures de la justesse, on insistera sur celles qui peuvent servir d'approximations de l'erreur totale d'enquête ou de l'erreur quadratique moyenne. Cela s'explique par deux raisons. D'abord, il est évidemment extrêmement difficile de mesurer l'une ou l'autre de ces données. En second lieu, le système d'enquête engendre déjà de nombreuses approximations de la justesse que l'on peut utiliser et analyser pour mesurer l'évolution de la justesse de l'enquête. En dernier lieu, quiconque entreprend des démarches pour améliorer la qualité d'une enquête doit connaître avec précision les ressources supplémentaires qu'il faudra engager à ce titre. Souvent, lorsqu'on lui propose une amélioration de la qualité, le gestionnaire d'enquête aura comme première question: "Combien est-ce que cela coûtera?". Pour pouvoir lui donner une réponse complète, il est indispensable de connaître les coûts actuels de l'enquête, non seulement globalement, mais aussi les coûts ventilés par source d'erreurs et par étape de l'enquête.

En partant de la structure des erreurs d'une enquête, on pourra mesurer la justesse de cette dernière de manière plus complète et mieux connaître les mesures dont la quantification s'impose. Aux fins du présent article, les erreurs d'enquête sont classées en erreurs d'observation et en erreurs de non-observation. Les premières comprennent les erreurs découlant de l'interaction entre le questionnaire, l'intervieweur et le répondant, les erreurs de contrôle, les erreurs de codage, etc. Les secondes comprennent le taux de non-réponse, l'erreur d'échantillonnage, la couverture, les quantités et taux d'imputation, etc. Il est possible que certaines des mesures de la qualité se rattachent aux deux sources d'erreurs: il peut alors être utile de déterminer dans quelle mesure chacune des deux sources a influé sur la qualité. On peut, par exemple, reviser les estimations initiales après leur publication. Ces révisions peuvent être rendues nécessaires par des erreurs d'observation, par la consignation de valeurs erronées ou par la non-observation, c'est-à-dire du fait que certains questionnaires aient été reçus trop tard pour être pris en considération lors de la formulation de l'estimation initiale et que les valeurs imputées aient différé suffisamment de la valeur juste pour qu'une révision soit nécessaire. Ce dernier exemple nous amène directement à examiner les étapes de l'enquête.

En procédant de manière quelque peu arbitraire, on peut décomposer l'enquête en une étape d'élaboration du plan d'échantillonnage et de sélection de l'échantillon, une étape de collecte des données, une étape de contrôle et d'analyse et une étape d'estimation et de tabulation. À chacune de ces étapes, le gestionnaire d'enquête peut relever d'importantes mesures de la qualité en examinant les critères de qualité et la structure des erreurs de l'enquête. La figure 2 présente une partie d'une matrice fictive correspondant à l'étape de la collecte des données.[2]

[2] La matrice complète comprend des colonnes correspondant aux erreurs de non-observation et aux erreurs d'observation relatives aux trois autres étapes de l'enquête.

Figure 2. Matrice de la qualité relative à l'étape de la collecte des données, période i

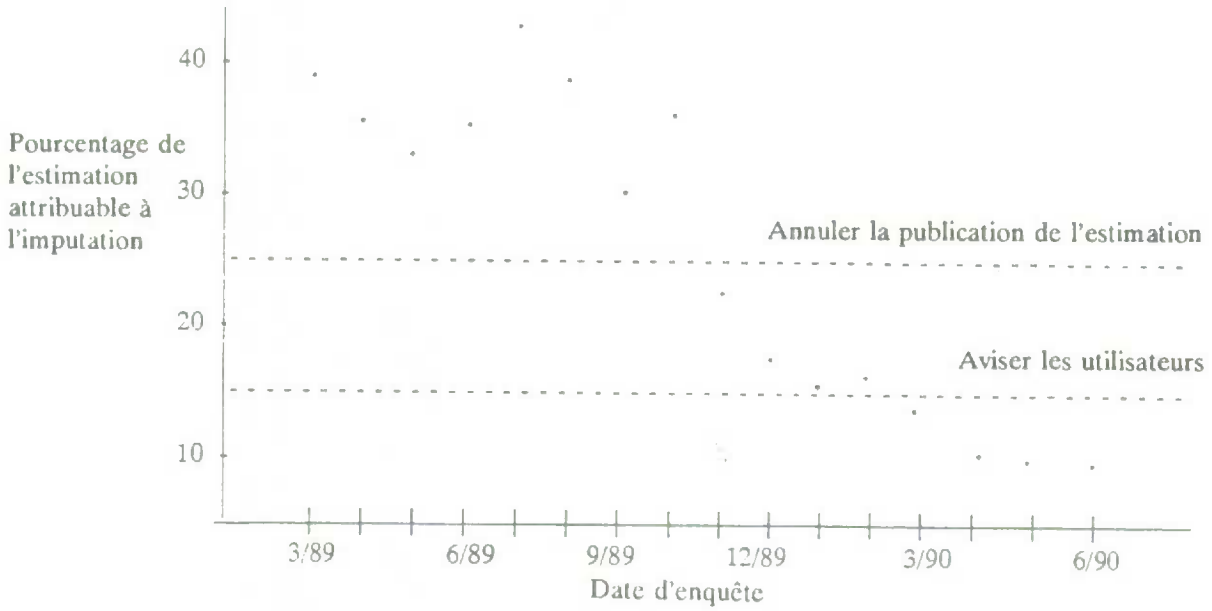
Exemple portant sur l'étape de la collecte des données
(éléments choisis)

	Structure des erreurs			
	Non-observation	Tolérance	Observation	Tolérance
Critères de qualité				
Respect des délais			Quartiles correspondant aux dates de retour du questionnaire	20e jour du mois
Pertinence			Pourcentage de chacune des questions nécessitant une révision	K pour cent
Justesse	Règle de non-réponse	NR pour cent	Estimation de l'erreur systématique des réponses	
	Pourcentage de l'estimation attribuable à l'imputation	I pour cent		
Ressources	Coût du suivi des non-réponses			

Certaines des cellules de la figure 2 renferment des données qui peuvent être obtenues directement à l'aide du système d'enquête. Il s'agit de renseignements tels les quartiles correspondant aux dates de retour des questionnaires, au pourcentage de révision que nécessite chaque question, au taux de non-réponse et au pourcentage de l'estimation attribuable à l'imputation. L'obtention de certains des renseignements peut nécessiter des efforts supplémentaires ou une étude spéciale, comme une estimation de l'erreur systématique des réponses.

Bien que la matrice offre des renseignements sur des éléments importants pour une date d'enquête donnée, les données sont plus utiles si elles sont présentées graphiquement de façon évolutive. Par exemple, la figure 3 illustre le pourcentage d'imputation relative à une estimation particulière dans le cadre d'une enquête auprès des entreprises.

Figure 3. Pourcentage de l'estimation attribuable à l'imputation dans le cadre d'une enquête auprès des entreprises



Le graphique chronologique qui figure ci-dessus présente deux limites de tolérance. Ainsi, si le pourcentage d'imputation est supérieure à vingt-cinq pour cent, la publication de l'estimation est annulée. Par contre, s'il se situe entre dix et vingt-cinq pour cent, l'utilisateur des données en est avisé au moyen d'une note infrapaginale dans la publication. (La tendance à la baisse de la proportion de l'estimation attribuable à l'imputation reflète les efforts particuliers déployés par les spécialistes en vue d'obtenir la collaboration des répondants.)

S'il élabore une matrice complète de mesures de la qualité et s'il analyse les représentations graphiques des données, le gestionnaire d'enquête pourra 1) relever les secteurs dans lesquels des études spéciales s'imposent pour obtenir de plus amples renseignements au sujet de la qualité générale de l'enquête, 2) effectuer une analyse de Pareto afin d'optimiser l'utilisation des ressources, 3) rendre compte de la qualité actuelle de l'enquête, 4) mettre sur pied des moyens d'améliorer l'enquête (et d'en mesurer l'amélioration).

4. DÉMARCHES ULTÉRIEURES

Tandis que le présent article dresse le cadre dans lequel doit s'inscrire une approche globale de la gestion de la qualité des enquêtes périodiques, il est clair qu'il reste encore beaucoup à faire avant qu'une telle approche puisse être mise en oeuvre. L'orientation future des travaux sur le sujet est donc toute tracée. On procède actuellement à la collecte des données nécessaires à l'élaboration d'une matrice pour les Current Industrial Reports du Census Bureau. Il reste à résoudre diverses questions, y compris le choix des enquêtes particulières pour lesquelles des matrices seront élaborées ainsi que la définition des éléments devant figurer dans ces matrices.

BIBLIOGRAPHIE

Brooks, C.A., et Bailar, B.A. (1978). An Error Profile: Employment as Measured by the Current Population Survey. Statistical Policy Working Paper Number 3. Washington, D.C.

Deming, W.E. (1986). Out of the Crisis. MIT, CAES, Cambridge, MA.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley & Sons.

Hansen, M.H., Hurwitz, W., et Bershad, M. (1961). Measurement Error in Censuses and Surveys. *Bulletin of the International Statistical Institute*, 38, 2.

Jabine, T. (1990). SIPP Quality Profile. U.S. Bureau of the Census, Washington, D.C.

Juran, J.M. (1964). Managerial Breakthrough. McGraw Hill, New York.

Mulry, M.H., et Spencer, B. (1990). Total Error in Post Enumeration Survey (PES) Estimates of Population: The Dress Rehearsal Census of 1988. Annual Research Conference Proceedings, U.S. Bureau of the Census, Washington, D.C.

Statistique Canada (1987). Lignes directrices concernant la qualité. Ministère des Approvisionnements et Services, Ottawa.

U.S. Energy Information Administration (1989). Standards Manual. Washington, D.C.

SESSION 2

Intégration des données

Recueil du Symposium de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

INTÉGRATION DES DONNÉES ÉCONOMIQUES: AVANTAGES ET PROBLÈMES

M. Colledge[1]

RÉSUMÉ

"Une question fondamentale qui se pose est celle de savoir si les recensements et les enquêtes intégrés constituent un instrument puissant et souple ou si "l'approche intégrée" n'a pas plutôt assujetti le Bureau à des complexités ou à des raffinements qui ont pour effet conjugué de rendre impossible la collecte et le traitement des données en peu de temps et à peu de frais", R. Cameron, statisticien en chef de l'Australie, 1981 (traduction).

Dans cette communication, nous examinons les avantages et les problèmes qui découlent de l'élaboration et de l'utilisation d'une approche intégrée en statistique économique. En bref, les avantages tiennent à ce qu'on peut juxtaposer des données provenant de diverses sources (et, ce faisant, augmenter leur valeur) et vérifier leur qualité. Les problèmes résultent du fait que, en tentant d'uniformiser les concepts et les procédures pour toutes les sources de données indépendamment des besoins des utilisateurs, on risque de simplifier à l'extrême et de déformer les réalités économiques. Cette thèse est illustrée à l'aide d'exemples provenant de Statistique Canada et de l'Australian Bureau of Statistics.

MOTS CLÉS: Données économiques; intégration.

1. INTRODUCTION

L'intégration des données et, en particulier, celle des données économiques, est un objectif largement encouragé par les organismes statistiques nationaux. Par exemple, le Comité des méthodes et des normes de Statistique Canada (1990) a déclaré que:

> "L'intégration améliore le contenu informationnel et la qualité des données et elle fournit un moyen, très souvent le seul moyen, de découvrir les lacunes dans nos séries de données et dans nos cadres de travail statistiques" (traduction).

Dans une étude récente dont le résultat final a été le remaniement en cours du programme de statistiques économiques de l'Australie, G. Sarossy (1987) a écrit:

> "L'intégration est et continuera d'être un des objectifs principaux de l'Australian Bureau of Statistics. L'appareil statistique requis pour atteindre cet objectif est un système intégré de recensements et d'enquêtes qui comprend toutes les branches d'activité industrielle et toutes les activités économiques" (traduction).

Dans une étude de l'appareil statistique des États-Unis, le rapport Bonnen (1981) a commencé sa discussion de l'intégration avec les remarques suivantes:

> "Historiquement, les statistiques ont été instituées à la suite de tentatives visant à répondre à la demande de secteurs spécialisés particuliers, par. ex., l'agriculture, la population, la santé, le

---

[1] M. Colledge, Division des méthodes d'enquêtes-entreprises, 11e étage, édifice R.H. Coats, Statistique Canada, Ottawa (Ontario) Canada K1A 0T6.

> travail, la fabrication et le commerce. L'importance accrue des comptes économiques nationaux et, plus récemment, les indicateurs sociaux ont aidé à élaborer une approche plus globale relativement aux programmes statistiques mais, dans l'ensemble, il y a encore une forte tendance à répondre aux besoins détaillés de secteurs particuliers à mesure qu'ils surgissent. Cependant, de plus en plus, les séries statistiques ne sont pas utilisées seules -- elles sont employées avec d'autres statistiques dans une gamme étendue d'analyses transversales. Il est donc essentiel de disposer d'un produit intégré, pour ce qui est des concepts, des mesures, de la classification et des méthodes analytiques" (traduction).

Plus loin, le rapport déclare que: "L'intégration peut, entre autres choses, entraîner des améliorations dans la qualité des données, dans leur pertinence à des fins de politiques et elle peut aider à réduire le fardeau des répondants" (traduction).

Selon les propos de Goldberg, alors qu'il était statisticien en chef adjoint à Statistique Canada, l'intégration devrait être: "... une philosophie directrice qui pénètre les actions et les politique d'un bureau de statistique" (traduction). Par la suite, à titre de directeur du bureau de statistique des Nations Unies, Goldberg a déclaré que la fonction d'intégration:

> "... devrait fournir une conscience globale, maintenir un système d'autocontrôle face à des pressions qui diffèrent provenant des diverses sections, favoriser la planification et la réalisation de projets interdisciplinaires et surmonter les obstacles, réels ou imaginaires, entre les diverses parties de l'organisme. Elle devrait assurer que les méthodes, les définitions, les classifications et les concepts communs sont non seulement disponibles mais effectivement appliqués dans les diverses divisions et sections de sorte que les séries statistiques représentent des éléments d'un cadre intégré et qu'elles sont aussi cohérentes et comparables que possible" (traduction).

Comme ces citations le montrent, les efforts visant à réaliser l'intégration ne sont pas du tout récents. Dans certains organismes, ils remontent à plusieurs années. L'intégration était, par exemple, le thème principal des recensements économiques intégrés de 1968-1969 en Australie qui ont constitué "... une restructuration importante visant à augmenter considérablement l'utilité et la comparabilité des genres de statistiques que l'on recueillait et publiait déjà " (ABS, 1970) (traduction). À Statistique Canada, elle était considéré comme "... une approche de base à l'organisation et l'emmagasinage des statistiques" (Gigantes et coll., 1970) (traduction).

En résumé, l'intégration est un mécanisme très puissant pour améliorer la qualité, d'où l'inclusion de la présente communication dans le symposium de 1990. L'intégration peut accroître la valeur et augmenter la pertinence des produits statistiques et elle offre la possibilité d'améliorer l'exactitude des données recueillies et de réduire le fardeau des répondants ainsi que les coûts de traitement. Toutefois, comme sa longue histoire le montre, l'intégration est un but insaisissable. Sa mise en oeuvre est loin d'être facile. Les résultats peuvent de fait être des coûts additionnels, la désuètude ainsi que l'ajustement forcé des données à un cadre conceptuel trop étroit. R. Cameron, le statisticien en chef de l'Australie, a fait allusion à ce problème quand il a demandé: "... si les recensements et les enquêtes intégrés constituent un instrument puissant et souple ou si l'approche intégrée n'a pas plutôt assujetti le Bureau à des complexités ou à des raffinements qui ont pour effet conjugué de rendre impossible la collecte et le traitement des données en peu de temps et à peu de frais" (traduction).

Dans cette communication, nous examinons les avantages et les problèmes qui découlent de l'élaboration et de l'utilisation d'une approche intégrée en statistique économique. Divers genres d'intégration et les avantages éventuels qui leurs sont associés sont décrits en détail dans la section 2. Les activités générales associées à l'intégration sont décrites dans la section 3. Certaines réalisations réelles portant sur l'intégration sont décrites dans la section 4 et dans la section 5 on résume les leçons apprises.

2. GENRES D'INTÉGRATION ET LEURS AVANTAGES

2.1 Remarques préliminaires

Un des problèmes reliés à la discussion de l'"intégration" est le fait que le terme lui-même n'est pas défini avec précision. Même dans le contexte relativement restreint de la collecte et de la production de statistiques économiques, l'intégration présente de nombreux aspects qui ont diverses significations. Les paragraphes ci-dessous portent sur cette question.

On peut considérer qu'un programme de statistiques économiques est un système qui obtient des données brutes de diverses sources et qui les transforme en données produites pour répondre aux besoins d'une gamme de clients et qui fonctionne dans un cadre conceptuel plus ou moins bien défini. En fonction de ce modèle simple, on peut déterminer et décrire quatre genres d'intégration: intégration des concepts, intégration des données brutes, intégration du traitement des données et intégration des données produites. Bien que ces aspects de l'intégration soient interdépendants, il est commode de les considérer séparément. L'intégration des concepts constitue le point de départ logique, car elle étaye tous les autres aspects, mais les efforts principaux visant à réaliser l'intégration tournent autour des données produites, c'est pourquoi nous commençons la description avec cet aspect.

2.2 Intégration des données produites

L'intégration des données produites laisse supposer beaucoup plus que le simple fait de rassembler en une seule base de données en sortie ou sous forme d'un seul véhicule de dissémination des ensembles de données recueillis séparément. Elle laisse supposer que l'on peut relier les ensembles de données et que l'on peut tirer de ces ensembles pris conjointement des renseignements qui ne peuvent être fournis par l'un quelconque de ces ensembles de données pris séparément. On peut donc caractériser et quantifier l'intégration des données comme l'augmentation du contenu informationnel d'ensembles de données combinés par rapport à la somme du contenu informationnel de chacun d'entre eux.

Il n'est pas difficile de montrer pourquoi l'intégration peut produire de la synergie. Considérons le contenu informationnel d'un ensemble de données de disons n champs où, pour simplifier, on suppose que chaque champ particulier est aussi informatif qu'un autre. Il est évident que le contenu informationnel global augmente avec n. L'augmentation sera probablement plus que linéaire à cause des renseignements incorporés dans les champs pris comme groupes. Par exemple, si l'on suppose que tous les champs de données pris individuellement et que toutes les paires de champs ont autant de valeur, alors leur contenu informationnel est proportionnel à $n(n+1)/2$. Il s'ensuit que si deux ensembles de données avec m et n champs sont intégrés alors, sous les mêmes hypothèses, le contenu informationnel combiné sera environ $(m+n)^2/(m^2+n^2)$ fois la somme de chacun des contenus informationnels. Si des groupes plus considérables de champs, disons des groupes qui en comptent 3 ou 4, sont considérés informatifs, alors les gains attribuables à l'intégration seront encore plus considérables.

L'exemple qui précède suppose que l'on peut relier deux à deux ou par groupes plus considérables les champs dans les deux ensembles de données. Dans la mesure où on ne peut le faire, le potentiel pour un gain d'information est réduit proportionnellement. Essentiellement, l'importance de l'intégration est déterminée par la mesure dans laquelle les ensembles de données respectent un cadre conceptuel commun, sujet dont on traitera plus en détail à la section 2.3.

L'intégration peut avoir lieu au niveau micro (unité) ou au niveau macro (agrégat). Plus le niveau est bas, plus le potentiel pour un gain d'information est élevé. Toutefois, il arrive souvent que l'intégration au micro niveau ne soit pas réalisable, par exemple, quand les données proviennent d'ensembles différents d'unités échantillonnées de la même population ou de populations différentes.

Les avantages que l'on peut attendre de l'intégration des données produites peuvent être résumés, en gros, dans les trois catégories suivantes.

Réponse aux besoins en données transversales

Une politique est rarement conçue en considération d'une variable unique. Elle va plutôt impliquer plusieurs dimensions, d'où la nécessité de l'intégration des données.

Plus la société est complexe, de même que plus le degré d'intervention gouvernementale est grand, plus la gamme des besoins en données statistiques est étendue. La façon traditionnelle de répondre à de nouveaux besoins consistait à établir de nouvelles collectes de données; c'est de cette façon que les programmes en statistique économique ont crû, petit à petit. Toutefois, dans le monde actuel, où les ressources additionnelles sont rares, il arrive souvent que l'élaboration de nouveaux véhicules de collecte ne soit pas réalisable du point de vue économique. Il reste la possibilité de tirer plus de renseignements des processus de collecte existants. Pour ce faire, l'intégration des données produites est un processus important. Les gains auxquels on peut s'attendre se retrouvent dans la réponse aux besoins qui recoupent les ensembles de données existants. Autrement dit, l'intégration est un outil qui permet à un organisme de réorienter la mise en valeur de ses sorties, antérieurement axées vers le processus, pour des données produites orientées vers les produits ou vers les clients.

Par exemple, l'intégration des données sur la production à celle des finances peut donner des renseignements sur la productivité qu'aucun des deux ensembles de données pris séparément ne pouvait fournir - la connaissance de la rentabilité des activités d'investissements individuelles est essentielle à une bonne allocation des ressources. Pour l'analyse et l'élaboration de politiques, il peut être souhaitable d'intégrer des données sur les importations avec des données sur la production intérieure, les dividendes versés avec les profits, l'investissement étranger avec les dépenses en capital, etc. L'intégration dans le temps, c.-à-d. la création d'ensembles de données longitudinaux, constitue une autre classe particulière d'intégration.

Considéré comme un client, le Système de comptabilité nationale exige des données relatives à toute la gamme de sorties de produits. Les données qui ne sont pas intégrées dans le cadre du programme de statistiques économiques doivent, de fait, être combinées à l'intérieur des comptes nationaux. Du point de vue de la maximisation du gain de renseignements, on est mieux de réaliser l'intégration comme partie intégrante du processus de production des données plutôt que par la suite dans les comptes nationaux où il y a moins de possibilités de faire une combinaison et une comparaison au micro niveau.

Amélioration de la cohérence et de l'exactitude

Ce n'est que lorsqu'on tente d'intégrer des données produites que l'on vérifiera leur cohérence. Le repérage des incohérences est une étape en vue d'améliorer la pertinence et l'exactitude des données. De plus, si l'organisme ne relève pas les incohérences, il se peut bien que ses clients le fassent. Le manque de cohérence peut découler de différences dans les unités, dans les systèmes de classification ou dans les définitions des données élémentaires, c.-à-d. dans la base conceptuelle sous-jacente, ou il peut indiquer qu'il y a des erreurs dans les ensembles de données particuliers. Cette dernière possibilité montre que l'intégration peut faire ressortir des erreurs qui ne seraient pas détectées autrement. Cela constitue la base du contrôle au macro niveau.

Détermination des lacunes et des répétitions dans les données

La découverte de lacunes ou de répétitions dans les données produites constitue une autre conséquence possible de l'intégration. Des renseignements de ce genre aident à déterminer où et quand les ressources relatives à un programme doivent être réorientées.

2.3 Intégration des concepts

Comme on l'a déjà mentionné, le fait de combiner des ensembles de données n'implique pas nécessairement qu'il y a intégration. L'importance de l'intégration qui peut être réalisée dépend de la mesure dans laquelle les ensembles de données sont construits à partir des mêmes "composantes d'information", c.-à-d. dans un cadre conceptuel commun.

On peut définir un cadre conceptuel approprié aux enquêtes économiques en fonction de quatre composantes de base:

- un modèle, c.-à-d. une collection de données économiques élémentaires ainsi que les rapports qui existent entre elles;
- des ensembles d'unités statistiques types, à propos desquelles il faut recueillir des données auprès des répondants et sur lesquelles les clients ont besoin de données;
- des systèmes types pour classifier les unités selon la branche d'activité, la région géographique et la taille;
- des ajustements et des définitions normalisés pour les données élémentaires, exprimés en termes compris par les répondants et acceptables pour les clients.

Modèle économique

Pour les statistiques économiques, le Système de comptabilité nationale fournit un cadre conceptuel complet qui n'a pas besoin d'être beaucoup élargi (Statistique Canada, 1989). De plus les principes de comptabilité généralement acceptés en ce qui concerne les déclarations des revenus et dépenses, les bilans, etc., fournissent un cadre partiel peu rigoureux pour les pratiques comptables des entreprises. Dans un tel cadre, une entreprise peut maintenir un certain nombre de comptes qui ne sont pas intégrés et qui peuvent ne pas être consistents les uns avec les autres ou avec le modèle de comptabilité nationale. Ceci doit être pris en considération lors de l'intégration des données; il en sera question ultérieurement.

Unités statistiques types

Quand on établit des unités statistiques types il faut tenir compte de trois exigences principales. Premièrement, les unités devraient couvrir complètement toutes les activités pour lesquelles on considère que des statistiques économiques sont requises. En particulier, les unités devraient faciliter l'intégration des données pour toutes les entreprises et toutes les institutions qui exercent leurs activités dans des branches d'activité différentes (intégration horizontale). Deuxièmement, les unités devraient permettre la collecte de divers genres de données économiques - sur l'emploi, sur la production, sur les finances, etc. Dans les grandes entreprises, différents genres de données sont conservées par les unités opérationnelles à différents niveaux. Par exemple, les salaires et les traitements sont disponibles à partir de chacun des centres de paye, alors qu'il se peut que les intentions en matière d'investissements ne soient disponibles qu'auprès des bureaux des divisions ou des sièges sociaux. On doit donc disposer de plus d'un ensemble d'unités statistiques types; il doit y en avoir un pour chaque niveau à partir duquel on doit obtenir d'importants groupes de données. Troisièmement, il devrait être possible d'intégrer des données recueillies à différents niveaux (intégration verticale); idéalement les ensembles d'unités devraient donc être hiérarchiques pour permettre le cumul des données d'un niveau inférieur à un niveau plus élevé.

Systèmes de classification types

La classification selon les branches d'activité est une exigence pour presque toutes les statistiques économiques, bien que le niveau de détail désiré pour la branche d'activité puisse varier d'un client à l'autre. On doit donc disposer d'un ou de plusieurs système(s) de classification type(s) des branches d'activité pour chaque ensemble d'unités statistiques types. Ce système devrait avoir une structure hiérarchique afin de permettre divers niveaux d'agrégation.

De même, on doit disposer d'un système de classification géographique des unités pour répondre aux exigences des clients pour des statistiques régionales, provinciales ou pour les petites régions. Encore une fois, il est souhaitable de disposer d'un système hiérarchique afin de faciliter le cumul des données des régions géographiques plus petites aux régions géographiques plus étendues.

Définitions et ajustements normalisés pour les données élémentaires

Idéalement, il devrait exister une définition normalisée simple pour chaque concept économique - emploi, revenu brut, valeur ajoutée, etc. En pratique, il peut exister des raisons impérieuses pour permettre et même pour normaliser certaines variantes. De plus, afin de permettre l'intégration des données et puisque les entreprises

ne peuvent pas nécessairement maintenir des comptes integrés conformément au cadre du système de comptabilité nationale, il devrait y avoir un ensemble d'adjustements normalisés pour micro-données comme il en existe un pour les macro-données dans le système de comptabilité nationale.

L'intégration des concepts en un cadre normalisé apporte certains avantages directs pour ce qui est de répondre aux besoins des clients qui, eux-mêmes, ont besoin d'une base conceptuelle pour la collecte, la classification ou l'analyse de données économiques. Toutefois, les avantages que l'on peut prévoir sont surtout indirects, découlant du fait qu'un cadre conceptuel commun constitue une condition préalable pour toutes les autres formes d'intégration, et des avantages qu'elles apportent.

2.4 Intégration des données brutes

Traditionnellement, les organismes statistiques ont partagé les besoins en donnée en groupes distincts de données élémentaires, par exemple la production, le financement. Les données sont recueillies séparément pour chacun de ces groupes. Les parties distinctes qui existent et qui définissent des "enquêtes" particulières sont un produit des pratiques comptables des entreprises et de l'élaboration historique du programme de statistiques économiques.

L'intégration des données brutes présente une autre façon d'aborder le problème. Les besoins en données sont groupés par répondant plutôt que par genre de données, c.-à-d. que l'accent est mis sur les répondants plutôt que sur les enquêtes. Ainsi, au lieu de, disons, une douzaine d'enquêtes dans le cadre desquelles on communique séparément avec des entreprises, les données qui répondent aux besoins de toutes ces enquêtes sont obtenues de chacun des répondants au moyen d'un véhicule de collecte intégré. C'est dans le cas des grosses entreprises avec une structure organisationnelle complexe que cette façon de procéder est la plus efficace.

L'intégration des données brutes trouve un proche équivalent dans l'intégration des données produites, en ce sens qu'elle oblige les entrées de l'organisme a être orientées vers les répondants plutôt que vers les processus. Ce faisant, l'intégration encourage l'établissement et l'utilisation de mécanismes de collecte qui sont bien adaptés aux pratiques comptables des répondants, reflétant la disponibilité des données et la facilité d'accès. Cela peut produire un certain nombre d'avantages. Premièrement, il est plus probable que l'organisme pourra s'assurer que les exigence en matière de données imposées par ses enquêtes sont cohérentes et qu'il n'y a pas de répétition. Deuxièmement, les mécanismes de collecte des données peuvent être choisis de façon à minimiser le fardeau des répondants. Troisièmement, il est plus probable que les données déclarées par un répondant seront compatibles et cohérentes, étant limitées à cet égard seulement par le niveau d'intégration et de cohérence des comptes de chaque répondant.

2.5 Intégration du traitement des données

L'intégration du traitement des données implique l'utilisation de méthodes génériques et de systèmes généralisés. Comme exemples de méthodes génériques citons la définition et l'utilisation de plans de sondage types parmi lesquels on choisirait un plan de sondage particulier pour toute nouvelle enquête, la définition et l'utilisation d'une méthode normalisée pour le traitement de la non-réponse, etc. Les systèmes généralisés peuvent être des méta-systèmes à partir desquels on peut créer un système spécifique approprié à toute enquête particulière, ou ils peuvent comprendre des programmes à usages multiples avec suffisamment d'options pour fournir les variations requises pour une gamme d'enquêtes.

L'intégration du traitement présente trois avantages évidents. Premièrement, si l'on utilise constamment des méthodes et des systèmes types, on en fait l'essai et la vérification complets et ils deviennent moins sujets à des erreurs. Deuxièmement, l'existence d'un ensemble de méthodes et de systèmes types réduit les coûts et le temps requis pour élaborer de nouvelles enquêtes. Troisièmement, l'utilisation de procédures normalisées facilite l'intégration des données brutes et des données produites.

3. ACTIVITÉS ASSOCIÉES À L'INTÉGRATION

Un certain nombre d'activités, qui en prouvent l'existence, sont requises pour réaliser l'intégration sous ses diverses formes. Elles sont résumées brièvement dans les paragraphes ci-après. Plus un organisme statistique s'implique dans ces activités, plus son engagement en matière d'intégration est important.

Cadre conceptuel. Comme on l'a déjà fait remarquer, un cadre intégré d'unités, de méthodes de classification et de définitions de données élémentaires est une condition préalable pour réaliser l'intégration. Un tel cadre doit constamment être amélioré pour suivre les changements économiques et technologiques.

Registre des entreprises. Un registre des entreprises qui fournit des ensembles classifiés d'unités et des renseignements sur les personnes-ressources pour toutes les enquêtes économiques est le mécanisme requis pour donner au cadre conceptuel une forme opérationnelle. Comme il y a constamment création et disparition d'unités économiques, on doit absolument disposer de procédures complètes pour tenir le registre à jour. En particulier, il faut dresser régulièrement le "profil" des grosses entreprises afin de s'assurer que la compréhension que l'organisme a des unités appropriées ainsi que des mesures relatives à la collecte des données est actuel.

Mesures coordonnées pour la collecte des données. La coordination des mesures de collecte des données pour toutes les enquêtes, du moins pour les grosses unités qui font partie de tous les échantillons d'enquête, constitue la base de l'intégration des données brutes.

Méthodes et systèmes génériques. L'élaboration et la tenue à jour de méthodes et de systèmes génériques constitue la base de l'intégration du traitement.

Analyses transversales. Par définition, les analyses transversales impliquent l'intégration de données produites. Quand l'organisme statistique effectue de telles analyses, ces dernières l'aident à évaluer la pertinence, la cohérence et l'exactitude des sorties.

4. RÉALISATIONS DANS LESQUELLES L'INTÉGRATION JOUE UN RÔLE

4.1 Remarques préliminaires

L'objectif de la présente section est de montrer les difficultés rencontrées quand on cherche à réaliser l'intégration. Trois réalisations, à Statistique Canada (SC), dans lesquelles l'intégration joue un rôle sont décrites. Trois autres exemples sont tirés de l'expérience de l'Australian Bureau of Statistics (ABS) afin de montrer tant les similitudes que les différences, dans la façon de procéder, par rapport à la situation à Statistique Canada.

Pour les deux organismes, les années 1960 ont marqué le début d'un effort important visant à réaliser l'intégration. Avant cette date, les programmes de statistiques économiques avaient pris de l'ampleur surtout à la suite de l'ajout de nouvelles enquêtes, souvent indépendamment les unes des autres, en réponse à des demandes particulières. Le Système de comptabilité nationale a été une force d'intégration au sein des deux organismes, mais ce fut virtuellement la seule. Il y avait peu de coordination globale. L'accent était mis sur les processus d'enquête plutôt que sur les produits. Par exemple, on effectuait un contrôle intensif des données au micro-niveau, alors que le contrôle au macro-niveau ainsi que la préoccupation relativement à la compatibilité pour des ensembles de données liés étaient très limités. Ce n'est que dans les comptes nationaux qu'il était probable que des données produites particulières allaient être rassemblées.

Dans les paragraphes ci-après on décrit les tentatives visant à réaliser l'intégration au cours des quelques vingt dernières années. Toutes les réalisations relatives à l'intégration pendant cette période ne sont pas incluses, et la description ne porte que sur les aspects des réalisations qui se rattachent spécifiquement à l'intégration, les détails complets n'étant pas fournis.

4.2 Remaniement du registre des entreprises: SC (1970-1975)

Avant le remaniement, on tirait les données relatives aux bases de sondage, pour les enquêtes économiques réalisées par Statistique Canada, de fichiers principaux des enquêtes qui étaient plutôt indépendants les uns des autres et on utilisait, à des degrés divers, les données qui pouvaient être extraites d'une liste d'entreprises dont la tenue à jour était centralisée. Les objectifs du remaniement relatifs à l'intégration étaient de remplacer la liste centrale et la majorité des fonctions des fichiers principaux des enquêtes existantes par un registre des entreprises qui assurerait une couverture complète et sans répétition (Sunter, 1971). Pour répondre aux besoins en données relatives aux bases de sondage dans le cas de toutes les enquêtes économiques, le nouveau registre devait fournir des ensembles d'unités statistiques, classés selon la Classification des activités économiques de 1970, avec des renseignements sur la personne-ressource initiale pour chaque unité. La tenue à jour du nouveau registre devait être basée, en grande partie, sur un traitement centralisé de données administratives tirées du système de retenues sur la paye de Revenu Canada. En résumé, les objectifs comprenaient l'introduction d'ensembles normalisés d'unités et un système normalisé de classification des branches d'activité ainsi que l'intégration des méthodes et des systèmes de traitement des données sur les bases de sondage.

Le remaniement a réussi au point de créer un nouveau registre des entreprises composé d'unités normalisées avec des renseignements sur la classification et sur les personnes-ressources qui pouvait être tenu à jour à l'aide de données sur les retenues sur la paye. L'utilisation de données tirées de ce registre a entraîné des améliorations importantes dans la couverture fournie par les bases de sondage. Toutefois, pour la majorité des opérations relatives aux enquêtes on a conservé les fichiers principaux correspondants; dans certains cas, ces fichiers ont été groupés en fichiers principaux "divisionnaires". Au cours des dix années suivantes, on a commencé à utiliser le registre des entreprises comme source de nouvelles unités pour des enquêtes, mais pas comme source principale pour la production de bases de sondage. Pour certaines enquêtes, les bases de sondage ont continué d'être fondées directement sur les fichiers de l'impôt sur le revenu. Par conséquent, on n'a jamais réalisé toute l'intégration envisagée.

Les problèmes rencontrés ainsi que les raisons mentionnées pour expliquer pourquoi le nouveau registre n'a pas été adopté en entier, se rapportaient à la qualité sous tous ses aspects - pertinence, exactitude, actualité et coût. Le nouveau registre renfermait des nombres assez élevés d'unités inactives ou hors du champ des enquêtes; de nombreuses unités étaient mal classifiées au niveau détaillé des branches d'activité; la mise à jour du système du registre était lente et lourde, et les coûts reliés à l'adoption de systèmes d'enquête ou au remplacement des systèmes existants étaient élevés.

Avec un certain recul, il y a peut-être trois leçons à tirer de cette expérience. Premièrement, l'utilisation de concepts, de méthodes et de systèmes intégrés doit être mise en application. Sans une telle décision et une surveillance constante, la multitude de difficultés particulières, locales et les sacrifices d'optimums locaux qui sont nécessaires, l'emporteront sur l'effort global en vue de réaliser l'intégration. Deuxièmement, on peut s'attendre à ce que l'intégration prenne beaucoup de temps. Il faut s'attaquer aux nombreuses difficultés locales, bien qu'on ne doive pas les laisser dominer. Troisièmement, il est essentiel que les délais d'exécution des traitements effectués par ordinateur soient très courts, même si cela veut dire qu'on doit sacrifier un peu l'aspect fonctionnel.

4.3 Projet d'intégration: SC (années 1970)

Le projet avait pour objectif d'intégrer toute la collecte des données pour les plus grosses entreprises du pays, qui réalisent près de 50% de toute la production économique.

Même si l'on a accordé beaucoup de temps à ce projet, on n'a jamais été près d'atteindre l'objectif visé et le projet a été abandonné. Avec un certain recul, il est évident que le but d'intégrer toute la collecte des données était trop ambitieux, compte tenu des ressources limitées, du point de vue de l'expertise en comptabilité, du personnel de bureau et de la puissance de traitement, consacrées au projet, et du fait que les entreprises elles-mêmes n'ont pas nécessairement des comptes intégrés. Le succès aurait été plus probable si l'on avait cherché à intégrer moins d'entreprises et d'enquêtes, ou si l'on avait engagé plus de ressources sous forme d'experts-conseils.

4.4 Projet de remaniement du registre des entreprises, Phase 1: SC (1985-1990)

Le projet était une entreprise majeure, qui avait un effet sur presque toutes les principales enquêtes économiques. Au début, on avait prévu qu'il durerait trois ans, mais sa réalisation a été étendue à six ans. A l'origine, c'est la nécessité de s'attaquer à certains problèmes sérieux relatifs aux données, qui surgissaient quand on intégrait des sorties du programme des enquêtes économiques à l'intérieur du système des comptes nationaux, qui a motivé la réalisation de ce projet.

Le projet visait principalement à normaliser et à intégrer les systèmes et les données, comme Colledge (1987) l'a écrit:

> "L'essentiel du projet est le remaniement des concepts, des procédures et des systèmes utilisés pour fournir les données sur les bases de sondage et pour employer les données fiscales. Le projet présente, de plus, une occasion d'examiner tout le programme des enquêtes économiques et d'élaborer des systèmes et des procédures polyvalents" (traduction).

Le remaniement a commencé où les efforts d'intégration antérieurs avaient cessé. On envisageait (Cain et coll., 1984) la création d'une nouvelle fonction et d'une nouvelle base de données pour le registre des entreprises. La fonction visait à:

> "... appuyer une gamme complète de services, y compris la fourniture de listes classées d'unités statistiques et déclarantes, la tenue à jour de ces données à l'aide de sources administratives et autres, la fourniture de normes, de lignes directrices et de procédures, l'échantillonnage dans des univers d'enquête et la production de listes d'adresses pour les enquêtes, ainsi que la coordination de l'échantillonnage de données fiscales et l'acquisition, le traitement et l'utilisation de telles données pour remplacer les données d'enquête" (traduction).

La nouvelle base de données devait servir à stocker, à tenir à jour, et donner accès aux données pour les bases de sondage ayant rapport aux enquêtes économiques des enregistrements à l'aide de trois genres d'enregistrements:

> "... enregistrements originaux qui renferment des données administratives ou sur les personnes-ressources vérifiées qui doivent subir un traitement ultérieur afin d'obtenir des enregistrements statistiques normalisés. Les sources des enregistrements en question sont les populations fiscale et des retenues sur la paye que Revenu Canada nous transmet régulièrement;
>
> enregistrements statistiques normalisés, sur lesquels sont fondées les bases de sondage. Ces enregistrements forment une hiérarchie d'entreprises, de sociétés, d'établissements et d'emplacements;
>
> enregistrements d'unités déclarantes, élaborés à la suite de négociations entre les gestionnaires d'enquête et les répondants et liés aux unités statistiques appropriées à l'enquête" (traduction).

Une caractéristique de la nouvelle base de données des bases de sondage, qui reflétait un compromis pour ce qui est de l'intégration, était la rationalisation de l'utilisation des ressources consacrées à la tenue à jour obtenue en divisant la base en deux parties - une "partie intégrée" et une "partie non intégrée" (Cain et coll., 1984).

> "La partie intégrée de la base de sondage fournira une couverture complète et sans répétition de toutes les unités statistiques de grande taille, complexes ou importantes pour une autre raison. La priorité, tant pour ce qui est de la fréquence que de la précision de la tenue à jour, sera accordée à ces unités. Les ajouts ou les modifications aux unités statistiques seront basés principalement sur le contact direct, c.-à-d. par enquête ou par demande portant sur les données d'une base de sondage particulière.
>
> La partie non intégrée de la BDRC (base de données) comprendra toutes les autres unités qui, à cause de leur petite taille, ne justifient pas le coût d'un soutien détaillé. Ces unités, bien qu'elles soient les plus nombreuses, ne correspondent qu'à une petite fraction de l'activité économique représentée par la partie intégrée. Pour ce groupe, la tenue à jour sera presque entièrement automatisée, on fera alors appel à

des versions améliorées de systèmes utilisés actuellement. La partie non intégrée de la BDRC sera divisée en deux groupes d'unités statistiques qui se chevauchent: celles qui proviennent des enregistrements originaux des données fiscales et de celles tirées des enregistrements originaux sur les retenues sur la paye " (traduction).

Des fonctions devaient permettre de suivre les unités dans les deux parties dans le temps, offrant ainsi la possibilité de réaliser l'intégration dans le temps et permettant d'effectuer une analyse longitudinale.

En résumé, les objectifs de l'intégration étaient:

- de créer un nouveau système d'unités statistiques et de mesures de collecte qui seraient mieux adaptées aux pratiques comptables des entreprises;
- de créer un registre des entreprises intégré, basé sur le nouveau cadre conceptuel, avec un ensemble complet de fonctions permettant d'effectuer le stockage et la tenue à jour de données sur les bases de sondage pour les grosses entreprises et de suivre les unités administratives et statistiques dans le temps;
- d'acquérir et de charger dans le registre des entreprises des données à jour relatives aux bases de sondage, obtenues par contact direct avec les grosses entreprises ("en en établissant un profil") et à partir de sources de données administratives pour les petites entreprises;
- d'utiliser le nouveau registre des entreprises comme source de données sur les bases de sondage pour chaque enquête économique, avec des versions appropriées disponibles à chaque étape, depuis une version préliminaire au moment du choix de l'échantillon initial jusqu'à une version finale au moment de l'estimation finale;
- de rationaliser et d'étendre l'utilisation de données fiscales pour ajouter à la collecte annuelle directe de données auprès des petites entreprises ou pour remplacer cette collecte.

Ces objectifs n'ont pas encore été entièrement atteints. Un modèle complet de données sur les bases de sondage a été élaboré (Statistique Canada, 1985). Un nouveau registre des entreprises basé sur ce modèle, avec des fonctions pour stocker des données sur les bases de sondage, pour en faire un suivi longitudinal et pour en fournir a été conçu, on y a chargé les données nécessaires et on l'a mis en exploitation (Cuthill, 1990). On a repris l'élaboration de fonctions utilisées pour faire l'échantillonnage de données financières tirées des déclarations d'impôt sur le revenu et pour utiliser ces données. Cependant, jusqu'ici, le profil des plus grosses entreprises n'a pas encore été réalisé, la stratégie adoptée pour les données fiscales n'a pas encore été mise en oeuvre et, pour certaines enquêtes, on n'utilise pas le nouveau registre comme source de leurs bases de sondage. De plus, il existe certains doutes à propos de la capacité opérationnelle requise pour tenir à jour, de la façon prévue à l'origine, les données relatives aux bases de sondage pour les grosses entreprises. De plus, les frais d'exploitation du registre sont élevés, compte tenu du fait que l'on n'a pas encore tiré pleinement profit de l'amélioration au plan fonctionnel, par exemple pour le suivi longitudinal. On abordera ces problèmes dans une seconde phase de l'élaboration.

Il est évident que l'on avait énormément sous-estimé le temps requis pour spécifier et mettre en oeuvre le système et les procédures complexes qu'implique la stratégie du projet. Avec un certain recul, on voit que l'on aurait dû considérer certaines simplifications. Par exemple, il se peut que le modèle conceptuel soit plus élaboré que cela est strictement nécessaire. De plus, l'obligation de disposer d'un ensemble complet de fonctions pour effectuer le suivi longitudinal a certainement augmenté le coût tant de la mise en oeuvre que de l'exploitation, sans produire d'avantages substantiels jusqu'ici.

Il faudrait ajouter que cette description du projet ne porte que sur les objections relatives à l'intégration, pour illustrer les avantages recherchés et les problèmes rencontrés au cours de la recherche de cette dernière. Il y a eu d'autres réalisations et d'autres objectifs importants, par exemple le remaniement de deux enquêtes-entreprises et l'introduction d'un ordonnanceur de tâches automatisé pour les registres, sujets dont on ne traite pas ici.

4.5 Recensements économiques intégrés: ABS (1962-1972)

Un effort considérable dans le but de réaliser l'intégration à l'Australian Bureau of Statistics a commencé en 1962 quand on a entrepris des travaux pour intégrer les recensements et les enquêtes économiques. Cela a

finalement mené à un remaniement considérable et à l'intégration des recensements pour la période de référence 1968-1969.

Le raisonnement de base qui s'appliquait à l'approche intégrée était: "d'augmenter considérablement l'utilité et la comparabilité des genres de statistiques déjà recueillies et publiées, aux fins de l'analyse économique générale et des études de marché" (ABS, 1970) (traduction), les objectifs du remaniement étaient d'intégrer la collecte de données sur la production annuelle, c.-à-d. de la réaliser conformément à la classification des branches d'activité ainsi qu'à un cadre commun d'unités déclarantes et de concepts relatifs aux données. Les statistiques pour les branches d'activité visées par les recensements devaient être produites sans les lacunes du chevauchement dans la couverture et de façon à ce que certaines données élémentaires importantes comme la valeur ajoutée, l'emploi, les salaires et les traitements, les dépenses en capital fixe et les stocks soient cohérentes pour toutes les branches d'activité. L'intégration verticale devait être facilitée par une définition appropriée des unités statistiques. De plus, les données du recensement devaient permettre la compilation du compte de production des comptes nationaux et elles devaient fournir une base pour concevoir ou pour ajuster des échantillons, particulièrement pour les dépenses en capital et pour les stocks (des composantes importantes des estimations des comptes nationaux trimestriels des revenus et des dépenses).

Les activités d'intégration requises pour atteindre ces objectifs comprenaient (ABS, 1970):

- la définition d'unités d'entreprise à un niveau normalisé en fonction des strates dans la structure des entreprises pour lesquelles divers genres de statistiques économiques étaient requises et pourraient être recueillies, et l'élaboration de règles normalisées pour reconnaître de telles unités d'entreprise;
- la détermination d'unités types pour toutes les entreprises et leur enregistrement dans un registre intégré qui servira à la réalisation des recensements et des enquêtes;
- l'adoption d'un système commun de classification des branches d'activité qui peut être utilisé pour tous les recensements et pour toutes les enquêtes;
- la définition, en termes communs, des données élémentaires de base pour lesquelles des statistiques étaient requises pour toutes les branches d'activité;
- la révision des questionnaires en tenant compte des définitions communes;
- la révision des mesures de collecte de façon à ce que les données soient obtenues des sièges sociaux, à chacun desquels on a demandé de produire des déclarations cohérentes pour chacun des établissements visés par les recensements et pour l'entreprise.

Les activités d'intégration étaient en grande partie terminées et un cycle de collecte des données réalisée à l'aide du nouveau cadre a eu lieu pour l'année de référence qui s'est terminée en juin 1969 quand les recensements des activités minières, des activités manufacturières, des services publics, du commerce et d'autres services choisis ont été réalisés pour la première fois sur une base intégrée. Toutefois, bien que ces activités aient été bien planifiées et appuyées dans tout le Bureau, elles n'ont pas atteint leurs objectifs dans un certain nombre de domaines.

La mise à jour du registre intégré n'a jamais été réalisée complètement, même pour les branches d'activité industrielle visées par le premier cycle de collecte. Bien que la disponibilité des données ait été vérifiée, on n'a pas entièrement tenu compte des coûts d'extraction pour les unités plus petites. Cela a entraîné une augmentation, plus élevée que prévu, du fardeau des répondants et des taux de non-réponse ainsi que l'obligation de détourner des ressources de l'ABS pour les consacrer au contrôle et à l'imputation afin de régler les problèmes causés par les données manquantes ou de mauvaise qualité. Les problèmes ont été aggravés par un ensemble de contrôles stricts et par des procédures de traitement encombrantes. Il y a donc eu de longs retards au niveau de la publication et une dégradation des rapports établis tant avec les répondants qu'avec les usagers. Les données étaient trop incomplètes pour qu'on les utilise, de la façon prévue, dans les comptes nationaux.

Les effets globaux furent une décision d'abandonner l'intégration complète comme on l'avait envisagé au départ et une répugnance, qui a duré longtemps, de la part de l'organisme à entreprendre d'autres travaux d'élaboration selon les mêmes plans et à la même échelle.

Avec un certain recul, on réalise que le projet était trop ambitieux à plusieurs titres. Premièrement, la base de connaissances opérationnelles était insuffisante parce que les concepts étaient généralement nouveaux.

Deuxièmement, le Bureau avait relativement peu d'expérience dans le traitement de collections de données aussi considérables et complexes que celles qui devaient être recueillies. Le fait de réaliser les recensements simultanément et d'utiliser une méthode de collecte à deux niveaux (entreprise-établissement) a entraîné des changements importants par rapport à ce qui était fait auparavant. Troisièmement, la technologie informatique de l'époque - le matériel et le logiciel - imposait beaucoup plus de restrictions que ce n'est le cas aujourd'hui.

Les réalisations du projet relativement à l'élaboration de concepts et de procédures d'intégration étaient considérables. Elles sont décrites en détail dans un excellent article publié dans l'Australian Year Book (ABS, 1970). Sauf pour des révisions mineures apportées afin de tenir compte de changements ultérieurs dans la structure organisationnelle des entreprises et dans la technologie du traitement des données, cet article est aussi pertinent aujourd'hui que lorsqu'il a été rédigé. Il fournit un résumé très complet et très lucide des éléments et des avantages de l'intégration.

4.6 Programme de rotation des recensements et des enquêtes économiques: ABS (1971-1987)

À la suite de l'abandon des efforts visant à réaliser l'intégration complète des recensements économiques, l'Australian Bureau of Statistics a entrepris un projet sur une plus petite échelle. Ce projet avait pour objectifs d'élaborer une autre méthode, plus tolérable, de collecte des données pour la période de référence annuelle et de préciser un système de traitement approprié.

En 1973, l'équipe avait élaboré une nouvelle stratégie. Au lieu de la méthode antérieure consistant à réaliser un instantané périodique de toute l'économie, l'équipe a proposé un programme comportant la rotation de recensements et d'enquêtes. Des données devaient être recueillies chaque année sur les activités manufacturières, sur les activités minières, sur les services publics et sur l'agriculture; les enquêtes visant d'autres divisions de l'activité économique devaient être réalisées par rotation. La première collecte effectuée dans le cadre du nouveau programme de rotation a porté sur le commerce de détail pour l'année de référence 1973-1974. La couverture a été augmentée graduellement au cours des 15 années suivantes. L'agriculture a été intégrée au programme pour l'année de référence 1974-1975. Les premières enquêtes économiques sur les industries de la construction, du commerce en gros et des transports ont été réalisées pour les années de référence 1978-1979, 1981-1982 et 1983-1984 respectivement. Deux groupes choisis d'industries de services ont été inclus en 1986-1987 et 1987-1988 respectivement.

Les travaux portant sur les spécifications relatives à un nouveau système de traitement ont finalement amené l'élaboration de l'Integrated Economic Statistics Information System (IESIS) (système intégré de renseignements sur les statistiques économiques) qui est encore utilisé actuellement. Le IESIS utilise le Generalised Interrogation System (système généralisé d'interrogation) qui, bien qu'il ait été produit par le Bureau il y a vingt ans, possède un bon nombre des fonctions d'un langage et d'une base de données de quatrième génération. Un système propre à chaque recensement ou à chaque enquête est élaboré en fonction du cadre normalisé du IESIS. La méthode du IESIS est explicitement conçue pour traiter la hiérarchie entreprise-établissement, à l'aide d'une méthode par phase, selon laquelle les données relatives aux établissements sont traitées en premier.

Du point de vue des objectifs de l'intégration, l'introduction du programme de rotation était un pas en arrière qui a été imposé par les nécessités de la pratique. Il est difficile d'établir un lien entre les données pour différentes branches d'activité recueillies au cours d'années différentes. La rotation est un obstacle à l'intégration. Par contre, la couverture des branches d'activité industrielle a été élargie grâce à l'utilisation d'un ensemble normalisé de concepts et d'unités. Par ailleurs, le IESIS s'est révélé un grand succès dans le contexte de l'intégration. Il est exploité depuis douze ans et il a contribué à la normalisation des enquêtes et des recensements annuels.

4.7 Stratégie pour les statistiques économiques: ABS (1987-1992)

L'intégration constitue le thème principal de l'élaboration de l'Economic Statistics Strategy (stratégie des statistiques économiques), qui a commencé en 1987. En un sens, l'élaboration de cette stratégie a commencé où l'élaboration des recensements intégrés s'était arrêtée. Voici quelles sont les principales composantes de la stratégie par rapport à l'intégration:

- la révision du cadre des unités statistiques et de leurs classifications afin qu'ils se rapprochent davantage des structures organisationnelles et des pratiques comptables des entreprises et qu'elles aient un rapport mieux défini avec les unités fiscales;

- la révision et l'amélioration des systèmes et des procédures relatifs au registre des entreprises, premièrement, pour appuyer le nouveau cadre des unités statistiques, deuxièmement, pour établir et tenir à jour des structures plus précises et de meilleures mesures de déclaration pour les grosses entreprises et, troisièmement, pour mieux utiliser les données administratives fournies par les autorités fiscales afin d'assurer la tenue à jour des petites unités d'entreprise;

- l'élaboration et la réalisation d'une collecte et d'une diffusion annuelles, pour l'ensemble de l'économie, de données élémentaires structurelles clés permettant au moins d'obtenir l'excédent brut d'exploitation;

- l'élaboration et la réalisation d'une collecte trimestrielle, pour l'ensemble de l'économie, de données sur les dépenses d'immobilisation, sur les stocks et sur l'excédent d'exploitation;

- l'élaboration et la réalisation de systèmes et de procédures afin de produire, pour l'ensemble de l'économie, des données régionales tirées du registre des entreprises;

- l'élaboration et la réalisation de systèmes et de procédures afin de traiter des données fiscales pour remplacer la collecte directe des données ou pour ajouter à cette collecte.

Le travail d'élaboration est une entreprise considérable qui est encore en cours. Les progrès réalisés jusqu'à ce jour peuvent être résumés de la façon suivante:

On a élaboré un nouveau modèle d'unités statistiques qui, bien qu'il soit plus simple, ressemble beaucoup au modèle élaboré pendant la réalisation du projet de remaniement des enquêtes-entreprises à Statistique Canada. Ce modèle fournit une bonne base pour réaliser l'intégration, bien que l'on n'ait pas exploré à fond tous les détails relatifs aux unités de niveau plus élevé pour les données financières.

Le système du registre des entreprises a été amélioré pour refléter les nouvelles unités et les procédures de traitement améliorées. Les modifications ont pris beaucoup plus de temps que ce que l'on avait prévu à l'origine. Le profil de toutes les grosses entreprises a été réalisé en fonction des nouvelles unités et du registre des données.

Une enquête annuelle auprès des entreprises, pour toute l'économie, a été élaborée et sa deuxième année d'essai est en cours. La combinaison de cette enquête avec les enquêtes, réalisées par rotation, en cours sur les établissements a posé de nombreux problèmes opérationnels qui ont été réglés avec plus ou moins de succès. En particulier, on a intégré les mesures de collecte des données pour les grosses entreprises. On en est encore aux premières étapes de l'élaboration des procédures pour la collecte intégrée trimestrielle de données, pour la production de données régionales et pour l'utilisation de données fiscales.

4.8 Observations

Les exemples qui précèdent se rapportent aux réalisations les plus significatives qui touchent à l'intégration à Statistique Canada et à l'Australian Bureau of Statistics au cours des vingt dernières années. Cependant, il y a eu un certain nombre d'autres projets et propositions, dont certains sont encore à l'étude, dans lesquels l'intégration jouait un rôle important. A Statistique Canada, ils comprennent le Programme d'analyse longitudinale de l'emploi, le projet de développement de fonctions générales d'enquête et l'intégration des opérations des enquêtes. La stratégie informatique actuelle met l'accent sur l'intégration des installations informatiques. A l'Australian Bureau of Statistics, on a remanié et amélioré le registre des entreprises entre 1979 et 1984, on a élaboré un systèmes de contrôle polyvalent pour l'expédition et la collecte avec une version spécialisée pour les enquêtes économiques annuelles. Des options permettant d'utiliser des dictionnaires de données centralisés ont été installées sur l'ordinateur central et la stratégie informatique met, là aussi, l'accent sur l'intégration.

5. CONCLUSIONS

Si l'on définit sommairement la qualité en fonction de quatre éléments - pertinence, exactitude, actualité et coût - alors l'intégration est un moyen qui permet d'aborder chacun de ces éléments. Par exemple, on peut améliorer la pertinence en intégrant les données produites afin d'étendre la portée des analyses transversales. L'exactitude peut être accrue par l'intégration des données brutes fournies par les répondants afin d'assurer une cohérence au micro-niveau pour les enquêtes, et (ou) par l'intégration des données produites dans le but de vérifier leur compatibilité mutuelle au micro-niveau et au macro-niveau. On peut réduire considérablement le temps nécessaire pour répondre à une nouvelle demande si l'on peut atteindre le résultat visé en intégrant des données déjà produites plutôt qu'en créant un nouveau véhicule de collecte des données. Les coûts d'élaboration peuvent être un peu réduits si l'on dispose d'un ensemble intégré de systèmes et de procédures polyvalents parmi lesquels on peut choisir.

Si l'on exprime les mêmes idées dans une perspective légèrement différente, on peut considérer l'intégration comme un outil valable pour atteindre les objectifs visés par un organisme. Lors de l'élaboration de son plan pour 1991-1992, Statistique Canada a défini plusieurs objectifs prioritaires, tant d'ordre administratif que technique, que l'intégration peut aider à atteindre. Voici certains de ces objectifs: la qualité des données et des produits, l'amélioration des sorties, la recherche de l'efficience, l'équilibre des programmes ainsi que des questions relatives aux répondants.

En résumé, l'intégration sous ses diverses formes offre des avantages éventuels énormes. Toutefois, comme les exemples de la section 4 le montrent, la voie qui mène à l'intégration est jonchée d'objectifs abandonnés, de dépassements de temps et des pertes de confiance qui en découlent. La recherche de l'intégration peut entraîner des tentatives pour construire des systèmes et afin d'introduire des procédures qui dépassent l'expertise et la puissance de calcul disponibles. Il y a toujours le risque que l'imposition de concepts normalisés ou de systèmes polyvalents amènera une perte inacceptable au niveau de la flexibilité ou de l'actualité pour des sorties ou des processus particuliers.

Pour conclure, bien que l'intégration soit un but insaisissable, elle peut donner des avantages tellement intéressants que l'on doit continuer de chercher à l'atteindre. Il ne peut faire autrement que d'exister une tension entre le cadre imposé par l'intégration et les exigences pour répondre à des besoins particuliers ou pour mettre au point des processus déterminés. En l'absence de toute ligne directrice générale, les préférences locales à court terme qui s'opposent à l'intégration l'emporteront sur les objectifs globaux à plus long terme. L'intégration doit donc être promue activement par un énoncé de politique, sinon par une unité fonctionnelle. "La conscience de l'intégration", c.-à-d. le fait que le personnel soit sensibilisé aux avantages potentiels de l'intégration est indispensable.

Il se peut que la meilleurs façon de progresser consiste à réaliser une série de petites étapes, en tirant profit des occasions comme elles se présentent, avec une effort important de temps à autre. Comme exemple de la réalisation du progrès graduel, les résultats de l'intégration des données dans les comptes nationaux ou au cours des analyses transversales devraient être réintroduits systématiquement dans les enquêtes d'où sont tirées les données, en vue de déterminer et de corriger les sources d'écarts. Lors de la planification et de la mise en application des principales initiatives en matière d'intégration, il est important d'assurer un certain équilibre entre les objectifs et les restrictions pratiques. Il faut reconnaître et permettre les écarts et les exceptions. Le cadre utilisé pour l'intégration doit permette une évolution future.

REMERCIEMENT ET REMARQUE

L'auteur désire remercier Gordon Brackstone, Jacob Ryten, Nanjamma Chinnappa et George Sarossy pour leurs commentaires utiles.

Bien que la présente communication renferme des exemples tirés de l'expérience de Statistique Canada et de celle de l'Australian Bureau of Statistics, les opinions exprimées sont celles de l'auteur et pas nécessairement celles de ces deux organismes.

BIBLIOGRAPHIE

ABS (1970). Australian Year Book, Chapter 31, 1970, Canberra.

Bonnen, J. (1979). (Project Director), Improving the Federal Statistical System: Issues and Options, The President's Federal Reorganization Project for the Statistical System, Février 1981, Washington, U.S.

Cain et coll. (1984). Infrastructure Development: Objectives, Policy and Strategy, Working Paper, Business Survey Redesign Project, Novembre 1984, Statistique Canada, Ottawa.

Colledge (1987). The Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada, presented to the U.S. Bureau of the Census Third Annual Research Conference, Mars 1987.

Cuthill (1990). The Statistics Canada Business Register, Working Paper, Business Survey Redesign Project, Août 1990, Statistique Canada, Ottawa.

Gigantes, T., Fellegi, I., Goldberg, S., et Podoluk, J. (1970). Micro Data Sets, Simulation and Statistical systems, Presented to the Workshop on Micro Data Sets, Washington, U.S., Statistique Canada, Ottawa.

Sarossy, G. (1987). Review of Economic Statistics, Working Paper, Industry Division, Australian Bureau of Statistics, Canberra.

Statistique Canada (1985). Business Statistics Survey Frame Model, Working Paper, Décembre 1985, Business Survey Redesign Project.

Statistique Canada (1990). Minutes of Methods and Standards Committee, Janvier 1990, Ottawa.

Sunter, A. (1971). On the Construction and Maintenance of a Central Register for Business Surveys, Working Paper, Statistique Canada, Ottawa.

SESSION 3

Mesure de la couverture dans les recensements de la population

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

LES ÉCARTS DE COUVERTURE DANS LE RECENSEMENT DE LA POPULATION DES ÉTATS-UNIS: ÉTUDE HISTORIQUE

J.G. Robinson et H. Hogan[1]

RÉSUMÉ

Cet exposé traite des constatations sur les écarts de couverture dans les recensements effectués aux États-Unis faites à partir des estimations dont on dispose. Les estimations du taux de sous-dénombrement net, établies à partir d'études démographiques pour tous les recensements depuis celui de 1880, indiquent une baisse de ce taux chez tous les groupes à l'exception de deux: les enfants de race noire et les adultes de sexe masculin, également de race noire. D'autres études révèlent l'existence d'écarts de sous-dénombrement pour certains autres groupes, comme la population hispanique, ainsi que pour certaines régions géographiques. Les résultats du recensement de 1980 font ressortir le fait que la réduction du sous-dénombrement net est due non seulement à la baisse du nombre de personnes oubliées lors du recensement mais aussi à la hausse probable du taux de surdénombrement brut. Étant donné la nature et la répartition des erreurs de couverture, les utilisateurs de données du recensement doivent tenir compte du taux de sous-dénombrement du recensement de 1980, même s'il est extrêmement bas.

MOTS CLÉS: Sous-dénombrement; écarts de couverture

1. INTRODUCTION

Depuis le premier recensement de 1790, on reconnaît l'existence d'un problème de sous-dénombrement des personnes dans les recensements décennaux des États-Unis. Entre 1790 et les années 1960, ce problème n'a vraiment constitué un sujet de préoccupation national qu'au moment des recensements de 1870 et de 1920, alors que les taux de sous-dénombrement supposés ont été élevés.

Cependant, à partir des années précédant le recensement de 1970, la question du sous-dénombrement au recensement décennal a pris de l'ampleur. Les préoccupations suscitées par l'existence possible de taux de sous-dénombrement élevés au recensement de 1970 étaient associées aux troubles sociaux de l'époque et à la méfiance grandissante envers le gouvernement (U.S. Bureau of the Census, 1976, chapitre 7). On comprenait fort bien les conséquences des décisions de la Cour suprême en faveur du principe de «suffrage universel pur et simple» sur la redéfinition des districts des États. En outre, le gouvernement fédéral distribuait les ressources aux localités, dans le cadre de programmes tels que le General Revenue Sharing, les Urban Block Grants et le Urban Mass Transit, en se basant en partie sur les chiffres de population. Dès 1980, le sous-dénombrement au recensement était devenu un enjeu politique et juridique aussi bien que statistique - le noeud du débat consistait à déterminer s'il fallait ou non redresser les chiffres officiels du recensement en vue de la répartition des sièges, de la redéfinition des districts et d'autres utilisations. Les poursuites judiciaires et les projets de loi du Congrès demandant qu'on procède à un redressement des données ont lancé le problème du sous-dénombrement dans l'arène politique et judiciaire.

Toute cette controverse au sujet du «sous-dénombrement» pourrait donner l'impression que c'est en raison d'une détérioration de la couverture des recensements décennaux au cours des dernières années qu'il est devenu

[1] J.G. Robinson et H. Hogan, U.S. Bureau of the Census, Washington, D.C. 20233 U.S.A.

nécessaire de redresser les chiffres de population. De fait, le taux de sous-dénombrement national net n'a jamais été aussi bas. En 1980, il se situait à environ 1.5%, comparativement à un taux de sous-dénombrement net estimatif d'un peu moins de 3% en 1970 et de plus de 5% en 1940. Pour comprendre la situation, il ne faut pas tenir compte uniquement du taux national net, mais aussi du fait que le sous-dénombrement net varie selon les groupes et qu'il est calculé à partir d'erreurs brutes plus importantes. Ainsi, même si le taux de sous-dénombrement net a atteint un nouveau plancher au recensement de 1980, le sous-dénombrement relatif pour la population de race noire (les écarts de sous-dénombrement) subsiste. Il nous faut aussi étudier le niveau relatif de couverture dans une perspective historique. La couverture est-elle vraiment meilleure qu'elle ne l'était il y a 20, 50 ou 100 ans? A-t-on toujours enregistré un taux de sous-dénombrement différent pour les personnes de race noire? Les taux de couverture relatifs selon la race, le sexe et l'âge ont-ils changé avec le temps? Les augmentations des taux de couverture sont-elles réparties également à travers le pays? Quelles sont les incidences des taux de sous-dénombrement et de surdénombrement bruts? Enfin, le taux de sous-dénombrement net étant très faible, les utilisateurs peuvent-ils, sans problème, éviter de tenir compte de ce facteur dans leurs travaux de recherche?

2. ESTIMATIONS DE LA COUVERTURE DES RECENSEMENTS DES ÉTATS-UNIS DEPUIS 1880

L'évaluation de la couverture, comme le processus de recensement lui-même, s'est améliorée avec le temps.[2] Il n'existe pas d'estimations analytiques de la couverture de la population des États-Unis avant 1880. Cependant, Ansley Coale et ses collaborateurs ont calculé des estimations de couverture distinctes pour les populations de race blanche et de race noire, pour les recensements effectués entre 1880 et 1940. De même, le U.S. Bureau of the Census a établi des ensembles cohérents d'estimations de couverture, c'est-à-dire des estimations calculées selon la même méthodologie, pour tous les recensements effectués depuis 1940. Ces estimations du taux de sous-dénombrement net selon la race, le sexe et l'âge sont fondées sur une analyse démographique.

Habituellement, la méthode démographique d'évaluation de la couverture consiste à établir des estimations de la population à la date du recensement, en analysant divers types de données démographiques qui proviennent, pour l'essentiel, de sources indépendantes du recensement, telles que les statistiques sur les naissances, les décès et l'immigration ainsi que les estimations de l'émigration et les données fournies par Medicare. C'est la différence entre l'estimation de la population et les chiffres du recensement qui mesure la couverture nette du recensement.

La méthode démographique repose sur la cohérence des données démographiques sous-jacentes et sur les relations qui existent entre ces données. Si on utilise les composantes de la variation démographique (les naissances, les décès, l'immigration nette), on peut reporter rétrospectivement ou prospectivement les chiffres de population estimatifs pour un groupe d'âge et un recensement (p. ex., les personnes âgées de 65 à 69 ans, en 1980), afin d'en tirer des estimations de la couverture pour un autre recensement (p. ex., les personnes âgées de 45 à 49 ans, en 1960). De cette façon, les nouvelles estimations de la couverture des recensements de 1940 à 1980, fondées sur l'analyse démographique, sont intrinsèquement cohérentes. En ce moment, nous sommes en train d'établir des estimations démographiques cohérentes de la couverture pour le recensement de 1990.

Diverses méthodes ont été utilisées pour établir des estimations de la couverture de groupes raciaux déterminés, avant 1940. Coale et Rives (1973) ont calculé des estimations de la couverture des personnes de race noire aux recensements effectués entre 1880 et 1970 et Coale et Zelnick (1963) ont établi des estimations similaires pour la population de race blanche née au pays, pour les recensements effectués entre 1880 et 1950. Chaque méthode nécessitait la reconstitution complexe de populations clés. Robinson (1988) a étendu la portée historique de ces séries d'estimations de la couverture à la période avant 1940, en combinant les estimations de Coale-Zelnick et celles de Coale-Rives pour établir des estimations composites de la couverture de la population totale des ans décennaux, entre 1880 et 1930.

---

[2] On trouvera une analyse plus approfondie de l'élaboration et de la qualité des estimations de la couverture dans le temps, dans Robinson (1988).

Il est évident que les estimations démographiques de la couverture sont sujettes à erreur. Toutefois, les estimations démographiques pour la période s'étendant de 1940 à 1980 sont d'une qualité supérieure à celle des estimations pour la période de 1880 à 1940, parce qu'elles sont fondées sur de meilleures données et sur des méthodes d'estimations améliorées.[3]

3. TENDANCES ET ÉCARTS DE COUVERTURE DANS LES RECENSEMENTS DE 1880 À 1980

Le tableau 1 présente les estimations de la couverture des onze recensements décennaux réalisés de 1880 à 1980. Ces estimations révèlent l'existence de deux tendances historiques générales: 1) une amélioration faible et irrégulière de la couverture globale, entre 1880 et 1940, caractérisée par des tendances différentes pour les populations de race blanche et de race noire et 2) depuis 1940, une amélioration continue et marquée de la couverture pour les deux groupes avec, toutefois, un écart entre leurs taux de sous-dénombrement respectifs pour chacun des recensements.

Selon le tableau 1, les taux de sous-dénombrement net, tant pour la population noire que pour la population blanche, étaient plus élevés dans les recensements effectués avant 1940, que dans les recensements plus récents. D'abord, les estimations de Coales-Rives pour la population de race noire indiquent que le sous-dénombrement net dépassait 10% dans tous les recensements effectués entre 1890 et 1940, pour atteindre des taux assez élevés en 1890 et en 1920. Il est intéressant de noter que les taux de couverture de la population noire n'affichent aucune tendance nette à l'amélioration jusqu'en 1940, pour ensuite s'accroître graduellement. Ensuite, les estimations de Coales-Zelnick pour la population de race blanche indiquent l'existence de taux de sous-dénombrement constants d'environ 6% jusqu'en 1920, puis une tendance à l'amélioration de la couverture à partir du recensement de 1930. Si on compare les estimations du sous-dénombrement selon la race pour chaque recensement (voir la dernière colonne du tableau 1), on constate que l'écart entre les taux de sous-dénombrement des populations blanche et noire date de longtemps.[4]

Les estimations démographiques de la couverture de 1940 à 1980, pour la population totale et pour certains groupes définis selon le sexe et la race, montrent que la couverture nette s'est améliorée avec chaque recensement consécutif depuis 1940. Dès 1980, le taux de sous-dénombrement net pour la population totale avait été réduit à moins de 2% (1.4%), comparativement à un sous-dénombrement global de plus de 5% (5.6%) en 1940. La réduction des taux de sous-dénombrement net a été particulièrement marquée entre le recensement de 1970 et celui de 1980.

Le taux de sous-dénombrement demeure beaucoup plus élevé pour la population noire que pour la population blanche. En 1980, 5.9% des personnes de race noire ont été oubliées comparativement à seulement 0.9% des personnes de race blanche; en 1940, les taux respectifs correspondants étaient de 10.3% et de 5.1%. Comme le montre le tableau 1 (dernière colonne), cet écart de 5 à 6% entre les taux de sous-dénombrement a persisté au cours des cinq derniers recensements. Si le nombre de personnes oubliées de race blanche a baissé avec chaque recensement, le chiffre net de personnes oubliées de race noire, par contre, est resté à peu près constant (entre 1.5 et 2 millions) pour chacun des cinq derniers recensements.

---

[3] La fiabilité des estimations démographiques de 1980 repose sur la façon dont on a tenu compte de la population d'étrangers pour laquelle on ne dispose pas de données. En outre, nous étudions actuellement un biais possible des résultats du Birth Registration Test de 1940, pour la population noire. Une sous-évaluation de l'exhaustivité de l'enregistrement des naissances a entraîné une surévaluation des naissances «corrigées» et des estimations exagérées de la couverture nette. Ce biais influe principalement sur les estimations des personnes de race noire âgées de 35 à 54 ans, en 1980.

[4] Un redressement en fonction du biais apparent des facteurs d'exhaustivité de l'enregistrement des naissances des personnes de race noire, en 1940 (voir note 3), aura pour résultat d'indiquer des écarts plus faibles entre les populations blanche et noire pour les recensements effectués de 1940 à 1980 que ceux présentés dans le tableau 1.

Tableau 1: Estimations du pourcentage des taux de sous-dénombrement net selon la race et le sexe:: 1880 à 1980

(Le pourcentage de sous-dénombrement est basé sur la population estimative. Reportez-vous au texte pour une description des différentes estimations.)

Année et source de l'estimation	Total	Sexe masculin	Sexe féminin	Population noire	Population blanche	Différence en points de pourcentage: Sexe masculin/ sexe féminin	Différence en points de pourcentage: Population noire/ population blanche
Census Bureau							
1980	1.4	2.4	0.4	5.9	0.9	2.0	5.0
1970	2.9	3.7	2.2	8.0	2.2	1.5	5.8
1960	3.3	3.8	2.8	8.3	2.7	1.0	5.6
1950	4.4	4.8	4.1	9.6	3.8	0.7	5.8
1940	5.6	6.1	5.2	10.3	5.1	0.9	5.2
Coale et Zelnick (population blanche) Coale et Rives (population noire)							
1940	5.0	5.4	4.5	12.7	4.1	0.9	8.6
1930	5.3	5.4	5.1	12.5	4.4	0.3	8.1
1920	6.7	6.2	7.2	15.1	5.6	-1.0	9.5
1910	6.5	5.9	7.1	12.1	5.7	-1.2	6.4
1900	6.7	5.9	7.7	10.9	6.2	-1.8	4.7
1890	7.4	6.4	8.4	14.6	6.3	-2.0	8.3
1880	6.5	5.5	7.4	9.2	6.1	-1.9	3.1

Source: J. Gregory Robinson, "Perspectives on the Completeness of Coverage of Population in the United States Decennial Censuses", article présenté à la Population Association of America, New Orleans, 1988.

Les estimations du tableau 2 (voir page suivante) reflètent les tendances historiques des taux de la couverture selon les groupes d'âge sous-jacentes aux tendances observées chez les groupes répartis selon la race et le sexe. On constate, en général, une évolution des taux de sous-dénombrement des recensements décennaux qui étaient relativement élevés pour la plupart des groupes d'âge et ont baissé considérablement chez tous les groupes à l'exception de deux: les enfants de race noire et les adultes de sexe masculin également de race noire.

Les estimations du tableau 2 révèlent que les taux de sous-dénombrement du recensement de 1980 pour les hommes et les femmes de race blanche de tout âge, pour les femmes de race noire âgées de plus de 5 ans ainsi que pour les hommes de race noire de 5 à 19 ans et de 65 ans et plus sont beaucoup plus bas que les taux de sous-dénombrement estimatifs des premiers recensements. Le sous-dénombrement net pour ces groupes est pratiquement éliminé, si on l'examine dans une perspective historique. Les taux de sous-dénombrement des enfants de race noire (moins de 5 ans) ont diminué de moitié à partir de taux élevés de plus de 20% pour atteindre un taux de 10% environ, en 1980 -- il s'agit toutefois d'un taux encore élevé, comparativement à ceux de la plupart des autres groupes. Les taux de sous-dénombrement des adultes de sexe masculin de race noire constituent la seule exception aux améliorations considérables de la couverture réalisées entre les premiers et les derniers recensements. En effet, la couverture des adultes de race noire âgés de 20 à 44 ans n'a pratiquement pas changé: des taux de sous-dénombrement d'environ 15%, de 1890 à 1980 (tableau 2). En outre, les estimations pour les hommes de race noire âgés de 45 à 64 ans indiquent une détérioration de la couverture au cours de la deuxième moitié du siècle, comparativement aux premières années: un taux de sous-dénombrement net dépassant les 10% dans tous les recensements effectués depuis 1950. Il faut insister sur l'incidence des taux de ces deux groupes sur l'écart souvent cité entre les taux de sous-dénombrement des populations de races blanche et noire -- en effet, si les taux de sous-dénombrement des adultes de sexe masculin de race noire et des enfants de moins de 5 ans également de race noire étaient identiques à ceux du reste de la population noire, l'écart entre les taux de sous-dénombrement des populations blanche et noire du recensement de 1980 serait de 1.5 point, de pourcentage plutôt que de 5 points de pourcentage.

Tableau 2: Estimations des pourcentages des taux de sous-dénombrement net de groupes d'âges sélectionnés, selon la race et le sexe: 1890, 1910, and 1950-1980
(Les pourcentages sont basés sur la population estimative)

Race, sexe et année	Âge				
	Moins de 5 ans	De 5 à 19 ans	De 20 à 44 ans	De 45 à 64 ans	65 ans et plus
Hommes de race blanche					
1980	0.2	0.3	2.7	2.5	0.3
1970	2.5	1.7	3.6	3.5	2.3
1960	2.1	2.5	4.3	2.7	1.3
1950	4.4	3.2	4.5	4.1	2.8
1930	7.1	2.9	5.5	4.4	4.1
1910	7.3	4.3	5.4	5.3	3.6
1890	4.6	4.9	6.9	6.1	3.6
Femmes de race blanche					
1980	0.1	0.1	0.2	-0.3	0.7
1970	2.1	1.3	1.2	1.8	3.6
1960	1.3	1.5	1.8	3.9	4.4
1950	3.8	2.6	2.4	6.1	4.9
1930	6.2	1.6	2.6	9.2	12.1
1910	6.9	2.8	4.7	14.8	10.2
1890	4.8	3.4	8.6	15.4	10.2
Hommes de race noire					
1980	9.6	2.6	13.2	13.6	-1.4
1970	10.4	5.3	17.6	12.1	-0.7
1960	7.1	6.6	17.4	11.7	-6.6
1950	9.8	11.0	14.7	13.9	-14.4
1930	23.4	11.7	12.9	7.6	10.4
1910	20.8	11.2	12.7	1.4	-22.4
1890	28.5	9.7	15.5	6.9	-27.9
Femmes de race noire					
1980	9.0	2.3	3.5	2.4	-0.3
1970	9.5	4.6	5.9	6.0	2.0
1960	5.4	5.1	6.7	11.2	-3.8
1950	9.0	8.5	5.2	16.5	-17.4
1930	22.4	8.7	4.9	25.3	27.0
1910	21.0	9.7	9.5	22.9	13.2
1890	28.9	8.2	14.3	24.3	4.3

Source: Robinson (1988)

4. ÉCARTS DE LA COUVERTURE DES GROUPES INFRA-NATIONAUX ET SOCIO-ÉCONOMIQUES

Nous disposons de moins de données sur les tendances et les variations de la couverture selon les régions ou selon les groupes socio-économiques que sur les taux et les tendances de la couverture du pays dans son ensemble. Dans cette section, nous examinerons certaines des données disponibles.

Presque toutes les évaluations des recensements de 1970 et avant indiquent que les taux de sous-dénombrement étaient plus élevés dans le Sud que partout ailleurs (tableau 3). Le fait que la population noire, dont les taux de sous-dénombrement ont toujours été plus élevés que ceux de la population blanche, est concentrée dans le Sud explique en partie cet écart. En outre, on enregistrait dans le Sud des taux de sous-dénombrement plus élevés, tant pour la population blanche que pour la population noire. En général, les études d'évaluation indiquent que les taux de sous-dénombrement de l'Ouest se situaient entre ceux du Sud et ceux du Nord, ces derniers étant les plus faibles au pays. En 1980, la couverture selon les régions avait changé -- le taux de sous-dénombrement net de l'Ouest était alors le plus élevé du pays et ceux du Nord et du Sud étaient relativement plus bas. La baisse importante du taux de sous-dénombrement du Sud (tant pour la population blanche que pour la population noire) est la cause principale de ce changement de tendance dans la couverture selon les régions, survenu entre 1970 et 1980, mais les variations du surdénombrement brut peuvent aussi avoir joué un rôle (voir section 5).

Tableau 3: Taux de sous-dénombrement selon les régions, d'après les résultats d'études d'évaluation: 1970-1980
(Les taux de la population représentent le pourcentage des personnes oubliées calculés à partir des chiffres de population corrigés)

Groupe et année	États-Unis	Le Nord-est	Le Centre nord	Le Sud	L'Ouest
Analyse démographique de 1980*					
Tous les groupes	1.4	1.1	0.4	1.0	3.5
Population blanche	0.9	0.6	0.3	0.5	3.0
Population noire et d'autres races	4.3	5.6	4.7	3.3	6.8
Programme post-censitaire de 1980					
Taux de sous-dénombrement net - ensemble 2-9					
Tous les groupes	1.6	1.1	1.3	1.2	3.1
Population blanche	0.7	0.1	0.5	0.1	2.5
Population noire	7.2	9.5	7.7	5.7	10.3
Analyse démographique de 1970					
Tous les groupes	2.6	1.4	1.4	4.0	3.3
Population blanche	1.9	0.9	1.1	3.0	2.8
Population noire et d'autres races	6.9	5.5	5.1	7.8	7.9

* J. Gregory Robinson, (1986)

Les résultats de recherches sur la relation entre la couverture et la situation socio-économique et d'autres variables sont résumés dans U.S. Bureau of the Census (1975), p. 8 à 10 et dans Citro et Cohen (1985), annexe 5.1. Ainsi, par exemple, selon l'étude d'appariement de l'Enquête post-censitaire de 1970, il y avait deux fois plus de personnes oubliées dans les familles dont le revenu était inférieur à $7,500 (4.4%) que dans les familles de revenu supérieur (2.1%). De même, le taux de personnes oubliées était nettement plus élevé parmi les personnes ayant un niveau de scolarité de 8 ans ou moins (3.8%) que parmi celles ayant poursuivi des études secondaires (2.5%) (U.S. Bureau of the Census, 1975, p. 8 et 9).

Tableau 4: Écarts des taux de sous-dénombrement estimatifs d'après le Programme post-censitaire de 1980

Séries	Population de race noire	Population hispanique (non de race noire)
2-8	5.0	3.6
3-8	4.7	3.5
2-9	5.8	4.3
3-9	5.5	4.2
14-9	2.8	1.7
2-20	5.9	4.2
3-20	5.7	4.2
14-20	3.0	1.7
5-8	2.8	4.9
10-8	2.5	3.4
5-9	3.6	5.7
14-8	2.1	1.0*
E. t. approx.		
(Ensembles 2, 3, 14)	0.6	0.8
(Ensembles 5, 10)	0.6	1.0

* Non significatifs au niveau de confiance de 90%

Le taux de sous-dénombrement peut être également plus élevé pour certains groupes ethniques, comme la population hispanique. Le tableau 4 présente les écarts de sous-dénombrement pour la population hispanique ainsi que pour la population noire, d'après les estimateurs du Programme post-censitaire (PP) de 1980. On a établi un ensemble d'estimations à partir d'un éventail d'hypothèses. Le U.S. Bureau of the Census, préoccupé par le biais des estimations, a établi les douze ensembles d'estimations sur des données et des hypothèses diverses afin de déterminer la sensibilité des estimations à de possibles non-respects des hypothèses.[5] Examinons les écarts des taux de sous-dénombrement indiqués par chaque ensemble d'estimations. Si on soustrait le taux de sous-dénombrement national estimatif, on élimine tout biais uniforme des ensembles d'estimations mais les biais qui varient selon les régions demeurent. On constate encore des écarts de sous-dénombrement. Nous pouvons en conclure, malgré le peu de données historiques, que le problème n'est pas limité à la population noire. En effet, selon le PP de 1980, les écarts de sous-dénombrement de la population hispanique s'apparente à celui de la population noire. (Voir aussi U.S. Bureau of the Census, 1979).

Les recherches sur la couverture tenant compte du type de domicile (urbain ou rural) ont révélé de façon assez constante que le taux d'omissions dans les logements occupés est plus élevé dans les secteurs ruraux que dans les secteurs urbains, bien que les observations sur les variations de la couverture selon la superficie des logements, dans les secteurs urbains, ne soient pas concluantes. Le PP de 1980 a permis, pour la première fois, l'évaluation de la couverture dans certaines villes. Le programme d'évaluation de la couverture de 1980 a établi des estimations du taux de sous-dénombrement de 16 grandes villes.

Le tableau 5 présente les écarts entre les taux estimatifs de sous-dénombrement et le taux national pour chaque ville et chaque ensemble d'estimations. Bon nombre des estimations individuelles ne sont pas significatives au niveau de confiance de 90%. Cependant, si les villes en tant que groupe présentent le même sous-dénombrement que le pays dans son ensemble, on pourrait s'attendre à ce que les estimations de seulement deux des seize villes soient significatives au niveau de 90% (0.1 X 16 < 2), selon les lois du hasard. On ne constate

[5] Pour une analyse des ensembles d'estimations du PP de 1980, voir U.S. Bureau of the Census (1988).

ce résultat que pour les estimations «14-8». Pour les autres ensembles, quatre villes ou plus présentent des taux de sous-dénombrement statistiquement significatifs.[6]

Tableau 5: Estimations des écarts de couverture dans 16 villes d'après le Programme post-censitaire de 1980

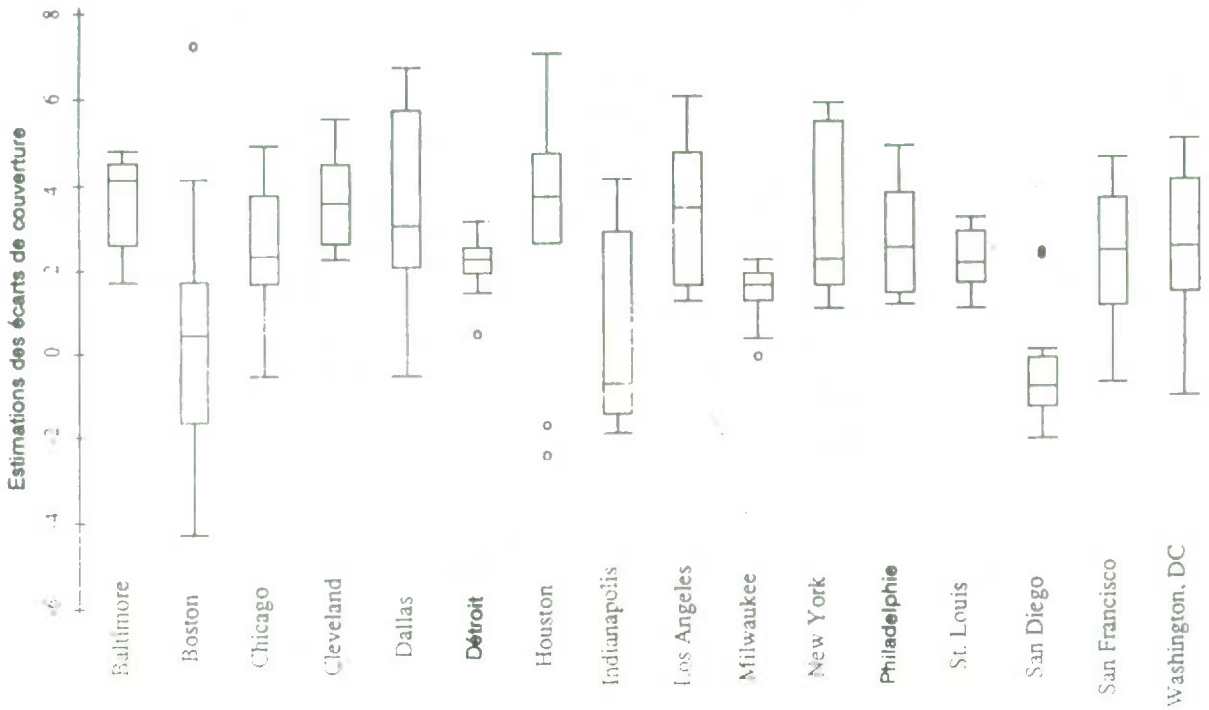
	Séries										Erreurs types	
	2-8	2-9	3-8	3-9	5-8	5-9	3-20	10-8	14-8	14-9	2,3,14	5,10
Baltimore	4.7	4.6	4.4	4.3	2.9	2.8	4.8	2.4	4.2	4.0	1.7	1.7
Boston	-1.9	0.8	-2.0	0.7	-0.3	2.6	0.7	0.1	-4.3	-1.5	4.1	7.2
Chicago	2.5	3.9	3.4	4.8	2.1	3.6	4.9	1.8	-0.6	0.8	1.5	1.8
Cleveland	3.6	3.5	3.0	3.8	5.5	5.4	3.5	5.0	2.7	2.5	2.2	2.5
Dallas	5.9	6.7	4.9	5.7	3.0	3.9	5.8	3.0	-0.6	0.2	2.0	1.8
Détroit	2.2	2.8	2.1	2.7	1.9	2.5	3.1	0.4	1.4	2.0	2.3	1.8
Houston	3.7	4.3	3.6	4.2	6.4	7.0	4.7	5.2	-2.5	-1.8	2.6	2.8
Indianapolis	-0.8	-1.2	-1.2	-1.6	4.1	3.7	-1.8	2.9	-1.5	-2.0	2.1	2.7
Los Angeles	4.2	6.0	3.6	5.4	1.6	3.4	5.3	1.6	1.8	3.5	1.4	1.2
Milwaukee	2.1	1.7	2.1	1.7	0.3	-0.1	1.5	1.0	2.2	1.7	1.4	1.4
New York	5.3	5.9	5.0	5.6	1.5	2.1	5.8	1.7	1.8	2.3	1.2	1.0
Philadelphie	4.8	4.9	3.7	3.8	1.1	1.3	3.8	1.2	2.5	2.5	1.9	1.5
Saint Louis	2.0	2.1	2.1	2.2	3.0	3.1	3.2	1.0	1.3	1.3	2.1	2.8
San Diego	-2.1	-1.0	-2.0	-0.9	-1.7	-0.6	-0.3	-0.9	-1.0	0.0	2.3	2.4
San Francisco	3.2	3.7	3.6	4.1	0.6	1.1	4.6	-0.8	1.0	1.4	2.9	1.9
Washington, DC	2.9	4.4	2.6	4.2	1.2	0.4	5.0	-1.1	2.4	3.9	1.6	2.2

Une autre méthode consiste à comparer les estimations des différentes séries pour chaque ville au moyen de diagrammes en boîte (figure 1). Il faut retenir que les estimations pour le mois d'avril (ensembles 2, 3, 14) présentent une forte corrélation entre elles tout comme celles du mois d'août (5,10) alors que la corrélation entre les estimations d'avril et d'août est faible: un coefficient de corrélation de .2 ou .3. Si on ne tient pas compte de biais, certaines tendances semblent se dégager. L'enquête post-censitaire (EP) de 1980 n'a révélé aucun écart de sous-dénombrement pour Boston, Indianapolis ou San Diego. Bien que la tendance pour Milwaukee, San Francisco, Saint Louis, Détroit et Washington ne soit pas marquée, elle semble indiquer un écart de sous-dénombrement. La tendance pour Baltimore, Chicago, Cleveland, Dallas, Houston, Los Angeles, New York et Philadelphie concorde étroitement avec la moyenne nationale. Certaines des données de l'EP de 1980 appuient donc la thèse d'un écart de sous-dénombrement dans les noyaux urbains de certaines grandes régions métropolitaines, bien qu'il faille être prudent et tenir compte des biais et des variances.

Les données révèlent que le taux de sous-dénombrement est plus élevé chez les pauvres, les célibataires et les chômeurs. De même, il est plus élevé chez les locataires que chez les propriétaires (Isaki et coll., 1987). On a constaté également des écarts en utilisant beaucoup d'autres variables de type social. On trouvera dans Fein et West (1988) une analyse du recensement de 1986 de l'East Central Los Angeles County.

[6] Les recherches de Schaefer (1989) ont remis en question l'exactitude des estimations de la série 14.

Figure 1. Estimations des écarts de couverture dans 16 villes d'après le Programme post-censitaire de 1980



5. ERREUR DE COUVERTURE BRUTE

Bien que cette recherche ait porté essentiellement sur le taux de sous-dénombrement net de la population, il faut savoir que les taux nets sont le résultat d'un taux de sous-dénombrement brut plus important, partiellement compensé par un taux de surdénombrement brut. Donc, les variations des taux de sous-dénombrement net proviennent d'une variation sous-jacente des taux de sous-dénombrement brut et de surdénombrement brut (personnes dénombrées plus d'une fois, falsifications faites par les recenseurs et autres erreurs). Bien qu'on dispose de nombreuses données sur l'évolution du taux de sous-dénombrement net, il est plus difficile d'obtenir des données claires sur l'importance et l'évolution des composantes distinctes (personnes oubliées et personnes dénombrées plus d'une fois) de ce sous-dénombrement.

En 1980, le taux de sous-dénombrement net était de 1.4% soit 3.2 millions de personnes. Selon l'évaluation de ce recensement, au moins 2.7 millions de personnes ont été recensées plus d'une fois (Cowan et Fay, 1984; Jones, 1986)[7] ce qui signifie qu'au moins 5.9 millions de personnes ont été oubliées. Ainsi, les données ne tiennent pas compte de millions de personnes réelles et incluent beaucoup de personnes recensées qui n'auraient pas dû l'être. Si la répartition géographique ou démographique des personnes recensées par erreur était différente de

[7] Cette estimation a été calculée en ajoutant les 2 492 900 personnes dénombrées deux fois, selon le PP, aux 214 000 personnes recensées deux fois, selon le Whole Household Usual Home Elsewhere Program. Le PP a compté seulement les personnes recensées deux fois repérées lors d'une recherche effectuée aux environs du secteur recensé. Cette estimation ne tient donc pas compte des personnes qui ont déménagé après le recensement et qui auraient été recensées à leur adresse le jour du recensement et, une autre fois, à leur nouvelle adresse.

celle des personnes oubliées, les répercussions pour les utilisateurs de données seraient beaucoup plus graves que ne le laisse supposer un taux de sous-dénombrement net de 1.4%.

Le problème du surdénombrement et du sous-dénombrement bruts est lié à celui des écarts entre les taux de sous-dénombrement net. Par exemple, la plupart des logements recensés deux fois se trouvaient dans les secteurs pour lesquels des listes avaient été préparées d'avance (U.S. Bureau of the Census, 1985), situés à l'extérieur des noyaux urbains des régions métropolitaines. Le taux de surdénombrement des logements occupés dans les secteurs à l'extérieur des Régions statistiques métropolitaines (Metropolitan statistical areas, MSA's) (1.20%) dépassait celui des secteurs à l'intérieur de ces régions (0.78%). Dans les régions recensées de façon traditionnelle (c'est-à-dire autrement que par la méthode d'envoi et de retour par la poste du questionnaire), on a enregistré un taux de surdénombrement extrêmement bas (0.11%). Une étude des variations de la couverture doit nécessairement inclure une analyse des erreurs brutes à moins que l'effectif des personnes dénombrées par erreur ne soit réparti également selon l'âge, la race, le sexe et les régions géographiques, comme l'effectif des personnes oubliées.

Les données sur la couverture des unités de logement apportent des éclaircissements sur les variations de la couverture selon les régions. Le tableau 6 présente les taux de surdénombrement et de sous-dénombrement bruts par région. Les taux de sous-dénombrement bruts du Sud et de l'Ouest s'équivalaient plus ou moins. Cependant, le taux de surdénombrement du Sud était le plus élevé tandis que celui de l'Ouest se situait parmi les plus faibles. Le Census Bureau a conclu: «Bien qu'aucune évaluation approfondie n'ait été réalisée pour les recensements précédents, les données indiquent que le nombre de personnes dénombrées plus d'une fois était beaucoup moins élevé. Il est donc regrettable qu'une part de l'amélioration du taux de couverture nette du recensement de 1980 soit imputable au nombre élevé de personnes dénombrées plus d'une fois.» (U.S. Bureau of the Census, 1980, p. 10).

Tableau 6: Erreurs de couverture brute estimatives réparties selon les régions: 1980

	États-Unis	Nord-est	Centre nord	Sud	Ouest
Omissions brutes	5.6	5.1	4.3	6.5	6.1
Taux de surdénombrements bruts	2.9	3.3	2.1	3.6	2.0

Charles D. Cowan et Robert E. Fay, (1984); (Ensembles 2-9)

6. IMPLICATIONS DES ÉCARTS DE SOUS-DÉNOMBREMENT

Un des objectifs de l'évaluation de la couverture est d'informer les utilisateurs de données du recensement des répercussions des écarts de couverture. Quelles sont pour eux les conséquences du sous-dénombrement? Les données du recensement servent surtout dans quatre domaines: politique, juridique, sciences sociales et planification.

De nombreuses utilisations politiques des données ne tiennent compte que des chiffres de population. Par exemple, les écarts des taux de sous-dénombrement n'ont une incidence sur la répartition des sièges que s'ils varient selon les régions. Cependant, selon les dispositions du "Voting Rights Act", des variations de la couverture selon la race, le groupe ethnique et même l'âge peuvent avoir des conséquences directes lors de la redéfinition des districts. En outre, de nombreux programmes gouvernementaux utilisent les chiffres de population dans les calculs de la distribution des subventions fédérales. Seul le taux de sous-dénombrement net a une importance pour la plupart de ces programmes.

Les chiffres de population servent aussi à des fins juridiques autres que l'affectation des ressources. Par exemple, le nombre de personnes de différents groupes raciaux dans un secteur d'emploi peut constituer une preuve dans les poursuites judiciaires pour discrimination dans l'emploi. En outre, les données des recensements décennaux permettent souvent d'évaluer la représentativité du système de sélection des jurés dans une juridiction spécifique.

La caractéristique raciale est la plus courante, mais dans certaines récusations on a mentionné le sexe, l'âge, la profession, le nivcau de scolarité et la situation économique (voir Rolph, 1986). Les écarts de couverture selon la race et l'âgc peuvent fausser le «bassin de main-d'oeuvre» ou le «bassin de jurés» estimatifs. En effet, le fait de tenir compte des personnes dénombrées plus d'une fois peut fausser tant la taille que la répartition des populations estimatives.

Les données du recensement servent aussi, directement ou indircctement, dans de nombreuses recherches en sciences sociales. Il suffit d'examiner quelques-unes des utilisations pour saisir l'importance d'une compréhension claire du sous-dénombrement du recensement.

Le rapport du taux brut de mortalité des personnes de sexe masculin de race noire au taux brut de mortalité des personnes de sexe masculin de race blanche en 1980 est de 1.05 (excédent de 5%). Le rapport réel équivaudra au rapport du taux de couverture pour les hommes de race noire (.912) au taux de couverture des hommes de race blanche (.983):

$$.912/.983 = .928$$

Le rapport des taux réels sera de 1.05 x .928 = .9744

Le National Center for Health Statistics a fourni quelquefois les instructions et les facteurs nécessaires au redressement des taux de sous-dénombrement du rccensement, dans le rapport sur les statistiques démographiques (U.S. Department of Health and Human Services, 1981).

Le taux de prévalence constitue une mesure couramment utilisée pour déterminer l'ampleur des problèmes sociaux. L'analyse de ce taux ressemble à celle des taux bruts de mortalité sauf qu'il n'est plus certain que les cas soient toujours déclarés. Il peut y avoir plus d'erreurs que dans les chiffres de population. Par exemple, on déclare peut-être beaucoup moins de cas de SIDA qu'il n'y en a en réalité. Néanmoins, un chercheur étudiant la prévalence des cas de SIDA dans les populations des quartiers déshérités devrait tout de même tenir compte de la couverture du recensement (voir Shai et Aptekar, 1990).

L'espérance de vie constitue une mesure plus complexe de la mortalité que le taux brut de mortalité. Le taux de sous-dénombrement du recensement de 1980 n'a pas d'incidence marquée sur l'espérance de vie à la naissance ou à 65 ans. Ce n'était pas le cas pour les rccensements précédents. Par exemple, le taux de sous-dénombrement du recensement de 1970 a réduit l'espérance de vie à la naissance observée pour les hommes de race noire de 1.5 an. En 1980, le redressement des chiffres de population, en fonction du taux de sous-dénombrement des hommes de race noire, a eu pour effet d'augmenter l'espérance de vie à 20 ans de plus d'une année.

Le rapport de masculinité calculé à partir des chiffres du recensement peut être trompeur. En effet, les personnes oubliées sont plus souvent de sexe masculin que de sexe féminin. C'est particulièrement vrai chez les adultes de race noire. Si nous examinons le rapport de masculinité du recensement et le rapport réel estimatif, nous pouvons percevoir les conséquences d'un tel écart. En 1980, le rapport de masculinité de la population de race noire âgée de 35 à 54, calculé à partir des chiffres de recensement, était inférieur de 13% à ce qu'il aurait du être. Quelles sont les conséquences d'une telle erreur sur notre perception des problèmes, des comportements sociaux, des conditions de logement de la population noire? En 1970, une étude ethnographique de petite échelle portant sur un quartier noir a démontré qu'alors que les données du U.S. Bureau of the Census indiquaient que 72% des ménages avaient un chef féminin, ces ménages ne représentaient, en réalité, que 12% du total. En se basant sur cette constatation, les ethnographes ont affirmé que des données du recensement biaisées créent et maintiennent une image fausse des ménages noirs ayant un chef féminin. (Voir Hainer et coll., 1988).

Dans un article sur les comportements matrimoniaux de la population noire, Goldman et ses collaborateurs (1984) ont dû tenir compte de cet aspect du problème.

> «À cause de la gravité du problème chez les hommes de race noire, nous avons redressé les chiffres de toute la population de célibataires, selon la race, l'âge et le sexe... Comme le sous-dénombrement est

probablement plus élevé chez les célibataires, les hommes, particulièrement ceux de race noire, sont peut-être davantage sous-représentés, dans nos estimations du nombre de célibataires.»

Les résultats auraient été différents s'ils n'avaient pas tenu compte de ce problème.

Le taux de sous-dénombrement influe sur les taux de croissance démographiques entre les recensements. Selon les chiffres de population officiels du recensement, la population résidant aux États-Unis aurait augmenté de 11.4% entre 1970 et 1980. Par contre, le taux de croissance calculé à partir de la population réelle estimative n'est que de 9.7% pour cette décennie. Sans aucun doute, on pourrait constater la même déformation des faits au niveau des localités et des États.

Les écarts de sous-dénombrement peuvent également fausser d'autres séries de données. Le recensement est utilisé pour délimiter des bases d'échantillonnage en vue d'autres enquêtes telles que l'Étude de la population actuelle (EPA). Le sous-dénombrement de personnes n'influe pas sur le choix de l'échantillon parce que celui-ci est constitué uniquement d'unités de logements. Toutefois, ces enquêtes rencontrent d'autres problèmes de sous-dénombrement. Dans l'ensemble, le taux de couverture de l'EPA est inférieur de 7% environ au taux de couverture du recensement. Pour les hommes de race noire il est inférieur de 17% et pour les hommes de 20 à 24 ans de race noire, de 27%. (Hainer et coll., 1988). Les chiffres du recensement servent aussi à corriger les problèmes de couverture d'autres enquêtes. Le U.S. Bureau of the Census contrôle statistiquement les données de L'EPA pour qu'elles concordent avec les données prévues; le sous-dénombrement du recensement est donc reporté dans l'EPA.

Le secteur privé utilise aussi largement les données du recensement et l'incidence du sous-dénombrement n'est pas passée inaperçue. Le maire de Kansas City, Richard Berkeley, a déclaré en 1988: «Cela peut arriver aussi dans le secteur privé où les décisions d'achat de publicité ou autres sont prises sur la base des chiffres de population dans une collectivité ou une région.» Le taux de sous-dénombrement peut influer sur les décisions d'annonceurs qui veulent acheter du temps sur des chaînes diffusant en espagnol, par exemple. L'incidence des variations des taux de sous-dénombrement sur le secteur privé constitue un domaine de recherche encore inexploré.

7. CONCLUSION

Le sous-dénombrement net à l'occasion du recensement constitue un sujet d'étude complexe: il correspond à la différence entre le nombre de personnes oubliées et le nombre de personnes dénombrées par erreur; il varie selon les groupes et les régions; ses répercussions diffèrent selon les utilisations qu'on fait des données du recensement; enfin, les méthodes utilisées pour l'évaluer sont complexes et sujettes à erreurs. Lors de la planification du recensement de 1990, nous avons conçu un programme à plusieurs facettes, dont les deux principales sont l'analyse démographique et l'enquête post-censitaire. En outre, nous avons élaboré un programme d'évaluation du programme de mesure de la couverture et un programme d'observation participante destiné à nous permettre de mieux comprendre les deux facettes principales du programme d'amélioration de la couverture. Ensemble, ces programmes devraient fournir des mesures plus précises de la couverture du recensement tant aux utilisateurs de données qu'aux chercheurs étudiant les méthodes de recensement.

BIBLIOGRAPHIE

Berkeley, R. (1988). *Interview on Morning Edition*, National Public Radio, Janvier 10, 1988.

Citro, C., et Cohen, M.L., eds. (1985). *The Bicentennial Census: New Directions for Academy Press*, Washington, D.C.

Coale, A.J., et Zelnik, M. (1963). *New Estimates of Fertility and Population in the United States*, Princeton University, Press, Princeton, New Jersey.

Coale, A.J., et Rives, N.W. (1973). A Statistical Reconstruction of the Black Population of the United States, 1880-1970: Estimates of True Numbers by Age and Sex, Birth Rates, and Total Fertility, *Population Index* 39(1):3-36.

Cowan, C., et Fay, R. (1984). Estimates of Undercount in the 1980 Census, *Proceedings of the American Statistical Association Annual Meetings.*

Fein, D.J., et West, K. (1988). Sources du sous-dénombrement lors du recensement: Résultats du recensement d'essai de 1986 à Los Angeles, *Techniques d'enquête*, 14, 2, 237-256.

Goldman, N., Westoff, C.F., et Hammerslough, C. (1984). Demography of the Marriage Market in the United States, *Population Index*, 50, 1.

Hainer, P. C. (1987). A Brief and Qualitative Anthropological Study Exploring the Reasons for Census Coverage Error Among Low Income Black Households, non publié.

Hainer, P., *et coll.* (1988). Research of Improving Coverage in Household Surveys, *Fourth Annual Research Conference Proceedings*.

Isaki, C.T., Schultz, L.K., Smith, P.J., et Diffendal, G.J. (1987). Small Area Estimation Research for Census Undercount: Progress Report, *Small Area Statistics, An International Symposium*, New York: John Wiley & Sons.

Jones, C.D. (1986). Note de service de Barbara Bailar, *Duplicate Enumerations*, 7 octobre, 1986.

Robinson, J.G. (1988). Perspectives on the Completeness of Coverage of Population in the United States Censuses, article présenté à Population Association of America, New Orleans, Louisiana.

Robinson, J.G. (1986). Temporal and Regional Variations of the Coverage of Population in the 1980 and 1970 Censuses, article présenté à annual meeting of the Southern Regional Demographic Group, 1986.

Rolph, J. Discussion. *Proceedings of the Second Annual Research Conference*, 344-346.

Schaefer, J. (1989). Missing Data Procedures in the 1980 Post Enumeration Program: Why PEP Series 14 Cannot Be Trusted, unpublished manuscript, Harvard University, Cambridge, Massachusetts.

Shai, D., et Aptekar, L. (1990). Factors in Mortality by Drug Dependence Among Puerto Ricans in New York City; *American Journal of Drug and Alcool Abuse*, 97-107.

U.S. Bureau of the Census (1973). *1970 Census of Population and Housing, Research and Evaluation Research Program, PHC(E)-5*, The Coverage of Housing in the 1970 Census, U.S. Government Printing Office: Washington, D.C.

U.S. Bureau of the Census (1975). *U.S. Census of Population and Housing: 1970*, Procedural History PHR(R)-1, U.S. Government Printing Office, Washington D.C.

U.S. Bureau of the Census (1979). *Coverage of the Hispanic Population of the United States in the United States in the 1970 Census*, Current Population Reports, Series P-23, No. 82, U.S. Government Printing Office: Washington, D.C.

U.S. Bureau of the Census (1985). *1980 Census of the Population and Housing, Evaluation and Research Reports, PHC80-E4*, The Coverage of Housing in the 1980 Census, U.S. Government Printing Office: Washington, D.C.

U.S. Bureau of the Census (1988). *1980 Census of Population and Housing, Evaluation and Research Reports, PHC80-E4*, The Coverage of Population in the 1980 Census, U.S. Government Printing Office: Washington D.C.

U.S. Department of Health and Human Services (1981). *Vital Statistics of the United States 1977*, Volume II, Part A, Technical Appendix et Table 6-4.

NOTE

Les opinions exprimées dans cet article sont celles des auteurs et n'engagent nullement le U.S. Bureau of the Census.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

ÉVALUATION DE L'ERREUR DE COUVERTURE NETTE DANS LES RECENSEMENTS CANADIENS

R.G. Carter[1]

RÉSUMÉ

La couverture des recensements de la population peut être évaluée grâce à des méthodes démographiques, ou à des estimations des erreurs de couverture fondées sur des enquêtes, ou encore à une combinaison de ces deux méthodes. Bien que la plupart des pays qui tentent d'évaluer directement les erreurs de couverture mènent à cette fin une enquête postcensitaire (EP), le Canada a adopté à cet égard une approche quelque peu différente. Depuis 1966, on évalue la couverture des recensements effectués au pays en procédant à la contre-vérification des dossiers (CVD), c'est-à-dire que l'on effectue l'appariement entre un échantillon de personnes qui répondent aux critères, sélectionnées à l'avance, et les documents du recensement pour déterminer si ces personnes ont été dénombrées ou non. Bien que cette méthode permette d'obtenir des estimations du sous-dénombrement, il n'existe aucune estimation directe de l'erreur de couverture nette (c'est-à-dire du sous-dénombrement moins le surdénombrement). On a toujours tenu pour acquis que le surdénombrement était infime, mais on n'a jamais tenté de l'évaluer avant le recensement de 1986. Toutefois, en 1986, on a mené une étude pilote sur le surdénombrement. En dépit de résultats restreints, cette étude a été fort utile pour expérimenter la mesure du surdénombrement et l'on prévoit en mener une semblable, mais plus importante et perfectionnée, dans le contexte du recensement de 1991. On envisage d'élaborer et de diffuser des estimations nationales et provinciales de l'erreur de couverture nette en 1991, fondées sur la contre-vérification des dossiers et sur l'étude sur le surdénombrement.

Le présent document renferme un aperçu du programme de mesure de la couverture du recensement de 1991. On y expose également les raisons pour lesquelles on a décidé de procéder à une étude sur le surdénombrement plutôt que de mener une enquête postcensitaire. En outre, on y étudie la possibilité d'utiliser les estimations de l'erreur de couverture obtenues au moyen de méthodes démographiques pour évaluer les résultats de la contre-vérification des dossiers et de l'étude sur le surdénombrement et pour estimer l'erreur de couverture nette des recensements antérieurs à celui de 1991.

MOTS CLÉS: Sous-dénombrement; surdénombrement; contre-vérification des dossiers; erreur en fin de période.

1. INTRODUCTION

Traditionnellement, un recensement a pour objet principal de dénombrer les personnes et les ménages qui vivent dans un pays et ses diverses sous-divisions administratives. Par conséquent, l'un des aspects les plus importants dont il faut tenir compte pour évaluer la qualité des données du recensement, c'est la couverture obtenue, c'est-à-dire la proportion de la population qui a vraiment été recensée. Les erreurs de couverture peuvent survenir à divers stades de la collecte et du traitement des données du recensement. La plupart de ces erreurs se traduisent par un sous-dénombrement, c'est-à-dire l'omission de ménages ou de certains membres de ménages. Toutefois,

[1] R.G. Carter, Division des méthodes d'enquêtes sociales, Statistique Canada, Parc Tunney, Ottawa (Ontario) K1A 0T6.

certaines erreurs de couverture, comme l'inclusion de personnes qui n'existent pas ou qui ne font pas partie de l'univers du recensement et le dénombrement de certaines personnes plus d'une fois, ont tendance à faire augmenter le chiffre du recensement par rapport au chiffre réel de la population; on parle alors de surdénombrement. Dans la plupart des cas, on s'intéresse surtout à l'écart entre le chiffre du recensement et le chiffre réel de la population, cette différence étant l'effet net du sous-dénombrement et du surdénombrement. Cependant, dans la mesure où des gens ou des ménages qui ont été oubliés ou qui ont été comptés en double présentent des caractéristiques différentes, que ce soit entre eux ou en comparaison avec ceux qui ont été dénombrés correctement, on peut tirer des conclusions erronées même si l'erreur de couverture nette est de zéro. Ainsi, lorsqu'on tente d'évaluer l'erreur de couverture, il est utile de mesurer tant le sous-dénombrement que le surdénombrement et de déterminer les caractéristiques des personnes ou des ménages en question.

Diverses méthodes ont été élaborées pour mesurer les erreurs de couverture. En général, on peut les classer de la façon suivante : les «micro-évaluations» et les «macro-évaluations», lesquelles peuvent être fondées soit sur une enquête-échantillon, soit sur des données administratives. Dans le cas d'une «micro-évaluation», on procède à l'appariement de dossiers individuels afin de repérer les personnes oubliées ou celles qui ont été mal recensées, tandis qu'une «macro-évaluation» repose sur la comparaison d'agrégats. De nombreux pays emploient plus d'une approche pour évaluer la qualité des données de leur recensement. Par exemple, le United States Bureau of the Census effectue à la fois une micro-évaluation fondée sur une enquête-échantillon (enquête postcensitaire) et à une macro-évaluation reposant sur les données administratives (analyse démographique). Statistique Canada se sert aussi de deux approches fondamentales pour évaluer les erreurs de couverture. D'une part, on a recours à des méthodes démographiques (macro-évaluations fondées sur les dossiers administratifs) pour estimer la variation du chiffre de la population d'un recensement au suivant. En comparant avec le chiffre du recensement correspondant, on obtient ainsi une estimation de la variation de l'erreur de couverture nette (voir la partie 3). D'autre part, on mesure le sous-dénombrement à l'aide d'une micro-évaluation fondée sur des échantillons tirés du recensement précédent et sur des dossiers administratifs (contre-vérification des dossiers). Au recensement de 1986, on a mis à l'essai une micro-évaluation fondée sur une enquête-échantillon afin d'évaluer le surdénombrement; cette méthode sera mise en application en 1991.

Les deux parties suivantes décrivent plus en détail les deux approches fondamentales adoptées par Statistique Canada. La partie 2 donne un aperçu de la technique de la contre-vérification des dossiers et de ses faiblesses et présente certains résultats. En outre, on y décrit brièvement l'étude sur le surdénombrement qui a été mise à l'essai en 1986. On y exposera les modifications à ces études qui sont prévues pour 1991 ainsi que les raisons pour lesquelles on a préféré cette approche aux enquêtes postcensitaires. La partie 3 décrit les méthodes démographiques, tandis que la partie 4 traite de la possibilité d'avoir recours aux méthodes démographiques pour évaluer l'erreur de couverture nette des recensements antérieurs à celui de 1991.

2. ESTIMATIONS, FONDÉES SUR DES ENQUÊTES, DU SOUS-DÉNOMBREMENT ET DU SURDÉNOMBREMENT

2.1 La contre-vérification des dossiers

Méthodologie

La contre-vérification des dossiers (CVD) consiste, à la base, à prélever un échantillon de personnes qui doivent être dénombrées au recensement, de retracer ces personnes à leur adresse le jour du recensement, puis de vérifier si elles ont été dénombrées ou non. Traditionnellement, l'échantillon est constitué à partir de quatre bases:

1) les personnes dénombrées au recensement précédent et
2) les personnes oubliées lors du recensement précédent.

Ensemble, ces bases couvrent la population d'il y a cinq ans; on ajoute ensuite:

3) les naissances survenues durant la période intercensitaire et
4) les immigrants reçus pendant cette période.

Sur le plan conceptuel, ces quatre bases couvrent la population cible du recensement actuel (plus les personnes qui sont décédées ou qui ont émigré et qui doivent par conséquent être dépistées comme ne faisant plus partie de la population canadienne). À l'exception de la base des personnes oubliées, un échantillon stratifié est tiré de chaque base. On apparie les échantillons à un dossier administratif pour obtenir des adresses plus à jour. Après le recensement, on apparie les échantillons aux documents du recensement, et les non-appariements font l'objet d'une opération de dépistage sur le terrain afin d'établir la bonne adresse le jour du recensement; on procède ensuite à un autre appariement. À la fin de ce processus, chaque personne choisie est classée dans la catégorie des personnes dénombrées, des personnes oubliées, des personnes décédées, des personnes ayant émigré ou des personnes non dépistées. On obtient les estimations définitives en repondérant les dossiers sur les personnes dépistées pour compenser pour celles qui ne l'ont pas été et en corrigeant les poids pour en arriver aux totaux connus pour chaque base.

Résultats

La figure 1 présente les taux globaux de sous-dénombrement qui ont été estimés grâce à la CVD pour les recensements de 1966 à 1986. (La CVD a débuté en 1961, mais les résultats sont incomplets étant donné que l'échantillon était limité à la base du recensement.) Comme on peut le voir, après s'être établi aux alentours de 2% dans les trois recensements de 1971 à 1981, le taux de sous-dénombrement s'est accru de façon importante en 1986. Comme la figure 2 le montre, des écarts considérables existent dans le sous-dénombrement de part et d'autre du pays. Le taux pour la Colombie-Britannique est toujours supérieur au taux national, tandis que les taux pour la région de l'Atlantique et la région des Prairies ont tendance à être inférieurs au taux national. (Cette tendance s'est maintenue dans les cinq recensements qui ont été réalisés depuis 1966.) Il existe également des variations importantes selon l'âge et le sexe (voir la figure 3). Notamment, le taux de sous-dénombrement des personnes âgées de 20 à 24 ans est beaucoup plus élevé que dans le cas de n'importe quel autre groupe, sans aucun doute un reflet de l'attachement moindre de ces personnes à un domicile habituel. (Une autre tendance qui se maintient d'un recensement à l'autre.)

Figure 1. Taux de sous-dénombrement, Canada 1966 à 1986

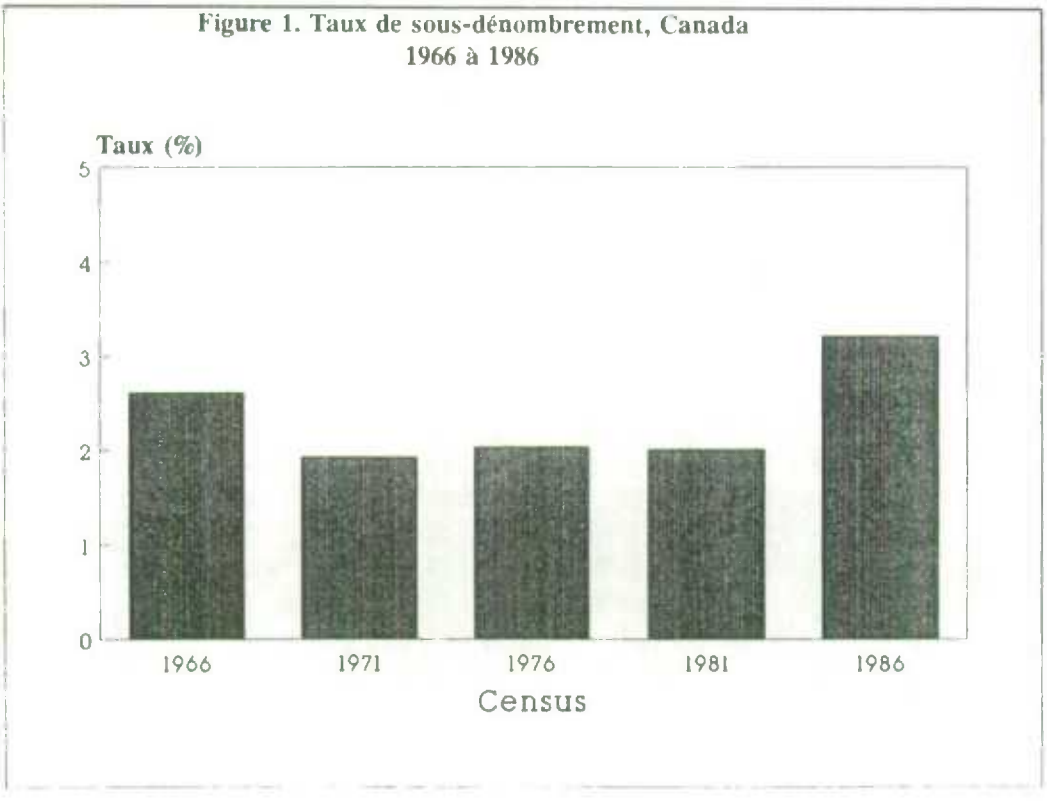


Figure 2. Taux de sous-dénombrement par province 1981 à 1986

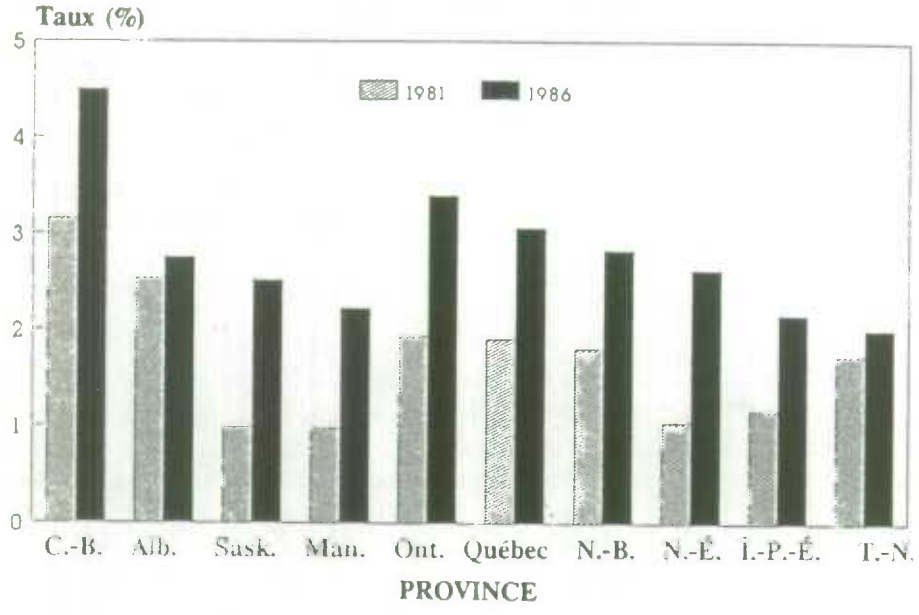
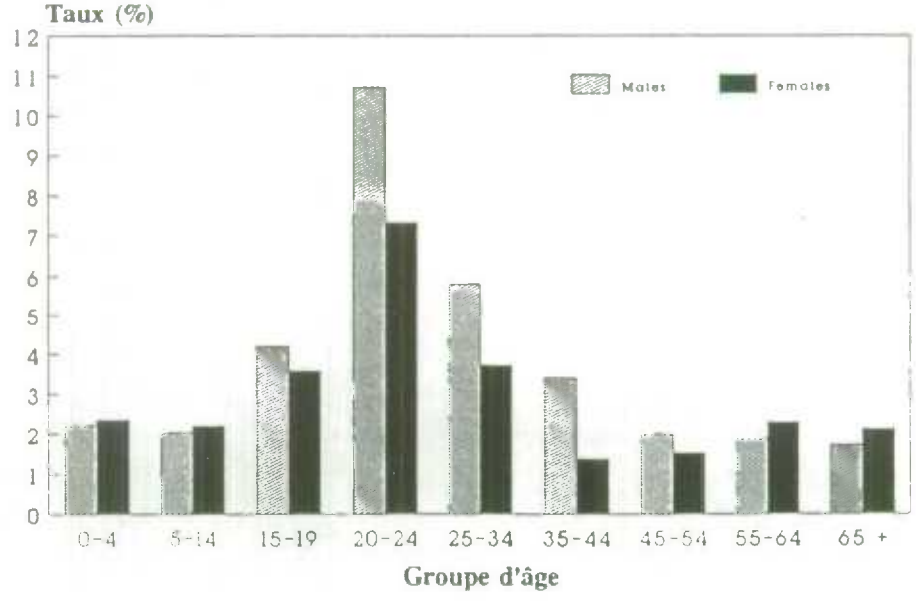


Figure 3. Taux de sous-dénombrement selon l'âge et le sexe 1986



Des résultats de la CVD de 1986 sont présentés plus en détail dans Statistique Canada (1990).

Faiblesses

Bien que la CVD a donné des résultats qui sont plausibles, qui sont raisonnablement compatibles avec d'autres sources de renseignements et qui montrent des tendances cohérentes, cette approche comporte certaines lacunes. D'abord, elle présente certaines faiblesses conceptuelles.

i) La base du recensement renferme des doubles comptes.
ii) Il n'existe, bien sûr, aucune liste complète des personnes oubliées lors du recensement précédent, et l'échantillon tiré de cette base conceptuelle ne se compose que des personnes qui ont été classées dans la catégorie des personnes oubliées au moment de la CVD précédente; par conséquent, les personnes qui n'ont jamais été recensées ne sont pas représentées dans l'échantillon utilisé pour la CVD.
iii) La base des immigrants est restreinte aux immigrants reçus; les citoyens canadiens de retour au pays après un séjour prolongé à l'étranger ne sont pas inclus dans aucune base.
iv) Jusqu'à maintenant, sauf à titre expérimental, aucun échantillon n'a été prélevé du Yukon et des Territoires du Nord-Ouest; par conséquent, les résultats pour les provinces individuelles ainsi que pour l'ensemble du Canada ne sont pas complets.

En étant fondés sur un échantillon (composé d'environ 36,000 personnes en 1986), les résultats sont sujets aux erreurs d'échantillonnage, ce qui limite la quantité de détails pour lesquels on peut présenter des résultats. Dans le cas des provinces moins populeuses, même les taux globaux de sous-dénombrement ont des erreurs d'échantillonnage relativement élevé. (Tandis que le coefficient de variation national se situait à environ 4% en 1986, celui des provinces variait de 6%, pour l'Ontario, à 37%, pour l'Île-du-Prince-Édouard.)

Toutefois, les erreurs possibles non dues à l'échantillonnage comme, par exemple, les biais dus à la non-réponse (causés par le taux de personnes non dépistées qui s'élève à environ 4%) et les erreurs de classement, sont peut-être plus graves. On en sait très peu au sujet de l'importance de ces erreurs. Cependant, il est prudent de dire que la non-réponse a tendance à créer des estimations du sous-dénombrement biaisées négativement (étant donné qu'il est plus probable que les personnes qui sont difficiles à dépister aient été oubliées lors du recensement), contrairement aux erreurs de classement qui créeront un biais positif (puisque les personnes peuvent être classées dans la catégorie des personnes dénombrées seulement si on les trouve sur une formule du recensement, tandis que des personnes qui ont été dénombrées peuvent être classées par erreur dans la catégorie des personnes oubliées lorsque l'adresse obtenue par le dépistage est inexacte).

On trouve une analyse plus complète des faiblesses de la CVD dans Burgess (1988).

2.2 L'étude sur le surdénombrement de 1986

Méthodologie

Malgré ses faiblesses, la CVD est devenue le principal outil pour évaluer la couverture du recensement du Canada. Toutefois, cette méthode permet d'évaluer seulement le sous-dénombrement brut, et non pas l'erreur de couverture nette. Au Canada, on a toujours tenu pour acquis que le surdénombrement était infime par rapport au sous-dénombrement; c'est pourquoi les efforts d'évaluation de la couverture sont concentrés sur ce dernier. Les évaluations de la couverture dans d'autres pays ont toutefois laissé supposer qu'on ne devrait peut-être pas ne tenir aucun compte du surdénombrement, même s'il faut reconnaître qu'il existe des différences considérables entre les pays à la fois dans la façon dont le recensement même est réalisé et dans les méthodes d'évaluation de la couverture. Par exemple, on peut soutenir qu'il y a plus de risques de doubles comptes au recensement des États-Unis, où les questionnaires sont envoyés par la poste, compte tenu notamment des nombreuses méthodes d'amélioration de la couverture qui sont en place (voir Ericksen, 1988). En outre, le rapport surdénombrement/sous-dénombrement tel qu'il est évalué dans l'enquête postcensitaire américaine a tendance à être plus élevé parce qu'on mesure la couverture par rapport à une unité géographique limitée. Par contre, dans la CVD, une personne est classée dans la catégorie des personnes oubliées seulement si elle n'a pas été dénombrée ailleurs au pays (une personne inscrite sur un questionnaire du recensement est considérée comme dénombrée même si elle n'a pas été recensée à l'endroit elle aurait dû l'être). Néanmoins, au cours des

dernières années, on s'est préoccupé davantage du niveau de surdénombrement possible au recensement du Canada et, en 1986, on a réalisé un programme expérimental pour examiner la question des doubles comptes.

Ce programme comportait quatre études:

1) une réinterview dans un échantillon de logements privés afin d'obtenir toutes les autres adresses auxquelles les membres des ménages ont pu être dénombrés, suivie d'une recherche dans les questionnaires du recensement pertinents afin de repérer les personnes qui ont été dénombrées plus d'une fois;
2) des visites dans un petit échantillon de logements collectifs afin d'obtenir les adresses d'autres membres de la famille ou les adresses précédentes des personnes qui ont été dénombrées lors du recensement comme des résidents habituels de l'établissement institutionnel, suivies également d'une recherche dans les formules du recensement pertinentes;
3) l'appariement informatisé d'un échantillon de dossiers du recensement pour déceler les doubles comptes à l'intérieur du même secteur de dénombrement (SD); et
4) l'appariement informatisé de dossiers dans un échantillon de SD contigus afin de repérer les doubles comptes effectués par deux recenseurs voisins (cas qui n'ont pas été détectés dans le petit échantillon étudié).

Résultats

Le tableau 1 présente les résultats de l'étude sur le surdénombrement de 1986. En tout, l'étude a permis de déceler environ 45 600 doubles comptes, dont presque la moitié (22 200) ont été repérés au moyen de la réinterview. (À noter que ce chiffre ne comprend pas les cas qui auraient pu être décelés par l'appariement automatisé.) L'appariement informatisé à l'intérieur des SD a permis de déceler 16 300 doubles comptes. Malgré la couverture restreinte de la composante des logements collectifs, 15% des doubles comptes ont été décelés dans les logements collectifs. Il faut noter que les erreurs types de toutes ces estimations sont assez importantes, et c'est pour cette raison qu'aucune autre répartition des résultats n'a été tentée.

Tableau 1

Étude sur le surdénombrement de 1986 - Résultats

Composante	Estimation	Erreur type
Logements privés - réinterview	22 200	6 050
- appariement automatisé	16 300	3 200
Logements collectifs	7 100	1 350
Total	45 600	6 950

Source : Statistique Canada (1990).

Faiblesses

Les études sur le surdénombrement de 1986 étaient expérimentales et, en tant que telles, très restreintes. L'échantillon de logements privés où l'on a procédé à une réinterview n'était composé que de 11 000 ménages; seulement trois types de logement collectif ont été inclus dans l'étude et l'échantillon était très petit (39 logements collectifs et 1 392 personnes). Les appariements informatisés mettaient chacun en jeu des échantillons de 400 SD. L'appariement avec des ménages avoisinants mettait en jeu un échantillon de 50 ménages dans

chaque SD. Par ailleurs, la méthode utilisée pour mener l'enquête par réinterview était loin d'être idéale. Le taux de non-réponse a été élevé, et il y a eu un plus grand nombre que prévu de personnes qui ont été dénombrées lors du recensement, mais pas durant la réinterview, pour lesquelles il était impossible de déterminer si elles ont été correctement recensées ou non.

Bien qu'on ait acquis de l'expérience très utile dans la détection du surdénombrement, on ne peut pas considérer les résultats comme définitifs. On devrait peut-être dire qu'ils permettent d'abaisser les limites.

2.3 Options envisagées pour 1991

Compte tenu de l'expérience considérable que Statistique Canada a acquise dans la contre-vérification des dossiers et des risques que comporte le passage à une autre méthodologie, on a décidé à un stade relativement précoce de la planification de conserver la CVD comme pierre angulaire du programme de mesure de la couverture de 1991. On a aussi décidé, bien entendu, d'examiner la possibilité d'apporter certaines améliorations à la CVD, notamment : l'inclusion du Yukon et des Territoires du Nord-Ouest, une augmentation générale de la taille des échantillons, des méthodes d'appariement et de dépistage plus efficaces et un suivi des personnes apparemment oubliées afin de vérifier l'adresse obtenue par le dépistage et peut-être afin de découvrir les raisons pour lesquelles ces personnes n'ont pas été recensées.

On a surtout étudié la question de l'inclusion des territoires dans la CVD. Le problème se divise en trois volets:

1) Le dépistage sur le terrain est plus difficile et plus coûteux dans ces régions éloignées et à population clairsemée.

2) La population des deux territoires, mais en particulier celle des Territoires du Nord-Ouest, se compose d'une proportion élevée d'autochtones, dont bon nombre sont difficiles à apparier en raison des variations et des erreurs dans la consignation des noms.

3) Une proportion relativement importante de la population des territoires consiste en des entrants provenant des provinces; sur le plan opérationel, aucune fraction de sondage dans les provinces ne pourrait permettre d'obtenir un échantillon satisfaisant de cette partie de la population des territoires.

Quoiqu'on ne puisse pas faire grand chose au sujet du deuxième problème, des améliorations récentes aux fichiers de l'assurance-maladie des territoires permettent d'espérer obtenir une base de sondage raisonnablement complète et actuelle pour pouvoir surmonter le troisième problème et minimiser le premier en rendant moins nécessaire le dépistage sur le terrain. Grâce à la collaboration de l'administration publique du Yukon, on a obtenu une copie du fichier de l'assurance-maladie de 1986 et, à partir d'un échantillon, on a procédé à un appariement avec le recensement de 1986 afin de déterminer s'il est faisable de mesurer le taux de sous-dénombrement en 1991.

L'autre question importante consistait à savoir quelle était la meilleure façon d'estimer l'erreur de couverture nette. On a songé, d'une part, à améliorer le programme d'études sur le surdénombrement et, d'autre part, à réaliser une enquête postcensitaire (EP) semblable à celle qui est effectuée aux États-Unis. L'avantage de la première option, c'est qu'on pourra profiter de l'expérience acquise en 1986. Les arguments en faveur de l'EP sont les suivants: il s'agit d'une méthode utilisée par la plupart des autres pays qui réalisent une étude sur la couverture par une micro-évaluation fondée sur une enquête-échantillon, la méthodologie utilisée a été considérablement perfectionnée à la suite de recherches menées récemment par le US Bureau of the Census, et une EP permet d'évaluer directement l'erreur de couverture nette et, ainsi, de minimiser le problème de l'intégration des résultats d'études ayant des structures d'erreurs différentes.

Toutefois, après un examen approfondi, on a décidé qu'une EP ne serait pas réalisable dans le contexte du recensement de 1991. L'approche de l'EP est fondée sur deux échantillons d'îlots géographiques se chevauchant. On dresse une liste indépendante des logements et des personnes dans un échantillon de ces îlots, puis on procède à un appariement bilatéral des répondants à l'enquête et des personnes dénombrées lors du recensement. Au besoin, on réalise une interview de suivi pour obtenir les renseignements qui manquent et pour réunir les entités non-appariées. Aux États-Unis, il y a une structure géographique globale fondée sur l'îlot.

Cet îlot, qui se compose en moyenne de 40 à 75 ménages, représente l'unité d'échantillonnage pour l'EP. Au Canada, la plus petite zone de recensement est le SD, qui se compose en moyenne de 350 ménages. Une telle zone serait trop grande pour servir d'unité d'échantillonnage pour l'EP. Le sous-échantillonnage de zones à l'intérieur de SD serait très difficile et sujet à des erreurs, particulièrement dans les régions rurales, sans élaboration et mise à l'essai considérables, ce que l'on a jugé comme étant impossible de réaliser à temps pour 1991. Un autre problème que présente cette approche, c'est l'intégration des procédures d'appariement et de suivi dans le cycle de production du recensement. En raison des contraintes imposées par le cycle de dépouillement du recensement, il ne serait pas possible de réaliser les interviews de suivi avant six à neuf mois après le jour du recensement, ce qui a été jugé inacceptable compte tenu du nombre de personnes ayant déménagé et des risques que les répondants se trompent en essayant de se souvenir. En outre, le coût d'une EP comme celle qui est menée aux États-Unis, en tenant compte du fait que nous nous étions aussi engagés à procéder à une CVD, aurait été prohibitif.

2.4 Les enquêtes d'évaluation de la couverture du recensement de 1991

Objectifs

En 1991, on vise à obtenir une estimation nationale du sous-dénombrement net la plus précise possible et, en même temps, des estimations acceptables à l'échelon provincial. Pour ce qui est de l'erreur d'échantillonnage, une estimation acceptable en serait une dont le coefficient de variation ne dépasserait pas 25%. En même temps, on devra s'occuper des principales faiblesses et des erreurs non dues à l'échantillonnage. Afin d'atteindre cet objectif, un certain nombre d'améliorations sont apportées.

Le sous-dénombrement

La taille de l'échantillon ordinaire pour la CVD passe de 36 000 à 50 000 personnes. La plupart des personnes qui composent cette augmentation résident dans les provinces moins populeuses, cela afin de s'assurer que les coefficients de variation ne dépassent pas 20%. Des bases supplémentaires font l'objet d'un échantillonnage afin de représenter les nouveaux groupes qui sont inclus dans la population cible pour 1991, à savoir les revendicateurs du statut de réfugié et les personnes qui demeurent au Canada en vertu d'un visa d'étudiant, d'un permis de travail ou d'un permis du ministre. Par suite des recherches effectuées dans le fichier de l'assurance-maladie du Yukon, on a décidé de tenter d'inclure dans l'échantillon pour la CVD des personnes vivant au Yukon et dans les Territoires du Nord-Ouest. On prélèvera des échantillons à partir des bases qui sont utilisées d'ordinaire dans la CVD et on les appariera aux fichiers de l'assurance-maladie des territoires. Les non-appariements seront conservés dans l'échantillon, tout comme des échantillons indépendants tirés des fichiers de l'assurance-maladie. Les échantillons ainsi obtenus devraient assurer une bonne couverture des populations des territoires, y compris les migrants intercensitaires et les personnes non protégées par les régimes d'assurance-maladie. On apporte présentement des améliorations à la façon dont on effectue la CVD, dont des modifications au questionnaire de la CVD complété pour les personnes dépistées sur le terrain, l'appariement automatisé avec les dossiers du recensement en fonction du jour, du mois et de l'année exacte de la naissance, et le suivi auprès des personnes qui ont pu être oubliées afin de vérifier les adresses.

Le surdénombrement

On a revu la conception de l'étude sur le surdénombrement en fonction de l'expérience acquise en 1986. Toutefois, comme en 1986, on procédera à une réinterview, à des visites dans des logements collectifs et à un appariement automatisé.

Pour la réinterview, on prélèvera en deux étapes un échantillon d'environ 30 000 ménages (15 ménages dans 2 000 SD). Si le taux de surdénombrement brut est de 0.5%, on prévoit que les coefficients de variation seront de 20 à 40% à l'échelon provincial et de 10% à l'échelon national. L'échantillon sera tiré d'une liste de logements établie par le recenseur dans les SD choisis. Les noms, les adresses, les numéros de téléphone et les renseignements démographiques de base seront transcrits dans les questionnaires de l'étude sur le surdénombrement. Une interview par téléphone permettra d'établir si la personne fait partie de la population cible, son domicile habituel le jour du recensement et toutes les autres adresses auxquelles elle a pu être recensée.

On dépistera les personnes ayant déménagé et on procédera à un suivi sur place auprès des non-répondants et des ménages sans téléphone. Dans les cas où on rejoint la personne et où celle-ci fait partie de l'univers du recensement, cette personne sera considérée comme correctement dénombrée et aucun autre appariement ne sera nécessaire à moins qu'on ait indiqué d'autres adresses, dans lequel cas on effectuera une recherche dans les formules pertinentes du recensement afin d'établir s'il y a eu double compte.

L'enquête dans les logements collectifs sera semblable à celle de 1986, mais elle couvrira tous les types de logement collectif et elle portera sur un plus grand échantillon (environ 500 établissements institutionnels et 1 200 personnes dans des logements collectifs non institutionnels).

Par suite de l'appariement automatisé des personnes faisant partie de ménages voisins qui a été effectué dans le cadre de l'étude de 1986, on a estimé à 16 000 le nombre de doubles comptes, en dépit du fait que seulement huit ménages voisins ont fait l'objet d'une vérification pour chaque ménage choisi. La raison de cette restriction était qu'on ne disposait que du mois et de l'année de naissance, du sexe et de l'état matrimonial comme variables d'appariement (on n'a bien sûr pas saisi les noms). En 1991, on saisira aussi le jour de naissance qui figure sur la formule du recensement, ce qui rendra possible l'appariement de tous les ménages dans le SD. Nous procédons actuellement à des études afin de déterminer la stratégie d'appariement la plus efficace, compte tenu de cette variable d'appariement plus précise. Nous devrions utiliser le même échantillon de SD pour l'appariement automatisé que pour la réinterview, ce qui facilitera l'intégration des deux composantes.

L'étude sur le surdénombrement de 1991 est décrite plus en détail dans Dibbs et Royce (1990).

Faiblesses

Bien que nous croyons que cette stratégie constitue la meilleure approche pour 1991 compte tenu du temps et des ressources dont nous disposons et des risques inhérents à la mise en oeuvre de nouvelles méthodes sans une analyse complète, nous reconnaissons qu'elle comporte certaines lacunes. Demeurent un bon nombre des faiblesses de la CVD que l'on a mentionnées auparavant, telles que les doubles comptes dans la base du recensement, l'absence de couverture des Canadiens de retour au pays et le renouvellement de l'échantillon des personnes oubliées. On réduira quelque peu les erreurs d'échantillonnage, et nous espérons abaisser également le nombre d'erreurs de classement. L'inclusion des territoires dans l'évaluation du sous-dénombrement permettra d'éliminer une autre lacune de la CVD. En outre, nous procédons présentement à un perfectionnement des procédures relatives à l'étude sur le surdénombrement et à l'augmentation de la taille des échantillons. Par ailleurs, nous étendons la portée de ces études à la détection des cas de dénombrement de personnes ne faisant pas partie du champ d'observation ainsi que des doubles comptes. Néanmoins, des erreurs et des biais demeureront. Une faiblesse de cette stratégie réside dans le fait qu'il est difficile, sinon impossible, de «contrebalancer» les erreurs afférentes aux diverses études. Par exemple, les erreurs de classement risquent de créer des estimations du sous-dénombrement biaisées positivement et les estimations du surdénombrement biaisées négativement. Par conséquent, les estimations du sous-dénombrement net peuvent être biaisées à la hausse (même s'il y a bien sûr d'autres composantes de l'erreur qui peuvent provoquer un biais dans le sens opposé). Le USBC prépare son EP de façon à s'assurer que les biais dans les estimations du sous-dénombrement et du surdénombrement aient tendance à se compenser l'un l'autre.

3. ESTIMATIONS DE L'ERREUR DE COUVERTURE OBTENUES PAR DES MÉTHODES DÉMOGRAPHIQUES

Comme il a été mentionné dans l'introduction, on peut aussi avoir recours à des méthodes démographiques pour évaluer l'erreur de couverture. Ces méthodes comptent principalement sur des sources de données administratives telles que l'enregistrement d'événements démographiques, les documents d'immigration et peut-être les fichiers de l'impôt, de l'assurance-maladie et de la sécurité sociale. Ces sources peuvent être utilisées pour obtenir des estimations des chiffres de population à un moment donné ou des estimations de la variation de la population au cours d'une période. Au Canada, bien que certains dossiers administratifs permettent de produire des estimations de la population de haute qualité pour certains sous-groupes de la population comme, par exemple, les dossiers sur les allocations familiales qui donnent de bonnes estimations de la population âgée de 14 ans et moins (Fortier et Raby, 1990), il n'existe aucun système de dossiers administratifs duquel on peut

produire des estimations du chiffre total de la population. Pour assurer une couverture intégrale, il est nécessaire de se fier au recensement; on se sert ensuite des dossiers administratifs pour produire des estimations de la variation de la population durant la période intercensitaire. Ces estimations sont établies à partir des soi-disant «composantes de la variation», c'est-à-dire les naissances, les décès et les migrations (immigration, émigration et migration interne). (Voir Statistique Canada, 1987a.) L'enregistrement des naissances et des décès est pratiquement complet au Canada, et les dossiers de l'immigration sont fiables en ce qui concerne les «immigrants reçus». (Jusqu'au recensement de 1986 inclusivement, seuls les citoyens canadiens et les immigrants reçus étaient inclus dans la population cible. En 1991, on inclura aussi d'autres immigrants ayant un statut temporaire mais relativement à long terme au Canada, comme les revendicateurs du statut de réfugié, les étudiants étrangers et les détenteurs de permis de travail.) Cependant, on ne possède pas beaucoup de données sur les Canadiens de retour au pays après un séjour prolongé à l'étranger (voir Fortier, 1990). Les estimations du taux d'émigration sont moins précises que les statistiques d'immigration, étant donné que ces estimations sont basées en partie sur les dossiers de l'immigration d'autres pays et sur des inférences (fondées sur des modèles) tirées des fichiers de l'impôt et des allocations familiales. Les estimations nationales de la variation de la population basées sur ces composantes sont jugées raisonnablement précises. Toutefois, en deçà de l'échelon national, on doit incorporer une composante additionnelle: la migration interne. Les estimations de la migration interne sont elles aussi basées sur des inférences tirées des fichiers de l'impôt et des allocations familiales. Les estimations provinciales de la variation démographique intercensitaire sont considérablement moins fiables que les estimations nationales, en particulier dans le cas des provinces ayant un taux de migration élevé. Il faut noter que les groupes très mobiles comme, par exemple, les jeunes hommes adultes, qui ont tendance à être oubliés au recensement sont aussi susceptibles d'être sous-représentés dans les fichiers de l'impôt et des allocations familiales et, par conséquent, il est probable que la migration interprovinciale de ces groupes soit sous-estimée.

Une évaluation de l'erreur de couverture peut être obtenue en calculant une «estimation de la population» dans la présente année de recensement et en la comparant avec le chiffre du recensement. On obtient l'estimation de la population en ajoutant les composantes de l'accroissement démographique au chiffre du recensement précédent. La différence entre les estimations de la population et les chiffres du recensement s'appelle l'«erreur en fin de période», et elle donne une estimation de la variation du sous-dénombrement net d'un recensement au suivant. L'estimation est sujette aux mêmes erreurs que celles qui sont afférentes à l'estimation démographique de l'accroissement. (Voir aussi Romaniuc, 1988, pour obtenir une autre estimation démographie de l'erreur de couverture.)

Le tableau 2 montre les valeurs de l'erreur en fin de période pour chacune des dix provinces canadiennes pour 1981 et 1986. Comme on l'a indiqué précédemment, il est en général admis que, à l'échelon national (c'est-à-dire le total des dix provinces), l'erreur en fin de période peut être considérée comme raisonnablement précise. D'après le tableau 2, le sous-dénombrement net a baissé entre 1976 et 1981, et il a augmenté en 1986. Les estimations provinciales sont moins fiables en raison des erreurs dans l'estimation de la migration interprovinciale, ce qui est particulièrement évident en ce qui concerne l'Alberta en 1981 et Terre-Neuve en 1986. Comme il a été mentionné auparavant, les erreurs dans l'estimation de la migration interprovinciale peuvent être dues en grande partie à la sous-représentation des jeunes adultes tant dans les fichiers de l'impôt que dans ceux des allocations familiales, quoiqu'il n'existe aucune preuve solide pour appuyer cette hypothèse.

Tableau 2

Erreur en fin de période pour 1981 et 1986

Province	1981		1986	
	#	%	#	%
T.-N.	7 100	1.25	11 500	2.02
Î.-P.É.	-400	-0.31	1 300	1.06
N.-É.	-200	-0.03	11 100	1.28
N.-B.	-2 000	-0.28	11 200	1.57
Québec	-37 600	-0.58	87 700	1.34
Ont.	31 500	0.37	66 900	0.73
Man.	8 600	0.83	6 100	0.57
Sask.	-5 000	-0.52	10 700	1.06
Alb.	-53 900	-2.41	19 400	0.81
C.-B.	-6 000	-0.22	16 800	0.58
Total*	-58 000	-0.24	242 700	0.96

Source : Statistique Canada (1987b, 1988)

* Total pour les dix provinces (c'est-à-dire le Canada à l'exclusion du Yukon et des Territoires du Nord-Ouest). En raison de l'erreur d'arrondissement, il se peut que l'addition des chiffres provinciaux ne soit pas égale au total.

\# Estimation démographique - chiffre du recensement.

% Erreur en fin de période en pourcentage du chiffre du recensement.

4. ESTIMATION DU SOUS-DÉNOMBREMENT NET DANS LES RECENSEMENTS ANTÉRIEURS À CELUI DE 1991

Compte tenu des erreurs afférentes à cette méthode, pour quelles raisons s'intéresse-t-on aux estimations de l'erreur de couverture obtenues au moyen de méthodes démographiques? Nous allons, après tout, évaluer directement le sous-dénombrement et le surdénombrement en 1991, et nous devrions nous attendre à obtenir des résultats suffisamment précis pour estimer le sous-dénombrement à l'échelon provincial. Il y a plusieurs raisons pour lesquelles nous envisageons d'autres approches comme les estimations produites à l'aide de méthodes démographiques. D'abord, on peut les obtenir plus tôt que les estimations fondées sur des enquêtes, et on peut être averti rapidement d'un problème possible. Ensuite, ces estimations permettent de vérifier les estimations fondées sur des enquêtes. Les estimations du sous-dénombrement obtenues au moyen de méthodes démographiques ont sûrement été très utiles en 1986 pour confirmer l'augmentation importante des estimations dérivées de la CVD. En outre, comme il est démontré ci-après, la comparaison des estimations produites par des méthodes démographiques et de celles qui sont fondées sur des enquêtes peut attirer l'attention sur des problèmes dans l'une des sources, ou dans les deux. Par ailleurs, bien que nous nous attendons à avoir pour le recensement de 1991 des estimations (fondées sur des enquêtes) du sous-dénombrement net qui sont fiables, nous ne disposons pas d'estimations comparables pour les recensements précédents. Pour de nombreuses raisons, les tendances dans l'erreur de couverture sont aussi importantes que les niveaux absolus, et il serait utile de disposer d'estimations du sous-dénombrement net pour les recensements antérieurs. Bien entendu, nous disposons d'estimations obtenues par la CVD pour les recensements à partir de celui de 1966. Les échantillons sont moins gros que pour le recensement de 1991, mais les principaux problèmes en ce qui concerne les estimations du sous-dénombrement sont l'absence d'estimations pour le Yukon et les Territoires du Nord-Ouest et les changements dans la population cible du recensement en 1991. Toutefois, le problème qui pose le plus de difficultés, c'est l'estimation du surdénombrement dans les recensements antérieurs à celui de 1991. L'étude de 1986 n'était pas de grande envergure et ses résultats ne sont pas jugés suffisamment fiables pour être utilisés

afin d'estimer le sous-dénombrement net en 1986; de plus, aucune étude sur le surdénombrement n'a été réalisée pour les recensements antérieurs. En l'absence d'estimations directes fondées sur des enquêtes, il faut envisager l'adoption de méthodes indirectes.

Il y a, en fait, plusieurs approches possibles:

a) Si le surdénombrement en 1991 est négligeable, nous pourrions supposer que le surdénombrement lors des recensements antérieurs l'était également.

b) Même si le surdénombrement n'est pas négligeable, nous pourrions supposer qu'il est demeuré constant.

c) Nous pourrions supposer que le surdénombrement est constant par rapport à une autre variable telle que le sous-dénombrement, le chiffre du recensement ou le chiffre réel de la population.

d) Nous pourrions tenter d'élaborer un modèle de surdénombrement en fonction de caractéristiques démographiques connues. Toutefois, cette approche ne s'est pas avérée fructueuse dans le cas du sous-dénombrement.

e) Nous pourrions utiliser une estimation produite au moyen d'une méthode démographique.

Si on adopte la dernière approche, il faudra alors soustraire les composantes de l'accroissement de l'estimation de la population en 1991 (c'est-à-dire le chiffre du recensement de 1991 plus le sous-dénombrement net tel qu'il est estimé au moyen de la CVD et de l'étude sur le surdénombrement). On comparerait ensuite l'estimation de la population ainsi obtenue avec le chiffre du recensement correspondant afin d'obtenir le sous-dénombrement net implicite. Une telle estimation serait sujette à deux composantes de l'erreur: l'erreur dans l'estimation du sous-dénombrement net au recensement de 1991 et l'erreur dans l'estimation de l'accroissement démographique. Même si cette approche permet de produire une estimation du sous-dénombrement net, on peut obtenir une estimation implicite du surdénombrement en soustrayant le sous-dénombrement net de l'estimation du sous-dénombrement obtenue par la CVD pour le recensement en question.

Cette approche équivaut à estimer la variation du surdénombrement en soustrayant l'erreur en fin de période du taux de sous-dénombrement tel qu'elle a été estimée par la CVD. Afin d'évaluer l'utilité de cette approche, examinons les estimations correspondantes de la variation du surdénombrement entre 1976 et 1981 et entre 1981 et 1986 (voir le tableau 3). (On trouve au tableau 4 les estimations du sous-dénombrement fournies par la CVD pour 1976, 1981 et 1986.) Ces estimations du surdénombrement sont sujettes à des erreurs dans les estimations de la variation du sous-dénombrement obtenues par la CVD ainsi qu'à des erreurs dans l'estimation démographique de l'accroissement. Le tableau 3 donne aussi les estimations de l'erreur type de la variation dans les estimations du sous-dénombrement obtenues par la CVD. Il est clair qu'il faut tenir compte de cette composante de l'erreur. En 1986, par exemple, seuls Terre-Neuve, l'Ontario et l'ensemble des dix provinces ont des estimations de la variation du surdénombrement plus élevées que deux erreurs types.

Tableau 3

Variations implicites du surdénombrement entre 1976 et 1981 et entre 1981 et 1986 (ainsi que les erreurs types approximatives pour la variation dans les estimations du sous-dénombrement obtenues par la CVD)

Province	1976 - 1981		1981 - 1986	
T.-N.	-3 300	(3 400)	-9 800 #	(3 200)
Î.-P.-É.	1 500	(7 600)	-200	(1 300)
N.-É.	2 000	(4 100)	3 500	(4 500)
N.-B.	-100	(3 400)	-3 400	(3 500)
Québec	-26 900	(21 300)	-6 800	(24 500)
Ont.	12 300	(19 000)	83 000 #	(22 100)
Man.	-9 500	(5 000)	8 100	(5 700)
Sask.	2 300	(2 400)	5 600	(5 300)
Alb.	84 400 #	(9 800)	-10 600	(11 900)
C.-B.	15 600	(12 300)	29 800	(15 500)
Total*	78 600 #	(33 700)	99 300 #	(40 900)

* Total pour les dix provinces (c'est-à-dire le Canada à l'exclusion du Yukon et des Territoires du Nord-Ouest). En raison de l'erreur d'arrondissement, il se peut que l'addition des chiffres provinciaux ne soit pas égale au total.

\# Statistiquement significatif au niveau 5%.

Tableau 4

Estimations (par la CVD) du sous-dénombrement aux recensements de 1976, 1981 et 1986 (erreurs types entre parenthèses)

Province	1976		1981		1986	
T.-N.	6 200	(2 200)	10 000	(2 600)	11 700	(1 900)
Î.-P.É.	400	(300)	1 500	(700)	2 800	(1 100)
N.-É.	7 200	(2 900)	9 000	(2 900)	23 600	(3 500)
N.-B.	15 000	(2 600)	12 900	(2 200)	20 700	(2 700)
Québec	189 700	(16 200)	125 200	(13 800)	206 100	(20 200)
Ont.	127 200	(14 200)	171 000	(12 600)	320 900	(18 200)
Man.	11 100	(3 500)	10 200	(3 600)	24 400	(4 400)
Sask.	12 400	(3 200)	9 700	(3 700)	26 000	(3 800)
Alb.	27 800	(4 900)	58 300	(8 500)	67 100	(8 300)
C.-B.	79 800	(7 800)	89 400	(9 500)	136 000	(12 300)
Total*	476 700	(23 900)	497 300	(23 800)	839 300	(33 300)

* Total pour les dix provinces (c'est-à-dire le Canada à l'exclusion du Yukon et des Territoires du Nord-Ouest). En raison de l'erreur d'arrondissement, il se peut que l'addition des chiffres provinciaux ne soit pas égale au total.

Malgré les problèmes que posent ces estimateurs, en particulier à l'échelon provincial, ces résultats sont utiles et peuvent aider à évaluer les estimations (fondées sur des enquêtes) de l'erreur de couverture. Par exemple, si nous pouvons considérer les erreurs dans les composantes de l'accroissement démographique à l'échelon

national comme négligeables, le sous-dénombrement net a diminué entre 1976 et 1981 de 58 000 personnes, et il a augmenté entre 1981 et 1986 de 242 700 personnes. Les estimations du sous-dénombrement dérivées de la CVD ont été les suivantes: 476 700 personnes (2.04%) en 1976, 497 000 personnes (2.01%) en 1981 et 839 000 personnes (3.21%) en 1986. Ainsi, le nombre de personnes dénombrées en trop a augmenté implicitement de 78 600 entre 1976 et 1981, et de 99 300 entre 1981 et 1986. Étant donné que le surdénombrement ne peut pas être inférieur à zéro, ceci laisse supposer que le surdénombrement a été d'au moins 177 900 personnes (0.7%) en 1986. L'étude expérimentale sur le surdénombrement menée en 1986 n'a permis de repérer que 45 600 personnes (0.2%) dénombrées en double (avec une erreur type de 6 950). Ou l'on a de beaucoup sous-estimé le surdénombrement dans l'étude, ou il y a une erreur importante à l'échelon national dans les composantes de la variation, ou, encore, on a surestimé l'augmentation du sous-dénombrement au moment de la CVD. (Il faut noter, cependant, que l'intervalle de confiance de 95% pour la hausse du sous-dénombrement entre 1976 et 1986 est de 362 600 ± 82 000. Si nous prenons le chiffre le moins élevé de l'intervalle, soit 280 600, l'augmentation implicite du nombre de personnes dénombrées en trop entre 1976 et 1986 ne serait que de 95 900. Entre 1981 et 1986, le même intervalle pour la hausse du sous-dénombrement est de 342 000 ± 81 800, le chiffre le moins élevé de cet intervalle laissant supposer une augmentation du nombre de personnes dénombrées en trop de seulement 17 500.)

À l'échelon provincial, seulement trois des estimations qui figurent au tableau 3 sont statistiquement significatives. Encore une fois, elles laissaient supposer soit un surdénombrement relativement élevé, soit des erreurs importantes dans les estimations de l'accroissement démographique. Comme nous l'avons vu pour l'Alberta en 1981 et pour Terre-Neuve en 1986, les erreurs d'estimation de la migration interprovinciale nette expliquent probablement l'estimation apparemment élevée de la variation du surdénombrement.

Par conséquent, il semblerait que les estimations implicites des variations du surdénombrement à l'échelon provincial ne sont pas fiables. En supposant que les estimations du sous-dénombrement aux recensements antérieurs à celui de 1991 qui ont été obtenues par la CVD ne doivent pas changer, l'approche démographique pour évaluer le sous-dénombrement net lors de ces recensements n'est par conséquent pas possible à l'échelon provincial. Toutefois, on pourrait peut-être utiliser cette approche à l'échelon national et désagréger l'estimation du surdénombrement parmi les provinces en proportion du chiffre du recensement ou de l'estimation du sous-dénombrement se rapportant à chacune d'elles (comme dans l'option c) ci-dessus).

5. CONCLUSIONS

Notre but est de disposer pour la première fois au Canada d'estimations fiables de l'erreur de couverture nette. Nous croyons que, si tout va comme prévu, le programme que nous avons mis en place permettra d'atteindre ce but. La question est à savoir ce que nous ferons avec ces estimations. Nous savons bien que le redressement des chiffres du recensement à partir de l'erreur de couverture nette est une question majeure au États-Unis. Même s'il est moins politisé au Canada, il y a ici aussi un débat pour savoir s'il est prudent de redresser les chiffres du recensement. Cependant, compte tenu du fait que nous sommes loin d'avoir amélioré nos méthodes d'évaluation de l'erreur de couverture nette, certainement à l'échelon infraprovincial, et compte tenu du temps qu'il faut pour réaliser des études sur la couverture et pour en évaluer les résultats, Statistique Canada n'a pas l'intention de redresser les chiffres du recensement de 1991 en fonction des estimations de l'erreur de couverture nette décrites précédemment. Ce redressement peut amener de meilleures estimations pour certaines fins, mais pas pour d'autres. Nous préférons divulguer les chiffres du recensement obtenus par les méthodes traditionnelles et mettre à la disposition des utilisateurs les renseignements que nous obtenons sur les erreurs de couverture de manière à ce qu'ils puissent faire leur propre interprétation informée des données du recensement. Par contre, nous examinons la possibilité d'incorporer les renseignements sur l'erreur de couverture nette dans les estimations démographiques officielles. À l'heure actuelle, ces estimations sont fondées sur les chiffres du recensement et, durant les périodes intercensitaires, sur les composantes de l'accroissement démographique qui ont été décrites précédemment. Des recherches sont en cours afin d'examiner la possibilité d'éliminer la contrainte voulant que l'estimation de la population pour le mois de juin d'une année de recensement devrait constituer le chiffre du recensement que l'on publie. On pourrait plutôt tenir compte de l'erreur de couverture. On étudie présentement les conséquences possibles d'un tel changement ainsi que la question du niveau d'unité géographique auquel de telles estimations sont fiables. Avant de prendre une décision définitive, il faut attendre les résultats des études sur la mesure de la couverture et l'évaluation de la qualité de ceux-ci.

REMERCIEMENTS

J'aimerais remercier les collègues suivants de Statistique Canada qui m'ont fait part de leurs observations utiles lors de la rédaction du présent document: Céline Fortier, Edward Pryor, Ronald Raby, Don Royce et Ravi Verma.

BIBLIOGRAPHIE

Burgess, R. (1988). Évaluation des estimations du sous-dénombrement obtenues par la contre-vérification des dossiers du recensement du Canada, *Techniques d'enquête*, 14, 2, 147-167.

Dibbs, R., et Royce, D. (1990). Measuring Overcoverage in the 1991 Census of Canada, présenté à la réunion de l'American Statistical Association, Government Statistics Section, Anaheim (Californie).

Ericksen, E. (1988). Bayes Estimates of Population Undercount for Local Areas, *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Fortier, C. (1990). Canadiens de retour: nombre (1981-1989) et caractéristiques (1988 et 1989), rapport interne, Division de la démographie, Statistique Canada.

Fortier, C., et Raby, R. (1989). Évaluation de la qualité des données par âge et sexe du recensement canadien de 1986, *Cahiers québécois de démographie*, 18, 2.

Romaniuc, A. (1988). Une approche démographique à l'évaluation du recensement de 1986 et des estimations de population pour le Canada, *Techniques d'enquête*, 14, 2, 169-185.

Statistique Canada (1987a). Méthodes d'estimation de la population, n° du catalogue 91-528.

Statistique Canada (1987b). Estimations annuelles postcensitaires de la population suivant l'état matrimonial, l'âge, le sexe et composantes de l'accroissement, Canada, provinces et territoires, 1er juin 1986, 4, quatrième numéro, n° du catalogue 91-210.

Statistique Canada (1988). Estimations annuelles postcensitaires de la population suivant l'état matrimonial, l'âge, le sexe et composantes de l'accroissement, Canada, provinces et territoires, 1er juin 1988», 6, sixième numéro, n° du catalogue 91-210.

Statistique Canada (1990). Guide à l'intention des utilisateurs sur la qualité des données du recensement de 1986: Couverture, n° du catalogue 99-135.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

PRÉDICTION DU SOUS-DÉNOMBREMENT POUR LES PETITES RÉGIONS À L'AIDE DU MODÈLE LINÉAIRE GÉNÉRAL

N. Cressie[1]

RÉSUMÉ

Supposons qu'on divise le pays en n petites régions, affichant chacune un certain taux de sous-dénombrement (non observable). Supposons que la variable de sous-dénombrement est une combinaison linéaire de variables explicatives, plus un vecteur d'erreurs de moyenne nulle dont la matrice des variances $\Gamma(\gamma)$ est fonction de paramètres γ. Toutefois, les données de sous-dénombrement comportent une autre composante d'erreur, ayant une matrice des variances Δ, connue à partir de facteurs ayant trait ou non à l'échantillonnage. Si on connaît $\Gamma(\gamma)$ et Δ, on peut calculer des prédicteurs linéaires optimaux pour les petites régions. Dans le présent article, nous ajustons un modèle spatial pour $\Gamma(\gamma)$ aux données du recensement des É.-U. de 1980 et de l'enquête postcensitaire de 1980 afin de permettre le calcul de prédicteurs pour les petites régions.

MOTS CLÉS: Facteurs de redressement; hétéroscédasticité; estimation par la méthode du maximum de vraisemblance; erreur quadratique moyenne de prédiction; dépendance spatiale.

1. INTRODUCTION

Le présent article débute par un examen des résultats relatifs à la prédiction du sous-dénombrement pour les petites régions, obtenus à l'aide du modèle linéaire général proposé dans Cressie (1990). Ce modèle est assez général pour qu'on puisse considérer tant l'hétéroscédasticité que la dépendance spatiale comme des cas particuliers. Nous étudierons donc l'incidence de l'introduction de la dépendance spatiale pour l'estimation du sous-dénombrement au recensement des É.-U. de 1980.

Actuellement, il existe deux approches étroitement reliées pour prédire le sous-dénombrement: celle de la stratification (Cressie, 1988, 1989) et celle de la régression (Ericksen et Kadane, 1985; National Academy of Sciences, 1985; Diffendal, 1988; Cressie, 1990). Supposons que le pays se divise en $i = 1,...,n$ régions (par ex. les États, y compris le District de Columbia). L'objectif visé consiste à prédire le sous-dénombrement dans chacune de ces régions à partir d'un échantillon obtenu au moyen d'une estimation de système dual (par ex., Wolter, 1986). Supposons en outre que chaque région est divisée en jusqu'à J strates, à l'intérieur desquelles on croit que le «mécanisme» de sous-dénombrement est homogène. Si le chiffre du recensement et le chiffre réel pour la sous-région (j,i) sont respectivement C_{ji} et T_{ji}, alors la proportion de la population n'ayant pas été denombrée est

$$U_{ji} = (T_{ji} - C_{ji})/T_{ji}\ , \tag{1.1}$$

et on définit le facteur de redressement par

[1] N. Cressie, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A.

$$F_{ji} \equiv T_{ji}/C_{ji} . \tag{1.2}$$

À l'évidence, $F_{ji} = 1/\{1 - U_{ji}\}$ et la première quantité est donc une fonction monotone de l'autre. Bien que les communiqués de presse et les rapports finaux fassent d'ordinaire état de la variable de sous-dénombrement (sous forme de pourcentage), il arrive souvent, pour des raisons de cohérence du modèle, que l'analyse statistique porte sur le facteur de redressement.

Définissons :

$$T_i \equiv \sum_{j=1}^{J} T_{ji} \; ; \; C_i \equiv \sum_{j=1}^{J} C_{ji} \; ; \; F_i \equiv T_i/C_i \; ; \; i = 1,...,n. \tag{1.3}$$

On notera que $T_i = \sum_{j=1}^{J} F_{ji}C_{ji}$, ce qui explique pourquoi F_{ji} est appelé facteur de redressement. Supposons qu'on obtienne au moyen d'une inférence statistique basée sur les données de l'enquête postcensitaire (EP) (ou sur toute autre source de données) des prédicteurs $\{F_{ji}^{prd}\}$; alors le prédicteur du chiffre réel de population de la région i est

$$T_i^{prd} \equiv \sum_{j=1}^{J} F_{ji}^{prd} C_{ji} . \tag{1.4}$$

La méthode de stratification suppose que $\{F_{ji} : i = 1,...,n\}$ sont des variables aléatoires dont la moyenne μ_j est un paramètre devant être estimé; $j = 1,...,J$. La méthode de régression habituelle suppose que F_i, définie par l'équation (1.3), est une variable aléatoire dont la moyenne est $\sum_{k=1}^{p} \beta_k z_{ki}$; $i = 1,...,n$; où $\{z_{ki} : k = 1,...,p\}$ sont les variables explicatives données et $\{\beta_k : k = 1,...,p\}$ sont les paramètres à estimer. Le modèle vraiment général, dont chacun de ces modèles constitue un cas particulier, suppose que la moyenne de F_{ji} est $\sum_{k=1}^{p_j} \beta_{kj} z_{kji}$; $j = 1,...,J$, $i = 1,...,n$. En d'autres termes, il permet de définir un modèle de régression distinct pour chaque strate. Par exemple, les noirs, les blancs, les non-noirs d'origine hispanique et les autres groupes similaires présentent des caractéristiques de sous-dénombrement distinctes qui devraient nécessiter l'utilisation de modèles de régression distincts. La méthode de stratification est le cas particulier où $p_j = 1$ et $Z_{1ji} = 1$, et la méthode de régression est le cas particulier où $p_j = p$ et $\beta_{kj} = \beta_k$ (ou le cas particulier où $J = 1$).

Une caractéristique importante du modèle de stratification proposé par Cressie (1988, 1989) est l'hypothèse

$$\text{var}(F_{ji}) = \tau_j^2/C_{ji} , \tag{1.5}$$

pour laquelle ce dernier donnait une justification à la fois bayésienne et fréquentiste (pour l'utilisation d'une pondération inversement proportionnelle aux chiffres du recensement). Intuitivement, on peut dire que l'équation (1.5) est sensée, puisqu'il est probable que les facteurs de redressement varient moins pour les régions les plus populeuses. Si on adopte l'équation (1.5) comme hypothèse, les variances du modèle de régression s'obtiennent par

$$\text{var}(F_i) = v_i^2 = (\sum_{j=1}^{J} (C_{ji}/C_i)\tau_j^2)/C_i \,. \tag{1.6}$$

Pour autant que $\sum_{j=1}^{J} (C_{ji}/C_i)\tau_j^2$ ne varie pas beaucoup sur l'ensemble des n régions il s'ensuit que, approximativement,

$$\text{var}(F_i) = \tau^2/C_i. \tag{1.7}$$

Cressie (1990) démontre que, pour les données de 1980, il importe peu que l'on choisisse l'équation (1.6) ou l'équation (1.7) comme hypothèse. Il existe une différence plus marquée entre l'équation (1.7) et l'hypothèse d'homoscédasticité,

$$\text{var}(F_i) = \tau^2. \tag{1.8}$$

Notons que les équations (1.6), (1.7) et (1.8) posent implicitement comme hypothèse que $cov(F_i, F_{i'}) = 0$ lorsque $i \neq i'$. Cependant, il est possible de réaliser des progrès considérables en supposant seulement que $var\{(F_1,...,F_n)'\}$ est définie non négative.

Après avoir examiné les résultats obtenus par Cressie (1990) à la section 2, nous consacrerons la section 3 à l'application de ces résultats à un modèle linéaire spatial pour $(F_1,..., F_n)'$, en utilisant la méthode du maximum de vraisemblance pour estimer les paramètres inconnus. La section 4 présente des prédicteurs optimaux des données de sous-dénombrement de 1980 fondés sur un modèle spatial (hétéroscédastique). Enfin, on trouve à la section 5 un exposé des conclusions qu'on peut tirer de la présente analyse.

2. PRÉDICTION EMPIRIQUE DE BAYES AVEC LE MODÈLE LINÉAIRE GÉNÉRAL

Nous considérons que la population réelle de toute strate définie au sein d'un État est inconnue. Une fois que les chiffres du recensement correspondants sont connus, on met à jour les estimations de la population réelle. Les modèles que nous allons construire et les inférences que nous allons tirer reposent donc sur les observations du recensement.

Dans la suite du présent article, nous allons adopter pour modéliser le sous-dénombrement la méthode de régression dont nous avons fait état à la section 1. Les personnes ayant contribué à l'élaboration et à l'étude de cette approche sont Ericksen et Kadane (1985), Freedman et Navidi (1986), Diffendal (1988), Ericksen, Kadane et Tukey (1989), et Cressie (1990). Le modèle linéaire général que nous adoptons dans cet article comprend tous les modèles élaborés par ces auteurs à titre de cas particuliers.

On se souviendra que le facteur de redressement pour la petite région i est,

$$F_i = T_i/C_i \,. \tag{2.1}$$

En conséquence, le sous-dénombrement pour cette région est

$$U_i = (T_i - C_i)/T_i \tag{2.2}$$

$$= 1 - F_i^{-1} \; ; \; i = 1,...,n \,. \tag{2.3}$$

Comme la variable de sous-dénombrement est une fonction monotone biunivoque de la variable du facteur de redressement, mathématiquement, il importe peu qu'on choisisse de modéliser l'une ou l'autre. Toutefois, d'un point de vue statistique, la modélisation et la prédiction des facteurs de redressement présentent certains avantages.

Supposons que $\underset{\sim}{F} \equiv (F_1,...,F_n)'$ est le vecteur nx1 des facteurs de redressement à prédire. Notons que si deux petites régions i et i' sont groupées pour former $i \cup i'$, alors

$$F_{i \cup i'} = \{C_i F_i + C_{i'} F_{i'}\}/\{C_i + C_{i'}\} . \tag{2.4}$$

Ainsi, l'agrégation est donc compatible avec les modèles linéaires et (de façon moins importante) avec les hypothèses de distribution gaussienne.

La notation, $\underset{\sim}{Z} \sim Gau(\underset{\sim}{\mu}, \Sigma)$, indique que le vecteur colonne $\underset{\sim}{Z}$ possède une distribution gaussienne à plusieurs variables avec une moyenne $\underset{\sim}{\mu}$ et une matrice des variances Σ. Supposons que la distribution a priori des facteurs de redressement est définie par,

$$\underset{\sim}{F} \sim Gau(X\underset{\sim}{\beta}, \Gamma(\underset{\sim}{\gamma})), \tag{2.5}$$

où X est une matrice nxp de variables explicatives, $\underset{\sim}{\beta}$ est un vecteur px1 de coefficients inconnus, et $\Gamma(\underset{\sim}{\gamma})$ est une matrice des variances nxn fonction d'un vecteur kx1 de paramètres $\underset{\sim}{\gamma}$ inconnus. Les variables explicatives considérées peuvent être le pourcentage de personnes faisant partie d'un groupe minoritaire, le pourcentage de ménages locataires, le degré de pauvreté, etc.

Comme on ne peut obtenir $\underset{\sim}{F}$ sous forme de données, il faut utiliser des observations imparfaites sur les facteurs de redressement. Par exemple, les estimateurs de système dual obtenus à partir d'une enquête postcensitaire (EP) nous donnent le vecteur d'observations $\underset{\sim}{Y}$. Supposons que, suivant $\underset{\sim}{F}$,

$$\underset{\sim}{Y} | \underset{\sim}{F} \sim Gau(\underset{\sim}{F}, \Delta), \tag{2.6}$$

où Δ est une matrice nxn de variances et de covariances d'échantillonnage connues.

Si $\underset{\sim}{p}_1(\underset{\sim}{Y})$ et $\underset{\sim}{p}_2(\underset{\sim}{Y})$ sont deux prédicteurs de $\underset{\sim}{F}$, on dit que $\underset{\sim}{p}_1(\underset{\sim}{Y})$ est aussi valable que $\underset{\sim}{p}_2(\underset{\sim}{Y})$ si

$$E\{(\underset{\sim}{F} - \underset{\sim}{p}_2(\underset{\sim}{Y}))(\underset{\sim}{F} - \underset{\sim}{p}_2(\underset{\sim}{Y}))'\} - E\{(\underset{\sim}{F} - \underset{\sim}{p}_1(\underset{\sim}{Y}))(\underset{\sim}{F} - \underset{\sim}{p}_1(\underset{\sim}{Y}))'\}$$

est non négative définie. Or, le prédicteur optimal est $E(\underset{\sim}{F}|\underset{\sim}{Y})$ (Cressie, 1990), qui s'exprime pour le modèle gaussien par,

$$\underset{\sim}{p}^*(\underset{\sim}{Y}) \equiv \{\Gamma(\underset{\sim}{\gamma})(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}\}\underset{\sim}{Y} + \{I - \Gamma(\underset{\sim}{\gamma})(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}\}X\underset{\sim}{\beta},$$

et,

$$E\{(\underset{\sim}{F} - \underset{\sim}{p}^*(\underset{\sim}{Y}))(\underset{\sim}{F} - \underset{\sim}{p}^*(\underset{\sim}{Y}))'\} = \{I - \Gamma(\underset{\sim}{\gamma})(\Delta + \Gamma(\underset{\sim}{\gamma}))^{-1}\}\Gamma(\underset{\sim}{\gamma}) . \tag{2.7}$$

D'un point de vue bayésien, $\underset{\sim}{p}^*(\underset{\sim}{Y})$ est un estimateur de Bayes de $\underset{\sim}{F}$ pour la matrice de perte $L(\underset{\sim}{F}, \underset{\sim}{p}) \equiv (\underset{\sim}{F} - \underset{\sim}{p})(\underset{\sim}{F} - \underset{\sim}{p})'$.

Supposons maintenant que $\underset{\sim}{\beta}$ est inconnu, mais que $\underset{\sim}{\gamma}$ est connu. Alors le prédicteur linéaire optimal non biaisé de $\underset{\sim}{F}$ est (Cressie, 1990),

$$\underset{\sim}{\hat{p}}(\underset{\sim}{Y})=[\{\Gamma(\underset{\sim}{\gamma})(\Delta+\Gamma(\underset{\sim}{\gamma}))^{-1}\}+(I-\Gamma(\underset{\sim}{\gamma})(\Delta+\Gamma(\underset{\sim}{\gamma}))^{-1}\}$$
$$X\{X'(\Delta+\Gamma(\underset{\sim}{\gamma}))^{-1}X\}^{-1}X'(\Delta+\Gamma(\underset{\sim}{\gamma}))^{-1}]\underset{\sim}{Y}. \tag{2.8}$$

Lorsque $\underset{\sim}{\gamma}$ est inconnu (ce qui est d'ordinaire le cas), il suffit de le remplacer par $\underset{\sim}{\hat{\gamma}}$, estimateur fondé sur $\underset{\sim}{Y}$. Alors on a dit que l'équation (2.8) (avec un estimateur $\underset{\sim}{\hat{\gamma}}$) était un estimateur empirique de Bayes (Ericksen et Kadane, 1985).

Dans Cressie (1990), l'auteur utilisait deux méthodes d'estimation de $\underset{\sim}{\gamma}$: la méthode du maximum de vraisemblance et la méthode des moments. Comme, pour les modèles spatiaux, il peut être difficile d'obtenir des estimateurs par la méthode des moments, nous utiliserons la méthode gaussienne du maximum de vraisemblance pour estimer $\underset{\sim}{\gamma}$ dans la suite de cet article.

3. MODÈLES SPATIAUX DE SOUS-DÉNOMBREMENT

Dans la présente section, nous supposons que la distribution des erreurs a priori, $\underset{\sim}{F} - X\beta$, obéit à un modèle spatial. (On peut définir la dépendance spatiale en fonction d'une distance euclidienne, d'une distance mesurée en pâtés de maisons, d'une distance «sociologique ou ethnographique», etc.) Le fait d'introduire la dépendance spatiale par l'intermédiaire d'un ou deux paramètres inconnus permet de se prémunir à bon compte contre la possibilité (évoquée par Freedman et Navidi, 1986) qu'une importante variable explicative ait été omise ou que la relation fonctionnelle linéaire ne soit en fait plus complexe.

En 1934, le statisticien F.F. Stephan écrivait «... Les données relatives aux unités géographiques sont liées les unes aux autres comme les raisins d'une grappe et non pas séparées comme des billes dans une urne. Bien sûr, la simple contiguïté dans le temps et dans l'espace ne constitue pas en soi l'indication d'un manque d'indépendance entre les unités d'une variable ou d'un attribut donné. Toutefois, lorsque nous traitons des données sociales, nous savons qu'en raison même de leur caractère social, les personnes, les groupes et leur caractéristiques sont interreliés et non pas indépendants. Il est possible qu'on en vienne à élaborer des formules de calcul de l'erreur d'échantillonnage s'appliquant à ces données, mais d'ici là, il convient d'utiliser les anciennes formules avec une grande circonspection. De même, il convient d'examiner les autres mesures statistiques avec soin lorsqu'on les applique à ces données...» (Stephan, 1934).

3.1 Le modèle linéaire spatial

Selon le modèle défini par les équations (2.5) et (2.6), n'importe quel des éléments de la matrice des variances peut être assujetti à une dépendance spatiale. Premièrement, supposons que, dans l'équation (2.5),

$$\underset{\sim}{\delta} = \underset{\sim}{F} - X\underset{\sim}{\beta}, \tag{3.1}$$

forme un processus gaussien à dépendance spatiale, comme par exemple le champ aléatoire gaussien de Markov avec conditions (Besag, 1974). En pareil cas, la distribution conditionnelle de δ_i étant donné $\{\delta_j : j \neq i\}$ est une distribution gaussienne qui est fonction uniquement de $\{\delta_j : j \in N_i\}$, où N_i représente un ensemble de régions «voisines» de la région i, $i = 1, \ldots, n$. Plus précisément,

$$\delta_i|\{\delta_j;\ j \neq i\} \sim Gau(\sum_{j \in N_i} q_{ij}\delta_j,\ \tau_i^2)\ ;\ i = 1,...,n\ , \qquad (3.2)$$

ce qui implique que,

$$\underset{\sim}{F} \sim Gau(X\underset{\sim}{\beta},\ (I - Q)^{-1}M)\ , \qquad (3.3)$$

où $M \equiv diag\{\tau_1^2,...,\tau_n^2\}$ et $Q \equiv (q_{ij})$ avec $q_{ij} = 0$ chaque fois que $j \notin N_i$ (y compris lorsque $q_{ij} = 0$). À l'évidence, la matrice I-Q doit être inversible et $q_{ij}\tau_j^2 = q_{ij}\tau_j^2$ $(i<j)$ est suffisant pour que la matrice $(I-Q)^{-1}M$ soit symétrique. En outre, il faut que la matrice $(I-Q)^{-1}M$ soit définie positive. Ces conditions définissent l'espace des paramètres pour tous les modèles spatiaux gaussiens avec conditions. Ainsi, supposons que $Q = \lambda H$ et que $M = \tau^2 D$, où H et D sont connus. Alors $\Gamma(\underset{\sim}{\gamma}) = \tau^2(I - \lambda H)^{-1}D$, où $\underset{\sim}{\gamma} = (\lambda,\ \tau^2)'$; voir, par exemple, Cressie et Chan (1989). Pour les champs aléatoires de Markov, $\underset{\sim}{\gamma}$ apparaît de façon non linéaire dans $\Gamma(\underset{\sim}{\gamma})$ et peut être facilement estimé à l'aide de la méthode du maximum de vraisemblance.

Nous pouvons également utiliser à titre d'exemple un modèle tiré de la géostatistique, comme $\Gamma(\underset{\sim}{\gamma}) = \{\gamma_1 I + \gamma_2\ (\exp(-d_{ij}))\}M$ où $(\exp(-d_{ij}))$ représente une matrice dont l'élément (i,j) est $\exp(-d_{ij})$, et où d_{ij} est une «distance» (pas nécessairement euclidienne) entre la petite région i et la petite région j. En pareil cas, les deux paramètres apparaissent de façon linéaire et il est aussi simple d'estimer le paramètre $\underset{\sim}{\gamma}$ à l'aide de la méthode du maximum de vraisemblance qu'à l'aide de la méthode des moments.

D'autres modèles géostatistiques, comme $\Gamma(\underset{\sim}{\gamma}) = \{\gamma_1 I + \gamma_2(\gamma_3^{d_{ij}})\}M$, prévoient une apparition non linéaire de $\underset{\sim}{\gamma}$, lequel peut être facilement estimé à l'aide de la méthode du maximum de vraisemblance.

Il convient de noter qu'il est possible que la carte des petites régions dressée par le sociologue ou l'ethnographe diffère considérablement de la carte dressée par le géographe. Ainsi, il peut arriver que Philadelphie, New York, Détroit, Chicago et Los Angeles soient considérées comme des villes voisines aux fins de l'étude du sous-dénombrement dans de petites régions comprenant des États, des villes et des parties résiduelles d'États. De même, il peut arriver que la ville de New York et le reste de l'État de New York ne soient pas considérés comme voisins. Des démographes, des sociologues et des ethnographes pourraient aussi élaborer ensemble une structure de voisinage des petites régions en se fondant sur leur connaissance des mécanismes de sous-dénombrement (par ex., consulter Hainer et coll., 1988, pour prendre connaissance d'une étude de certains de ces mécanismes).

Le terme Δ de l'équation (2.6) peut également constituer une seconde source de dépendance spatiale. Aux fins de l'estimation de système dual du vrai $\underset{\sim}{F}$, on prélève un échantillon d'îlots de recensement à l'intérieur des petites régions. Le calcul des variances et covariances d'échantillonnage nous permet d'obtenir Δ, qui peut (ou non) être modélisé et estimé à partir de l'emplacement géographique des îlots : c'est ici qu'il serait particulièrement approprié d'avoir recours aux méthodes géostatistiques. À cette fin, le système TIGER du U.S. Census Bureau pourrait s'avérer un outil puissant pour toute modélisation spatiale.

En sus de sa généralité supplémentaire et de son élégance, le modèle linéaire spatial a aussi pour caractéristique d'être pragmatique. Il permet de tenir compte de la possibilité qu'une importante variable explicative de $E(\underset{\sim}{F}) = X\underset{\sim}{\beta}$ ait été omise ou que la relation fonctionnelle linéaire soit en fait plus complexe. Idéalement, toutes les variables importantes sont choisies ou (de façon moins idéale) des variables subrogatives apparaissent dans la droite de régression. Ces variables, les variables omises et la variable dépendante (facteur de redressement) varient toutes dans l'espace. On peut donc tirer un avantage considérable de l'ajustement d'un ou de deux paramètres de la dépendance spatiale. On peut ainsi réduire l'erreur systématique locale (à l'intérieur des

diverses petites régions) et obtenir des estimateurs plus efficaces des paramètres de régression $\underset{\sim}{\beta}$ (par ex., Doreian, 1980; Dow et coll., 1982; Anselin et Griffith, 1988; Dubin, 1988).

3.2 Estimation des paramètres de la matrice des variances

Supposons,

$$\underset{\sim}{Y} \sim Gau(X\underset{\sim}{\beta}, \Sigma(\underset{\sim}{\gamma})), \tag{3.4}$$

où on se rappelera que $\underset{\sim}{Y} \equiv (Y_1,...,Y_n)'$ est un vecteur nx1 de données, X est une matrice nxp de variables explicatives, $\underset{\sim}{\beta} = (\beta_1,...,\beta_p)'$ est un vecteur px1 de coefficients de régression inconnus, $\Sigma(\underset{\sim}{\gamma})$ est une matrice nxn symétrique définie positive et $\underset{\sim}{\gamma} \equiv (\gamma_1,...,\gamma_k)'$ est un vecteur kx1 de paramètres inconnus de la matrice des variances. Par conséquent, la fonction de vraisemblance logarithmique négative de $\underset{\sim}{\beta}$ et $\underset{\sim}{\gamma}$ est :

$$L(\underset{\sim}{\beta},\underset{\sim}{\gamma}) = (n/2)\log(2\pi) + (1/2)\log(|\Sigma(\underset{\sim}{\gamma})|) + (1/2)(\underset{\sim}{Y}-X\underset{\sim}{\beta})'\Sigma(\underset{\sim}{\gamma})^{-1}(\underset{\sim}{Y}-X\underset{\sim}{\beta}). \tag{3.5}$$

Si on minimise cette fonction par rapport à $\underset{\sim}{\beta}$ et $\underset{\sim}{\gamma}$, on obtient les estimateurs du maximum de vraisemblance (m.v.) $\underset{\sim}{\hat{\beta}}$ et $\underset{\sim}{\hat{\gamma}}$. En pratique, il faut d'ordinaire trouver les solutions par optimisation numérique, l'algorithme le plus fréquemment utilisé à cette fin étant l'algorithme itératif de Gauss-Newton (ou algorithme de caractérisation).

Définissons le vecteur px1:

$$\underset{\sim}{L}_\beta \equiv X'\Sigma(\underset{\sim}{\gamma})^{-1}X\underset{\sim}{\beta} - X'\Sigma(\underset{\sim}{\gamma})^{-1}\underset{\sim}{Y}, \tag{3.6}$$

et l'élément i du vecteur kx1 $\underset{\sim}{L}_\gamma$

$$(\underset{\sim}{L}_\gamma)_i \equiv (1/2)tr(\Sigma(\underset{\sim}{\gamma})^{-1}\Sigma_i(\underset{\sim}{\gamma})) + (1/2)\underset{\sim}{\delta}'\Sigma^i(\underset{\sim}{\gamma})\underset{\sim}{\delta}, \tag{3.7}$$

où

$$\underset{\sim}{\delta} \equiv \underset{\sim}{Y} - X\underset{\sim}{\beta}, \tag{3.8}$$

$$\Sigma_i(\underset{\sim}{\gamma}) \equiv \partial\Sigma(\underset{\sim}{\gamma})/\partial\gamma_i \tag{3.9}$$

$$\Sigma^i(\underset{\sim}{\gamma}) \equiv \partial\Sigma(\underset{\sim}{\gamma})^{-1}/\partial\gamma_i = -\Sigma(\underset{\sim}{\gamma})^{-1}\Sigma_i(\underset{\sim}{\gamma})\Sigma(\underset{\sim}{\gamma})^{-1} \; ; \; i = 1,...,k. \tag{3.10}$$

(La dérivée d'une matrice est égale à la matrice des dérivées des éléments de cette matrice et l'opérateur matriciel tr (G), appelé la trace, est la somme des éléments de la diagonale de la matrice carrée G.) Définissons la matrice pxp :

$$J_\beta \equiv X'\Sigma(\underset{\sim}{\gamma})^{-1}X \; , \tag{3.11}$$

et l'élément (i,j) de la matrice kxk J_γ :

$$(J_\gamma)_{ij} = (1/2)t_{ij} = (1/2)tr(\Sigma(\underset{\sim}{\gamma})^{-1}\Sigma_i(\underset{\sim}{\gamma})\Sigma(\underset{\sim}{\gamma})^{-1}\Sigma_j(\underset{\sim}{\gamma})). \tag{3.12}$$

Alors, l'algorithme de caractérisation est :

$$\begin{bmatrix}\underset{\sim}{\beta}^{(t+1)}\\ \underset{\sim}{\gamma}^{(t+1)}\end{bmatrix} = \begin{bmatrix}\underset{\sim}{\beta}^{(t)}\\ \underset{\sim}{\gamma}^{(t)}\end{bmatrix} - (\boldsymbol{A}^{(t)})^{-1}\begin{bmatrix}\underset{\sim}{L}_\beta^{(t)}\\ \underset{\sim}{L}_\gamma^{(t)}\end{bmatrix} \tag{3.13}$$

où

$$\boldsymbol{A} = \begin{bmatrix}J_\beta & O\\ O & J_\gamma\end{bmatrix}, \tag{3.14}$$

et $\boldsymbol{A}^{(t)}$, $\underset{\sim}{L}_\beta^{(t)}$ et $\underset{\sim}{L}_\gamma^{(t)}$ désignent respectivement les quantités $\boldsymbol{A}$, $\underset{\sim}{L}_\beta$, et $\underset{\sim}{L}_\gamma$, évaluées à $\underset{\sim}{\beta} = \underset{\sim}{\beta}^{(t)}$ et $\underset{\sim}{\gamma} = \underset{\sim}{\gamma}^{(t)}$. On peut réduire l'équation (3.13) à:

$$\underset{\sim}{\beta}^{(t)} = (X'\Sigma(\underset{\sim}{\gamma}^{(t)})^{-1}X)^{-1}X'\Sigma(\underset{\sim}{\gamma}^{(t)})^{-1}\underset{\sim}{Y} \tag{3.15}$$

$$\underset{\sim}{\gamma}^{(t+1)} = \underset{\sim}{\gamma}^{(t)} - (J_\gamma^{(t)})^{-1}\underset{\sim}{L}_\gamma^{(t)}, \tag{3.16}$$

où $J_\gamma^{(t)}$ est égal à J_γ évalué à $\underset{\sim}{\beta} = \underset{\sim}{\beta}^{(t)}$ et $\underset{\sim}{\gamma} = \underset{\sim}{\gamma}^{(t)}$. Vous trouverez plus de renseignements sur la mise en oeuvre de ces algorithmes dans Mardia et Marshall (1984) et dans Kitanidis et Lane (1985).

Lorsqu'il n'y a qu'un seul paramètre de dépendance spatiale, il est facile d'obtenir une évaluation directe de l'estimateur du maximum de vraisemblance de ce paramètre à partir de sa fonction profil de vraisemblance. Par exemple, considérons le modèle de dépendance spatiale (3.3), où

$$Q = \lambda H \text{ et } M = \tau^2 D\ , \tag{3.17}$$

H et D étant connus. Alors, dans l'équation (3.4),

$$\Sigma(\underset{\sim}{\gamma}) = \tau^2(I - \lambda H)^{-1}D + \Delta \quad ; \underset{\sim}{\gamma} = (\lambda, \tau^2)' \tag{3.18}$$

Supposons, pour l'instant, que λ est connu. En maximisant l'équation (3.5) pour $\underset{\sim}{\beta}$ et τ^2, on obtient les estimateurs du maximum de vraisemblance $\hat{\underset{\sim}{\beta}}(\lambda)$ et $\hat{\tau}^2(\lambda)$. Il suffit de substituer ces valeurs dans l'équation (3.5) pour obtenir la fonction profil de vraisemblance logarithmique négative pour le paramètre de dépendance spatiale λ:

$$L^*(\lambda) = L(\hat{\underset{\sim}{\beta}}(\lambda), \hat{\tau}^2(\lambda), \lambda), \tag{3.19}$$

où $L(\underset{\sim}{\beta}, \tau^2, \lambda)$ est donné par les équations (3.5) et (3.18).

On obtient l'estimateur du maximum de vraisemblance $\hat{\lambda}$ en minimisant L^* par rapport à λ. En outre, grâce à Whittle (1954), on obtient pour λ une région approximative de confiance à 100 (1-α) % (fondée sur des considérations asymptotiques) par:

$$\{\lambda: L^*(\lambda) \leq L^*(\hat{\lambda}) + (n/(n-p-2))x_1^2(\alpha)/2\}, \tag{3.20}$$

où $x_1^2(\alpha)$ est le point 100 (1-α) % supérieur de la distribution chi carré avec un degré de liberté.

De façon typique, lorsque $n \to \infty$, les estimateurs du maximum de vraisemblance sont cohérents, asymptotiquement gaussiens et asymptotiquement efficaces. La matrice des variances asymptotiques de $n^{1/2}(\hat{\underset{\sim}{\beta}}', \hat{\underset{\sim}{\lambda}}')'$ est A^{-1}, où A est donné par l'expression (3.14). Il convient de noter que $\hat{\underset{\sim}{\beta}}$ et $\hat{\underset{\sim}{\lambda}}$ sont asymptotiquement indépendants. En pratique, on obtient les erreurs types en calculant les racines carrées des éléments de la diagonale de A^{-1} et en les évaluant aux points $\hat{\underset{\sim}{\beta}}$ et $\hat{\underset{\sim}{\lambda}}$.

4. UNE ANALYSE SPATIALE DES DONNÉES PEP 3-8 DE L'EP DE 1980

Nous allons illustrer la construction et l'ajustement d'un modèle spatial à l'aide des données PEP 3-8 relatives aux n = 51 États (y compris le District de Columbia). Ces données sont présentées dans Cressie (1988). Les données relatives aux huit variables explicatives données par Ericksen, Kadane et Tukey (1989) ont été agrégées au niveau des 51 États (à partir des 66 petites régions comprenant des villes, des parties résiduelles d'États et des États). Ces variables sont:

1. Pourcentage de personnes faisant partie d'un groupe minoritaire
2. Taux de criminalité
3. Pourcentage de personnes pauvres
4. Pourcentage de personnes éprouvant des difficultés linguistiques
5. Scolarité
6. Logement
7. Proportion de la population habitant l'un ou l'autre de 16 centres-villes prédéterminés
8. Pourcentage de personnes habituellement recensées.

Pour trouver un sous-ensemble de ces variables fournissant un bon modèle de sous-dénombrement, nous avons utilisé la méthode de sélection d'Ericksen, Kadane et Tukey (1989), mais en attribuant aux données un poids proportionnel aux racines carrées des chiffres du recensement des petites régions. Nous avons ainsi sélectionné les variables 1 (appartenance à un groupe minoritaire) et 5 (scolarité), ainsi que le terme constant. Dans la suite du présent article, ces trois variables seront les seules variables considérées dans le modèle linéaire général, c.-à-d. que seuls les coefficients de régression β_0, β_1, et β_5 seront ajustés.

La dépendance spatiale sera modélisée à l'aide du champ aléatoire gaussien de Markov défini par les expressions (3.3) et (3.17). Dans l'expression (3.17),

$$H = \begin{cases} 1 \ ; \ d_{ij} \leq 700 \text{ milles}, \ i \neq j \\ 0 \ ; \ \text{ailleurs} \end{cases} \tag{4.1}$$

et

$$D = \text{diag}\{1/C_1,\ldots,1/C_{51}\} \ . \tag{4.2}$$

La forme de la matrice diagonale D nous est suggérée par l'équation (1.7). Nous avons obtenu la forme de H au moyen d'une analyse spatiale exploratoire des données; on trouve une justification plus étayée du choix de H dans Cressie et Chan (1989). Dans l'équation (4.1), d_{ij} est la distance entre les centres de gravité des États

i et *j*. La figure 1 présente les centres de gravité des 49 États continentaux (y compris le District de Columbia) et illustre un cercle d'un rayon de 700 milles dont le centre se trouve en Iowa.

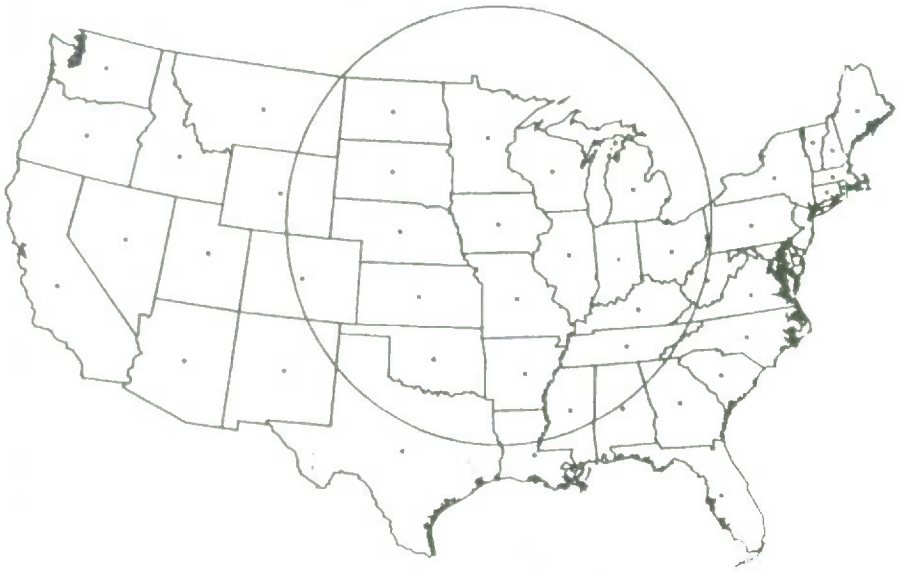


Figure 1 : **Carte des États continentaux des États-Unis illustrant leurs centres de gravité ainsi que le cercle de voisinage d'un rayon de 700 milles de l'État de l'Iowa.**

Il convient de souligner que la figure 1 n'illustre qu'une seule façon d'attribuer des h_{ij}'s non nuls. On peut choisir n'importe quel autre mode d'attribution, sans qu'il soit nécessairement fondé sur la distance : la seule condition à satisfaire est que $h_{ij}/C_j = h_{ji}/C_i$ $(i < j)$.

Afin de déterminer l'incidence de la modélisation spatiale, nous avons ajusté aux modèles (2.6), (3.3) et (3.17) les configurations suivantes :

(1,1): $\beta_5 = 0, \lambda = 0$ (1,2): $\lambda = 0$
(2,1): $\beta_5 = 0$ (2,2): aucune restriction.

Ainsi, la variable 5 est délibérément omise (ou non) et la dépendance spatiale est définie égale à zéro (ou non). Nous pourrions utiliser $\hat{\tau}^2$ pour déterminer de façon sommaire l'efficacité de chaque modèle à rendre compte des caractéristiques importantes (l'efficacité du modèle étant inversement proportionnelle à la grandeur de $\hat{\tau}^2$). Si on utilise une notation simple, les valeurs obtenues sont : $\hat{\tau}^2_{11} = 109.1$, $\hat{\tau}^2_{12} = 47.32$, $\hat{\tau}^2_{21} = 18.39$, and $\hat{\tau}^2_{22} = 0$. On notera qu'un modèle spatial avec variable omise possède un $\hat{\tau}^2$ plus petit qu'un modèle non spatial avec variable incluse, ce qui vient étayer mon assertion selon laquelle la modélisation spatiale permet de compenser l'omission de variables.

Pour la configuration spatiale (2,1), nous avons obtenu les estimateurs du maximum de vraisemblance suivants (les erreurs types estimatives figurent entre parenthèses)

$\hat{\beta}_0$: 1.00703 (0.00254) $\hat{\beta}_1$: 0.0004453 (0.0001208)
$\hat{\lambda}$: 0.04950 (0.00016) $\hat{\tau}^2$: 18.39 (20.65) .

La fonction profil de vraisemblance logarithmique négative de λ, définie par l'expression (3.19), est présentée à la figure 2.

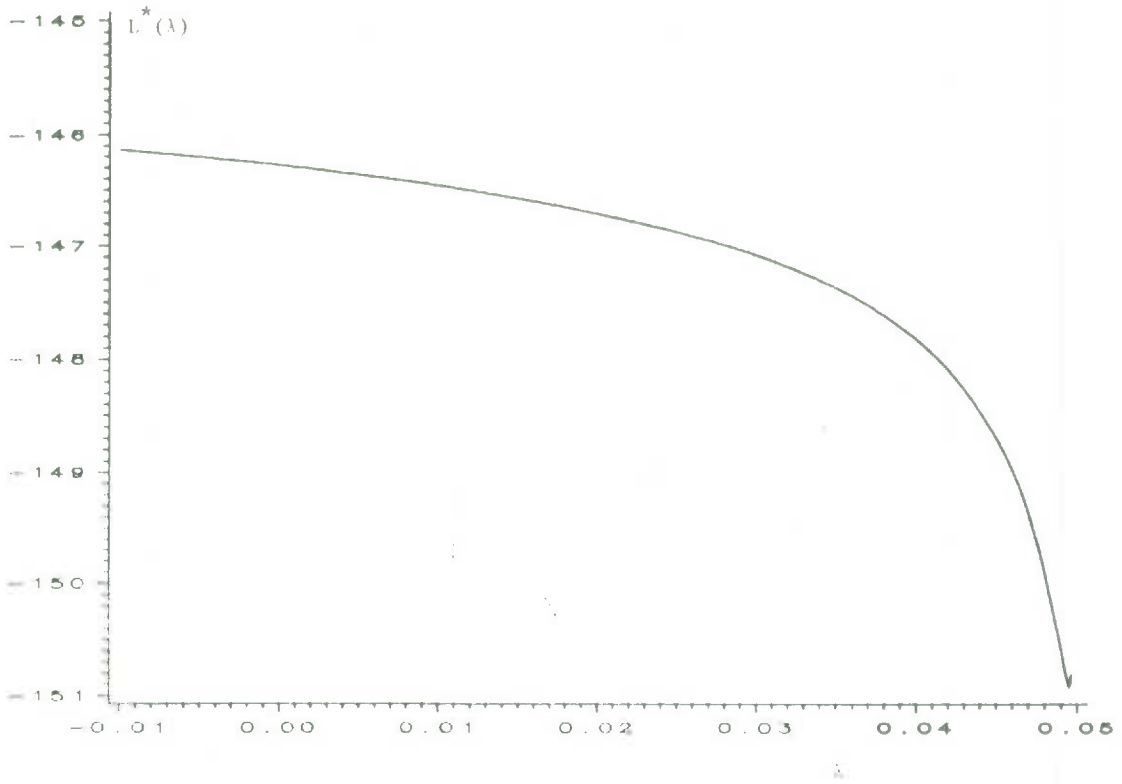


Figure 2 : **Fonction profil de vraisemblance logarithmique négative de λ définie par l'expression (3.19). L'ajustement du modèle est donné par les expressions (2.6), (3.3) et (3.17), où $\beta_s = 0$ (c.-à-d. que β_0, β_1, τ^2, et λ sont ajustés). Le domaine des valeurs de λ est déterminé par l'exigence que $(I-\lambda H)$ soit définie positive.**

L'important pour l'étude du sous-dénombrement, ce sont les prédictions empiriques de Bayes fondés sur un modèle obtenus à l'aide de l'équation (2.8) (en remplaçant $\underset{\sim}{\gamma}$ par $\underset{\sim}{\hat{\gamma}}$). En adoptant l'approche proposée dans Cressie (1990), on peut décomposer les différences $\{Y_i - 1: i = 1,...,51\}$ en différences qui correspondent à l'effet de lissage, à l'effet de la modélisation spatiale, à l'effet de l'omission d'une variable explicative et à l'effet de l'ajustement. Ainsi,

$$Y_i - 1 = (Y_i - F_{11,i}^{prd}) + (F_{11,i}^{prd} - F_{21,i}^{prd}) + (F_{21,i}^{prd} - F_{22,i}^{prd}) + (F_{22,i}^{prd} - 1). \tag{4.3}$$

Ces valeurs sont indiquées au tableau 1 avec les valeurs obtenues en multipliant chacune des différences par C_i; $i = 1,...,51$. La dernière ligne du tableau 1 indique la somme des carrés (SC) et la somme pondérée des carrés (SPC) de ces différences; ainsi, dans le cas de la différence entre les facteurs de redressement brut et l'absence de redressement, SC est $\sum_{i=1}^{51} (Y_i - 1)^2$ et SPC est $\sum_{i=1}^{51} C_i(Y_i - 1)^2$.

À l'évidence, le choix de lisser ou non les chiffres bruts redressés et de redresser ou non les chiffres du recensement importe au premier chef. Par ailleurs, l'omission d'une variable ou la présence d'une dépendance spatiale n'ont qu'une incidence relativement faible sur les redressements prévus.

Tableau 1 : **Les colonnes de gauche indiquent les composantes de la décomposition (4.3) pour chaque État. Lorsqu'on multiplie ces colonnes par les chiffres du recensement, on obtient les colonnes de droite, qui indiquent les différences par rapport à la population estimée pour chaque État. Les colonnes sont élevées au carré, (pondérées) et totalisées, pour obtenir la SC (et la SPC).**

ÉTAT	Y-1	Y-F11	F11-F21	F21-F22	F22-1	RECENSEMENT	(Y-1)C	YC-T11	T11-T21	T21-T22	T22-1
ala	-0.003528	-0.012484	-0.001182	0.005967	0.004172	3856169	-13603	-48142	-4560	23010	16088
ska	0.028754	0.009413	0.005502	-0.002486	0.016326	398316	11453	3749	2191	-990	6503
ariz	0.020377	0.009467	-0.004609	0.000464	0.015054	2699377	55005	25556	-12442	1254	40637
ark	-0.010508	-0.014640	-0.000425	0.007382	-0.002824	2258342	-23732	-33063	-960	16670	-6379
cal	0.030674	0.015399	-0.003956	-0.002078	0.021308	23417367	718296	360598	-92628	-48660	498987
col	0.003251	-0.003132	-0.005277	-0.007055	0.018714	2861207	9300	-8960	-15098	-20186	53545
con	-0.011418	-0.014126	-0.000396	-0.004594	0.007698	3065974	-35008	-43310	-1215	-14085	23603
del	-0.006211	-0.005859	-0.009368	-0.020133	0.010413	588903	-3658	-3450	5517	-11857	6132
fla	0.014384	0.002854	-0.002720	0.002561	0.011689	9654302	138867	27555	-26259	24726	112845
gga	-0.004519	-0.013227	-0.002086	0.005739	0.005055	5404451	-24425	-71486	-11272	31013	27319
hai	0.011061	0.003561	-0.003582	0.001063	0.010019	959215	10610	3416	-3436	1020	9610
idh	0.012545	0.002909	-0.000092	0.003012	0.006716	935683	11739	2722	-86	2818	6284
ill	0.021051	0.008822	0.001811	0.000351	0.010067	11291203	237687	99611	20448	3962	113665
ind	0.006387	-0.008655	0.000879	-0.001389	0.002778	5428596	-34670	-46982	4771	-7538	15080
iow	-0.006784	-0.006468	-0.001265	-0.002367	0.003317	2866173	-19443	-18539	-3626	-6785	9506
kan	0.005581	0.000785	-0.000014	-0.003934	0.008744	2326235	12983	1826	-32	-9151	20340
kty	0.015506	-0.014840	0.001105	0.009423	-0.011194	3623499	-56185	-53773	4004	34144	-40561
lou	0.023384	0.007688	-0.000584	0.007465	0.008815	4157253	97215	31963	-2426	31034	36644
mne	0.020065	0.009081	0.009147	0.002312	-0.000475	1110298	22278	10083	10155	2567	-527
mld	0.024187	0.010805	0.005306	-0.005404	0.013480	4171724	100902	45075	22135	-22545	56236
mas	0.011820	-0.010630	-0.003878	-0.003863	0.006551	5661335	-66918	-60181	-21952	-21870	37085
mch	0.007866	0.000278	0.000951	-0.001315	0.007953	9158324	72041	2543	8705	-12047	72840
min	0.011081	0.005456	-0.000024	0.000895	0.004755	4013408	44474	21899	-98	3590	19083
mis	0.009652	-0.006049	0.002150	0.004703	0.008848	2497274	24104	-15106	5369	11745	22097
mou	0.008028	0.001473	0.000874	0.004227	0.001455	4858439	39004	7156	4244	20535	7069
mon	0.014405	0.006089	0.000396	0.002283	0.005637	778046	11207	4737	308	1776	4386
neb	0.000761	-0.000737	-0.002372	-0.002945	0.006815	1546576	1176	-1140	-3668	-4555	10539
nev	0.026546	0.011352	0.000995	-0.000037	0.014236	793841	21074	9012	790	-29	11301
nwh	-0.015835	-0.009294	0.002056	-0.011849	0.003251	911430	-14433	-8471	1874	-10799	2963
nwj	0.013048	0.002963	0.001343	-0.001475	0.010217	7298030	95225	21624	9799	-10762	74564
nwm	0.023583	0.002630	-0.001859	-0.001562	0.024374	1292790	30488	3400	-2403	-2019	31510
nwy	0.016552	0.004626	-0.001020	0.001352	0.011594	17335623	286944	80188	-17677	23436	200996
noc	0.011849	0.000348	0.003226	0.007732	0.000543	5811925	68864	2020	18749	44936	3158
nod	0.000480	-0.000042	-0.001076	0.004132	-0.002534	642418	309	-27	-692	2655	-1628
oho	0.010760	0.003916	0.002032	0.000281	0.004530	10678666	114898	41815	21702	3001	48379
okl	-0.002305	-0.005148	-0.001848	0.002648	0.002044	2984855	-6879	-15367	-5515	7903	6100
ore	0.002697	0.000102	-0.005614	-0.000174	0.008389	2604411	7025	266	-14622	-453	21833
pen	-0.002806	-0.006170	-0.001637	0.003342	0.001659	11736077	-32927	-72412	-19210	39226	19469
rhi	0.008858	0.002519	0.009480	0.002039	-0.005181	934631	8279	2355	8860	1906	-4842
soc	0.063208	0.046182	0.007060	0.005336	0.004631	3088114	195193	142614	21801	16477	14301
sod	0.000831	-0.000044	-0.000512	0.002515	-0.001129	678440	564	-30	-347	1706	-766
ten	-0.028315	-0.031807	0.000013	0.005944	-0.002468	4545692	-128711	-144584	58	27021	-11206
tex	0.003708	-0.011249	-0.005440	0.005826	0.014570	14069400	52171	158260	-76532	81969	204995
uth	0.003966	0.000478	-0.004322	-0.005410	0.013221	1451408	5756	693	-6273	-7853	19199
vmt	-0.011135	-0.001267	-0.001987	-0.013766	0.001911	505424	-5628	-640	1004	-6957	966
vir	0.000911	-0.006515	0.001096	0.000362	0.005968	5291281	4819	-34470	5797	1913	31580
was	0.014188	0.008573	-0.004066	-0.001643	0.011325	4084245	57949	35014	-16607	-6712	46254
wev	-0.005796	-0.006316	0.008472	0.002882	-0.010835	1935011	-11216	-12221	16394	5578	-20966
wis	0.017349	0.008130	0.004049	0.001887	0.003284	4639675	80495	37719	18788	8753	15235
wyo	0.036087	0.014581	0.009379	0.000181	0.011946	466633	16839	6804	4376	85	5574
dcl	0.037495	0.001987	0.020905	-0.030369	0.044973	638428	23938	1252	13179	-19146	28352
SC	0.016525	0.006172	0.001343	0.002466	0.006721	SPC	70421	27838	2618	4258	28602

5. DISCUSSION

L'intérêt de la décomposition (4.3) tient au fait qu'elle démontre que le redressement au moyen de l'estimateur empirique de Bayes est relativement insensible aux modifications des hypothèses de modélisation. On notera que les composantes de l'expression (4.3) ne sont pas orthogonales, bien qu'elles soient près de l'être. Abstraction faite de cette absence d'orthogonalité, on peut conclure que la réalisation d'un redressement à l'aide des facteurs de redressement bruts se traduit par une modification grossière des chiffres du recensement, mais que l'utilisation d'un redressement au moyen de l'estimateur empirique de Bayes fondé sur un modèle offre une solution de compromis aux chercheurs selon lesquels il faut éviter de redresser les chiffres du recensement.

Ces résultats concordent avec les conclusions qu'a tirées Cressie (1990) d'une comparaison entre le modèle homoscédastique (1.8) et les modèles hétéroscédastiques (1.6) et (1.7). Toutefois, Cressie a démontré que l'erreur quadratique moyenne des estimations relatives aux petites régions est plus sensible aux modifications des hypothèses de modélisation. Il en va de même, dans la présente étude, de la prédiction spatiale des facteurs de redressement. Une comparaison des méthodes d'estimation des paramètres $\underset{\sim}{\gamma}$ a démontré qu'on observait une courbe de sensibilité similaire pour les prédicteurs et pour l'erreur quadratique moyenne des prédictions. Les estimateurs du maximum de vraisemblance peuvent comporter un important biais négatif pour les échantillons de petite taille, condition entraînant un lissage excessif des données brutes $\underset{\sim}{Y}$. Alors, l'estimateur empirique de Bayes est souvent $X\underset{\sim}{\hat{\beta}}$; cette dépendance excessive à l'égard du modèle (2.5) constitue une source compréhensible d'embarras pour le Census Bureau.

À l'évidence, il est nécessaire d'élaborer des méthodes d'estimation du paramètre $\underset{\sim}{\gamma}$ exemptes du problème de biais susmentionné. À cet égard, Cressie (1990) a étudié la possibilité d'estimer $\underset{\sim}{\gamma}$ à l'aide de la méthode des moments. Suivant le modèle simple des composantes de la variance (1.7), (2.5) et (2.6), il a découvert que l'estimateur du paramètre des composantes de la variance τ^2 obtenu à l'aide de la méthode des moments était, en général, moins biaisé vers zéro que l'estimateur m.v. Cependant, l'estimateur obtenu par la méthode des moments ne suit aucune distribution asymptotique simple. Il sera démontré ailleurs qu'on peut satisfaire à ces deux conditions (estimateur de $\underset{\sim}{\gamma}$ ayant un faible biais et une distribution asymptotique) au moyen d'une estimation par la méthode du maximum de vraisemblance avec contraintes (Patterson et Thompson, 1971).

REMERCIEMENTS

L'auteur tient à remercier Robert Parker, qui l'a aidé à effectuer les calculs et à préparer les figures. Cette recherche a été rendue possible en partie grâce aux conventions sur la statistique n^os 89-23 et 90-41 entre le U.S. Bureau of the Census et l'Iowa State University. Les opinions exprimées dans cet article sont celles de l'auteur et ne reflètent pas nécessairement la position du Census Bureau. Cette recherche a également été rendue possible en partie grâce à une subvention de la National Science Foundation (n° DMS-8902812).

BIBLIOGRAPHIE

Asselin, L., et Griffith, D.A. (1988). Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association*, 65, 11-34.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 36, 192-225.

Cressie, N. (1988). Estimating census undercount at national and subnational levels, dans *Proceedings of Bureau of the Census Fourth Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., 127-150.

Cressie, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 393-401.

Cressie, N. (1990). Weighted smoothing of estimated undercount, dans *Proceedings of Bureau of the Census 1990 Annual Research Conference*, Bureau of the Census, Washington, D.C., 301-325.

Cressie, N., et Chan, N.H. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association*, 84, 393-401.

Diffendal, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County, *Techniques d'enquête*, 14, 75-92.

Doreian, P. (1980). Linear models with spatially distributed data. Spatial disturbances or spatial effects? *Sociological Methods and Research*, 9, 29-60.

Dow, M.M., Burton, M.L., et White, D.R. (1982). Network autocorrelation: A simulation study of a foundational problem in regression and survey research. *Social Networks*, 4, 169-200.

Dubin, R.A. (1988). Estimation of regression coefficients in the presence of spatially autocorrelated error terms. *Review of Economics and Statistics*, 70, 466-474.

Ericksen, E.P., et Kadane, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-109.

Ericksen, E.P., Kadane, J.B., et Tukey, J.W. (1989). Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association*, 84, 927-944.

Freedman, D.A. et Navidi, W.C. (1986). Regression models for adjusting the 1980 Census. *Statistical Science*, 1, 3-17.

Hainer, P., Hines, C., Martin, E., et Shapiro, G. (1988). Research on improving coverage in household surveys, in *Proceeding of Bureau of the Census Fourth Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., 513-539.

Kitanidis, P.K., et Lane, R.W. (1985). Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss-Newton Method. *Journal of Hydrology*, 79, 53-71.

Mardia, K.V., et Marshall, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71, 135-146.

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

Patterson, H.D., et Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545-554.

Stephan, F. (1934). Sampling errors and interpretations of social data ordered in time and space, in *Proceedings of the American Statistical Journal*, New Series No. 185A, ed. F.A. Ross. *Journal of the American Statistical Association*, 29, Supplément, 165-166.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41, 434-449.

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

SESSION 4

Mesure de l'erreur d'enquête

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

MESURE DE LA QUALITÉ DES DONNÉES DU RECENSEMENT DE 1990

H.F. Woltman et K.F. Thomas[1]

RÉSUMÉ

Dans le cadre du programme de recherche portant sur la mesure de l'erreur non due à l'échantillonnage dont sont entachées les données recueillies au recensement de 1990, on mène des études visant à mesurer l'erreur attribuable aux répondants, aux recenseurs et aux données manquantes. On a également mis au point un modèle mathématique conçu en fonction des modes de collecte et de traitement des données du recensement de 1990 afin de pouvoir combiner les estimations de l'erreur produites à partir de ces données pour obtenir une estimation approximative de l'erreur "totale". Notre communication donne une vue d'ensemble de ce programme de recherche.

MOTS CLÉS: Réinterview; erreur non due à l'échantillonnage; erreur totale.

1. INTRODUCTION

La mesure, au moyen d'études conçues expressément à cet effet, de l'erreur dont sont entachées les données recueillies au recensement fait et continuera de faire partie intégrante de l'étape du recensement décennal consacrée à la recherche, à l'évaluation et à l'expérimentation. Nous allons donner ici une vue d'ensemble des plans du US Census Bureau relatifs à la mesure de cette erreur en ce qui concerne le recensement de 1990. Nous allons commencer par un bref exposé du processus d'élaboration du contenu du recensement, au cours duquel on détermine les sujets sur lesquels portera le recensement et les questions qui seront incluses dans le questionnaire. Nous parlerons ensuite des objectifs du programme d'évaluation de l'erreur de mesure ainsi que de la méthodologie et du plan de sondage de chacune des études menées dans le cadre de cette évaluation. Enfin, nous donnerons un aperçu du modèle d'erreur mis au point pour produire des estimations de l'erreur totale, soit à la fois l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage, que comportent les statistiques produites à partir des résultats du recensement.

2. VUE D'ENSEMBLE DU PROCESSUS D'ÉLABORATION DU CONTENU DU RECENSEMENT

2.1 Généralités

Le vingt et unième recensement décennal a eu lieu le 1er avril 1990. Le Census Bureau a dû, bien avant cette date, prendre des décisions importantes concernant les sujets qui y seraient abordés et les questions qui y seraient posées. Cet organisme était tenu par la loi de présenter au Congrès avant le 1er avril 1987 les sujets sur lesquels porterait le questionnaire du recensement de 1990, tandis qu'il avait jusqu'au 1er avril 1988 pour présenter les questions elles-mêmes.

Au cours du processus d'élaboration du contenu du recensement, on détermine les sujets et les questions qui seront inclus dans le questionnaire du recensement. Ce questionnaire est produit en deux versions: la formule abrégée, qui contient un nombre limité de questions de base relatives aux variables démographiques et au logement, et la formule complète, qui contient, en plus, quelques questions supplémentaires. Il existe des questionnaires spéciaux comme le rapport individuel, le rapport destiné aux militaires et celui s'adressant aux

[1] H.F. Woltman et K.F. Thomas, Statistical Support Division, US Bureau of the Census, Washington, DC 20233.

marins, dont on se sert pour le dénombrement de certaines parties de la population. On y trouve les mêmes questions de nature démographique que dans le questionnaire abrégé et dans le questionnaire complet.

2.2 Principales étapes de la planification du contenu

On a commencé par passer en revue les questions posées au recensement de 1980. On a ensuite déterminé les besoins en données prévus pour 1990 et examiné les lois en vigueur et en voie d'être promulguées qui pourraient avoir une incidence sur les données de recensement nécessaires. Le Census Bureau a alors organisé une structure complexe de rencontres, de conférences et de groupes de travail pour faire en sorte que les diverses parties de la communauté des utilisateurs de données soient consultées pendant le processus d'élaboration du contenu du recensement. Il en a résulté une liste des questions le plus susceptibles d'être retenues pour la série d'essais qui allait avoir lieu. Dans la plupart des cas, au lieu de faire des recommandations précises quant au contenu du recensement, les personnes consultées ont indiqué de façon générale les données dont elles auraient besoin, et les questions ont été formulées en conséquence par le Census Bureau.

Diverses rencontres ont eu lieu, dont le but était de produire des recommandations; les premières ont été les rencontres publiques locales au printemps de 1984, suivies par celles des comités consultatifs et des organismes fédéraux, qui se sont réunis périodiquement tout au long du cycle du recensement pour discuter des activités du recensement de 1990.

Les analystes du Census Bureau ont examiné les recommandations concernant le contenu du recensement et y ont appliqué les critères dont nous parlerons plus loin afin de déterminer les questions qu'il conviendrait de considérer et qui devraient être incluses dans les essais. Les recommandations reçues de l'ensemble des sources étaient bien trop nombreuses pour qu'on puisse tenir compte de toutes dans le questionnaire, et beaucoup d'entre elles portaient sur des renseignements qui n'étaient pas exigés par la législation fédérale. Les questions les plus prometteuses ont été mises à l'épreuve au cours du test du recensement national (National Content Test, NCT), de recensements d'essai et d'essais spéciaux. On a ensuite procédé à des évaluations exhaustives pour s'assurer que les données recueillies au moyen des questions étaient exactes et fiables. La liste des sujets qu'on prévoyait inclure dans le recensement de 1990 a été envoyée au Congrès le 27 mars 1987.

Le contenu du recensement a été perfectionné après la tenue d'essais supplémentaires qui ont eu lieu en été et en automne 1987. L'aboutissement de ce processus a été la présentation au Congrès, le 1er avril 1988, des questions proposées pour le recensement de 1990 et des catégories de réponse correspondantes.

2.3 Critères utilisés pour déterminer le contenu du recensement de 1990

Pour choisir les questions, le Census Bureau s'est fondé sur cinq critères de base. Premièrement, on n'a considéré que les données essentielles, soit celles qui répondaient à un vaste besoin dans la société et celles qui sont exigées en vertu de lois à l'échelle du pays, des États et des municipalités ou qui sont nécessaires pour l'application de programmes gouvernementaux.

Il fallait que ces données soient requises pour des régions relativement petites (administrations municipales et petites régions statistiques) ou pour des populations à faible effectif. En effet, les enquêtes par sondage sont un moyen plus approprié à la collecte de données requises seulement à l'échelle nationale ou régionale.

Deuxièmement, on a repris en 1990 un bon nombre des questions posées en 1980 parce qu'elles permettent de recueillir des informations fondamentales pour l'ensemble du pays et de constituer des séries chronologiques de données essentielles sur le plan socio-économique et en ce qui concerne le logement. La stabilité relative du contenu du recensement au cours des quelques dernières décennies tient, en partie, à la pertinence et à l'utilité de certaines questions de base et à la nécessité de mesurer l'évolution dans le temps de la population et du parc de logements.

Troisièmement, le nombre de questions posées au recensement de 1990 n'a pas été augmenté de façon appréciable par rapport au recensement de 1980. Le Census Bureau a dû trouver le bon équilibre entre la quantité d'informations nécessaires et la longueur du questionnaire. Ce facteur est important parce que la

collaboration du public, dont dépend la réussite du recensement, pourrait être moins grande si les recensés trouvaient le fardeau de réponse trop lourd.

Quatrièmement, les formules ne contenaient pas de questions indiscrètes, choquantes ou très controversées. Il est illégal, par exemple, d'obliger les recensés à répondre à des questions concernant leur appartenance ou leurs croyances religieuses. On a évité d'autres sujets susceptibles d'influencer ou de réduire le taux de réponse au recensement.

Cinquièmement, il fallait que le Census Bureau puisse formuler, sur chacun des sujets, des questions qui soient claires et concises et qui permettent d'obtenir des données exactes. Le libellé et la présentation étaient particulièrement importants, car, dans la plupart des cas, les questionnaires du recensement étaient envoyés par la poste et remplis par les recensés eux-mêmes. S'il ressortait des essais qu'une question risquait d'être mal interprétée, ne serait-ce que par une partie de la population, il y avait peu de chances qu'elle soit incluse dans le questionnaire.

On a fait divers tests pour évaluer l'effet de changements dans le libellé, l'ordre et la présentation des questions. La plupart ont été effectués dans le cadre du NCT. On a cependant eu recours à d'autres moyens pour la conception et l'élaboration du questionnaire du recensement de 1990, notamment à des recensements d'essai et à des essais spéciaux (département du Commerce des É.-U, 1987).

2.4 Sommaire

Le recensement de 1990 représentait le deux centième recensement effectué aux États-Unis. Il a fallu faire de nombreux choix difficiles au sujet de son contenu pour en terminer la préparation. La tâche importante que constitue la sélection des sujets sur lesquels porterait le recensement a été accomplie après qu'on ait eu soigneusement examiné et testé les recommandations émanant d'un large éventail d'utilisateurs des données de recensement, parmi lesquels des organismes fédéraux et autres organismes publics, des comités consultatifs, des associations professionnelles et des membres du grand public.

3. OBJECTIFS DU PROGRAMME DE RECHERCHE, D'ÉVALUATION ET D'EXPÉRIMENTATION DE 1990

Le Census Bureau a groupé la plupart des études composant le programme de recherche, d'évaluation et d'expérimentation de 1990 (1990 Research, Evaluation and Experimentation Program, REX) en trois principaux volets: contenu, couverture, procédures et traitement.

Les études dans le volet contenu du programme REX sont celles dont le but était de faire en sorte qu'on obtienne des renseignements sur les questions suivantes:

1. Nous évaluerons la qualité des données tirées du recensement de 1990 afin de pouvoir fournir des renseignements sur l'origine et l'ampleur des erreurs non dues à l'échantillonnage commises au cours de la collecte et du traitement des données.
2. Nous évaluerons l'effet, sur l'utilisation des données, de l'erreur d'échantillonnage, de l'erreur non due à l'échantillonnage, de l'erreur de couverture et de l'erreur géographique. Nous déterminerons les principales utilisations des données et évaluerons l'incidence qu'ont sur les applications des données de recensement les erreurs dont celles-ci sont entachées.
3. En nous fondant sur les données recueillies au recensement et sur les évaluations, nous déterminerons des façons plus efficaces et plus précises de recueillir des données de recensement.
4. Nous adopterons une façon de procéder qui permettra de contrôler, de décrire et, s'il y a lieu, de corriger les erreurs qui se produisent pendant la collecte et le traitement des données.

Les objectifs premiers du programme REX de 1990, comme ceux de tout programme de recherche et d'évaluation, sont:

1. de produire des données que le Census Bureau pourra utiliser pour évaluer et améliorer les méthodes à utiliser et les opérations à exécuter au cours des recensements à venir;
2. de fournir des renseignements aux utilisateurs des données de recensement concernant l'origine et l'effet des erreurs dont ces données sont entachées.

Le Census Bureau s'est en outre fondé sur les critères suivants pour choisir les composantes du programme REX de 1990:

1. La proposition devrait s'insérer dans la limite des ressources humaines et financières dont on disposerait pour l'ensemble du programme REX.
2. La proposition devrait nécessiter un contexte de recensement réel, sans lequel les mesures ne serait pas adéquates. (Autrement, il conviendrait mieux d'y donner suite dans le cadre d'un recensement d'essai ou d'un autre essai spécial).
3. La proposition ne devrait pas avoir pour effet de retarder au delà des dates limites prévues par la loi la production des chiffres servant à la redistribution de la population entre les circonscriptions électorales et à la redéfinition de ces circonscriptions.
4. La proposition ne devrait pas influer sur la qualité des données de recensement de quelque manière risquant de compromettre les principales utilisations de ces données.

Compte tenu de ces critères, le programme d'évaluation du contenu du recensement comporte quatre grands projets d'évaluation, qui comprennent l'enquête de réinterview (Content Reinterview Survey, CRS), l'étude de la variance due au recenseur (Enumerator Variance Study, EVS), l'évaluation de l'imputation et la recherche s'y rapportant (Imputation Evaluation and Research) et le suivi du traitement des enregistrements (Master Trace Study, MTS). De plus, on a mis au point un modèle d'erreur de mesure qui tente d'expliquer les principales erreurs de procédures opérationnelles et les sources d'erreurs du recensement.

4. LES ÉTUDES

Le programme d'évaluation du contenu du recensement comporte quatre grandes études visant à mesurer l'erreur entachant les données recueillies qui serait attribuable aux recensés, aux recenseurs et à l'imputation des données manquantes. La méthodologie, le plan de sondage et la taille de l'échantillon de chacune de ces études sont décrits ci-après.

4.1 Enquête de réinterview

L'enquête de réinterview est toujours menée après le recensement décennal et elle a pour but de mesurer l'erreur de réponse associée à certains postes ayant trait aux variables démographiques et aux variables concernant le logement. Son objectif est de mesurer la variance de réponse simple et le biais de réponse associés aux données recueillies au moment de la réinterview. Celle de 1990 aura aussi pour objet d'évaluer la qualité des données, mais elle portera principalement sur les données recueillies au moyen de nouvelles questions et comportera des innovations dans le domaine des techniques de collecte des données sur le terrain.

Dans cette étude, on comparera les réponses individuelles recueillies au recensement à celles recueillies pendant la réinterview. On produira des mesures de l'erreur de réponse à partir de l'échantillon pour estimer l'erreur systématique et l'erreur aléatoire non due à l'échantillonnage associées aux données recueillies au moyen du questionnaire de recensement complet.

L'échantillon a été limité aux ménages qui ont rempli le questionnaire complet, et on réinterviewera aussi bien les ménages qui ont renvoyé leur questionnaire par la poste que les autres.

On a utilisé un plan d'échantillonnage systématique à un degré. Pour obtenir un échantillon final de 12,800 logements occupés, on a prélevé dans le fichier d'adresses du recensement un échantillon initial de 15,500

logements. Ce nombre de ménages est comparable à celui fixé pour les échantillons de l'enquête de réinterview de 1970 et de 1980.

Pour toutes les interviews initiales, on procède par interview téléphonique assistée par ordinateur (ITAO). En effet, comme on se sert des résultats de cette étude pour mesurer la validité des réponses fournies au recensement, il est essentiel que ceux-ci soient de la plus grande qualité. Or, l'ITAO permet d'obtenir des données de meilleure qualité, car le biais dû à l'intervieweur y est réduit en raison, premièrement, du rôle joué par l'ordinateur (lequel fait en sorte que les instructions "passez à" sont respectées) et, deuxièmement, du contrôle exercé par les superviseurs (qui peuvent surveiller aussi bien le déroulement de l'interview au téléphone que l'inscription des données à l'écran).

Les taux de réponse à l'enquête de réinterview ont toujours été élevés parce qu'on y procède à un suivi très complet. En 1990, on enverra des lettres de présentation aux logements échantillonnés avant le début des interviews téléphoniques. On fera des ITAO auprès des ménages jusqu'à ce qu'on ait obtenu tous les renseignements nécessaires concernant chaque personne et le logement (ou jusqu'à ce qu'un certain nombre de tentatives d'interview se soldent par un échec parce qu'on n'aura pas réussi à rejoindre le ménage). Les données sur les caractéristiques des personnes âgées de 15 ans et plus seront fournies par les personnes elles-mêmes. Les ménages qu'on n'aura pas réussi à rejoindre au téléphone feront l'objet d'un suivi sur place.

Pour l'échantillon de l'enquête, il fallait pouvoir prendre dans le questionnaire le numéro de liste des ménages et leur numéro de téléphone. L'ITAO l'exige, et cela présente en outre des avantages lorsqu'on veut faire le couplage des données de recensement et des données de réinterview se rapportant aux mêmes personnes.

On fera des recoupements entre les réponses obtenues au recensement et pendant la réinterview au sujet des mêmes personnes et des mêmes logements pour calculer des mesures de la variance de réponse et du biais de réponse.

Pour certains postes, le questionnaire de l'enquête de réinterview contient des questions d'approfondissement qui permettent de recueillir des données d'un degré d'exactitude impossible à atteindre au recensement. Autrement dit, on peut considérer l'enquête de réinterview comme la technique de mesure «privilégiée». En comparant les données recueillies au moment de la réinterview à celles provenant du recensement, on obtiendra une estimation du biais de réponse associé aux données de recensement. Parmi les variables pour lesquelles on évaluera le biais de réponse en posant une question d'approfondissement à la réinterview, mentionnons :

Variables démographiques	Variables relatives au logement
Race	Mode d'occupation
Lieu de naissance	Loyer mensuel
Citoyenneté	Repas avec le loyer
Niveau d'instruction	Installations sanitaires
Origines	Nombre de voitures et de camionnettes
Langues parlées	Année de construction
Service militaire	
Situation d'emploi	
Incapacité	

Pour certaines variables, on estime la variance de réponse en posant la ou les mêmes questions qu'au recensement. Il s'agit notamment des variables suivantes :

Origine espagnole	Description de l'immeuble
Année d'immigration	Taille du terrain
Fréquentation scolaire	Ventes agricoles
Employeur: Genre de commerce et type d'entreprise	

4.2 Étude de la variance due au recenseur

Les recenseurs chargés du suivi de la non-réponse sont une cause importante d'erreur non due à l'échantillonnage dans les statistiques produites à partir des résultats du recensement décennal de 1990. L'étude de la variance due au recenseur a pour objet d'estimer la part de l'erreur totale non due à l'échantillonnage attribuable à ces recenseurs.

Pour l'étude de la variance due au recenseur, on s'est servi d'un plan de sondage à trois degrés pour recueillir des données auprès des recenseurs chargés du suivi de la non-réponse dans les régions métropolitaines.

1. On a formé quinze strates de bureaux de district (district offices, DO) et on a prélevé deux bureaux de district dans chaque strate.
2. On a prélevé dix-sept secteurs géographiques de registres d'adresses (address register areas, ARA) -- qui correspondent généralement à des secteurs de recensement -- dans les bureaux de district prélevés au premier degré.
3. On a réparti au hasard les îlots de recensement prélevés dans les secteurs de registres d'adresses de l'échantillon entre les tâches des recenseurs pour favoriser un mélange des tâches.

Cette étude a certaines limites:

- On s'attend qu'un seul recenseur fasse la majeure partie de chaque tâche de suivi de la non-réponse. La variance de l'erreur associée à la tâche de suivi de la non-réponse devrait être plus faible que la variance de l'erreur due au recenseur dans le cas des tâches accomplies par plus d'un recenseur.
- Pour réduire au maximum les difficultés et les coûts opérationnels, on a décidé de mélanger les îlots de recensement plutôt que les ménages. On ne pourra pas estimer la variance associée à la tâche du recenseur en ce qui concerne les unités géographiques inférieures à deux îlots. La variance associée à la tâche à l'intérieur des îlots sera confondue avec la variance associée à la tâche entre les îlots.
- Les résultats seront limités principalement aux régions urbaines du pays.

4.3 Recherche relative à l'imputation

L'imputation est l'attribution de données là ou des questions sont restées sans réponse dans le questionnaire. L'effet de l'imputation sur la qualité des données et l'ampleur de l'erreur qu'elle introduit dans les statistiques de recensement publiées sont des choses que l'on ne comprend pas bien à l'heure actuelle.

L'évaluation des méthodes d'imputation utilisées au recensement nous renseignera sur la quantité de données de recensement qui ont été imputées et sur les façons dont l'imputation a influé sur les données. En faisant de la recherche dans le domaine, nous serons mieux placés pour choisir entre les méthodes d'imputation qui s'offrent à nous. À terme, le fait d'essayer diverses méthodes et de les adapter donnera lieu à une meilleure manière de procéder en matière d'imputation.

4.3.1 Évaluation de la méthode du «hot deck» actuellement utilisée

La méthode du «hot deck», servant à imputer des valeurs là où elles sont manquantes consiste à trouver dans le fichier la dernière personne ou le dernier ménage présentant des similarités avec la personne ou le ménage en cause. L'ordre des ménages dans le fichier joue donc un rôle important, et il correspond plus ou moins à l'agencement des unités géographiques dans l'univers physique. On a eu recours à cette méthode séquentielle pour deux raisons principales : 1) elle est pratique du point de vue calcul parce qu'elle ne nécessite qu'on parcoure le fichier de données du recensement qu'une fois; 2) selon les résultats des analyses de données, il existe pour beaucoup de variables de recensement de fortes corrélations entre les personnes qui sont voisines, et ces corrélations s'atténuent à mesure que la distance entre les ménages augmente; un ménage voisin est donc généralement un meilleur prédicteur qu'un ménage qui est éloigné dans le fichier de données (Thomas et coll., 1984).

L'évaluation de la méthode du «hot deck» se fera en plusieurs étapes (Schafer, 1989, Schafer, 1990).

a. Évaluation descriptive

Il s'agit de décrire l'ensemble des règles selon lesquelles cette méthode permet d'imputer des données là où il n'y en a pas. Ces règles se divisent en deux grandes catégories: les règles de contrôle de la cohérence et les règles d'imputation. Il importera de faire la distinction entre ces deux catégories de règles, tant au moment d'évaluer la méthode du «hot deck» telle qu'elle est utilisée actuellement que de proposer des améliorations et des solutions de rechange.

b. Estimation des taux d'erreur d'imputation

Le modèle d'erreur totale actuellement utilisé pour le recensement (Woltman, Johnson, 1989) tente d'intégrer en un seul modèle les diverses sources d'erreur et permet de calculer approximativement l'erreur quadratique moyenne associée à une statistique de recensement. La statistique, p, est une estimation, selon les résultats du recensement, de la proportion P de particuliers ou de ménages dans la population possédant certaines caractéristiques. Autrement dit, p est la proportion de «positifs» observés au recensement tandis que P est la proportion réelle de positifs dans la population.

Le modèle d'erreur totale précise deux paramètres de l'erreur due à l'imputation de données pour des caractéristiques manquantes: θ_1, la probabilité d'imputer une valeur négative lorsque la valeur réelle est positive et ϕ_1, la probabilité d'imputer une valeur positive lorsque la valeur réelle est négative. On peut se servir des données provenant de l'enquête de réinterview et, peut-être, de celles tirées de l'enquête postcensitaire pour produire des estimations ponctuelles des données intégrales et des données d'échantillon. Schafer (1989) propose également de redéfinir les paramètres des taux d'erreur de manière à séparer l'effet du contrôle de cohérence de l'effet de l'imputation.

c. Analyse de données

Les analyses proposées par Schafer (1990) sont semblables, du point de vue philosophique, à l'étude de Thomas et collab. (1984), en ce sens qu'elles visent à déterminer s'il y a des relations spatiales entre les données de recensement observées, dans l'espoir que ces relations seront également valables dans le cas des données manquantes (et imputées par la suite). Si l'on trouve un rapport spatial qui permet de prédire correctement la valeur des données observées, alors il est raisonnable de penser qu'on pourrait peut-être aussi s'en servir comme outil d'imputation.

L'idée de base consiste à appliquer la règle d'imputation de la méthode du «hot deck» aux données de recensement et à voir dans quelle mesure cette règle prédit les données effectivement observées. Supposons que Y est une variable clé de recensement (p. ex. origine hispanique) dont nous allons imputer les valeurs. Nous pouvons appliquer la règle d'imputation à chaque unité (ménage) dans le fichier de données pour trouver l'enregistrement donneur correspondant; cet enregistrement pourrait être défini, par exemple, comme étant celui du dernier ménage partageant une autre caractéristique, telle que le nombre de personnes qui le constituent. En comparant la valeur de Y dans l'enregistrement donneur à la valeur de Y dans l'enregistrement receveur, nous pouvons évaluer la performance de cette règle d'imputation relativement à notre ensemble de données. Nous produirons des statistiques sommaires et leur représentation graphique, qui permettront de juger et de comparer la performance des diverses règles d'imputation.

4.3.2 Élaboration de méthodes d'imputation améliorées

L'algorithme actuel de la méthode du «hot deck» est fondé sur la grande expérience que nous avons acquise à propos de la structure des données de recensement, tant au niveau des particuliers que des ménages, et à propos des mécanismes qui font que des données sont manquantes. Schafer (1989) présente une reformulation de la méthode du «hot deck» dans un modèle statistique formel. Nous espérons que ce modèle nous permettra de mettre au point une méthode d'imputation plus raisonnée, moins empirique, que la méthode du «hot deck», sans rien sacrifier de l'efficacité des règles d'imputation de celle-ci.

Schafer (1989) examine d'autres méthodes qu'on peut employer pour évaluer les solutions de rechange en matière de règles d'appariement séquentiel selon le principe du «hot deck» afin de trouver le compromis optimal entre l'appariement selon des covariables et la proximité géographiques. Pour finir, nous nous pencherons sur les possibilités autres que la méthode d'imputation séquentielle du «hot deck», qui est fortement tributaire de l'information géographique implicite dans le classement des unités dans le fichier de données. Ainsi, nous évaluerons davantage de méthodes fondées sur les grappes: pour définir le bassin de donneurs dans des îlots ou des groupes d'îlots, par exemple (Schafer, 1989 et Blodgett, 1990).

4.4 Suivi du traitement des enregistrements

Le suivi du traitement des enregistrements a pour objet la constitution et la documentation d'une base de données exhaustive sur laquelle pourront s'appuyer les travaux en cours en matière de recherche, d'évaluation et d'expérimentation et la production de données qui pourront servir aux évaluations futures ayant trait au recensement. Ces données devraient pouvoir fournir des données-échantillon de nature géographique et démographique à différentes étapes de la collecte et du traitement des données de recensement. Ces «clichés» d'un échantillon de questionnaires du recensement (formules complètes et abrégées) à diverses étapes du traitement et les résultats de la réinterview postcensitaire permettront d'évaluer la qualité des réponses individuelles aux questions et l'effet des méthodes de collecte et de traitement ainsi que du processus de contrôle et d'imputation sur les réponses aux questions prises individuellement.

L'objectif du suivi du traitement des enregistrements est la production d'une base de données contenant une entrée pour tous les postes du questionnaire, à chaque étape du traitement, pour chaque enregistrement correspondant à une unité de l'échantillon du recensement. Les résultats permettront d'évaluer la qualité et l'exactitude des changements apportés, pendant le traitement des données de recensement et jusqu'à la production des dernières totalisations, aux réponses inscrites dans les questionnaires par les répondants ou les recenseurs. Ils permettront aussi de trouver la source des erreurs introduites dans les données de recensement pendant le traitement des données, et donc de mettre au point des procédures et des processus améliorés qui feront en sorte que l'erreur soit moins grande à l'avenir. Ce suivi produira enfin des données dont on pourra se servir pour évaluer l'ampleur de l'erreur dont sont entachées les données de recensement et l'effet de cette erreur sur l'utilisation qui est faite de ces données.

Pour le suivi du traitement des enregistrements, nous avons prélevé dans l'ensemble du pays un échantillon aléatoire de 15,500 questionnaires complets et de 15,500 questionnaires abrégés. L'échantillon de questionnaires complets est le même que celui que nous avons utilisé pour l'enquête de réinterview de 1990. On estime qu'environ 60 à 70 formules par district feront l'objet du suivi.

4.5 Modèle de l'erreur totale

Un modèle d'erreur de mesure a été mis au point pour la production d'estimations de l'erreur totale (erreur d'échantillonnage et erreur non due à l'échantillonnage) dont sont entachées les statistiques de recensement (Woltman et Johnson, 1989). L'expression «erreur totale» implique l'intégration de toutes les sources d'erreur que l'on retrouve dans les statistiques produites à partir des résultats d'un recensement ou d'une enquête. Or, ce modèle-ci à lui seul ne tient pas compte de toutes les sources d'erreur. L'expression «modèle de l'erreur totale» utilisée pour le désigner est donc employée un peu librement.

Les applications d'un tel modèle sont les suivantes:

1. L'estimation de l'erreur totale dont est entachée une statistique de recensement comprenant l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage. Il est alors possible de fonder les inférences statistiques sur l'exactitude globale des statistiques plutôt que sur la précision de l'échantillonnage seulement.
2. Le modèle permet de mesurer l'incidence relative de différentes sortes d'erreur sur l'erreur totale. Il est alors possible de déterminer en connaissance de cause les processus ou les opérations qui ont besoin d'être améliorés.
3. Il permet d'appliquer ces connaissances à la répartition efficace des ressources pour les recensements à venir, c'est-à-dire de manière à réduire au maximum l'erreur totale compte tenu d'un coût fixe.

Le modèle cherche à rendre compte des sources d'erreur suivantes:

- Non-exhaustivité du recensement
- Erreur d'échantillonnage (dans le cas des statistiques produites à partir des données-échantillon)
- Erreur de réponse
- Erreur due au recenseur
- Erreur d'imputation

Le modèle cherche en outre à rendre compte du lien entre ces sources d'erreur et les processus et opérations de recensement suivants:

- Le fait qu'un questionnaire soit retourné par la poste ou non
- La vérification des données manquantes dans les questionnaires retournés par la poste
- Le suivi des questionnaires rejetés au contrôle en raison des règles de tolérance appliquées
- Le suivi des questionnaires non retournés par la poste
- La méthode d'imputation appliquée lorsqu'il manque encore des données après les suivis.

En termes précis, le modèle de l'erreur totale cherche donc à rendre compte des processus et des opérations qui donnent lieu à des erreurs d'échantillonnage ou à des erreurs non dues à l'échantillonnage dans la proportion de ménages ou de particuliers qui possèdent un certain attribut selon les valeurs observées et les valeurs imputées relativement à un poste donné du questionnaire.

Nous nous servirons des estimations des divers paramètres d'erreur qui découleront de chaque étude conjointement avec le modèle pour simuler des estimations de l'erreur totale associée à diverses statistiques produites, à divers niveaux géographiques, d'après les résultats du recensement.

BIBLIOGRAPHIE

Blodgett, R. (1990). *Changing Point*, note de service interne du Census Bureau.

Schafer, J.L. (1989). *Proposal for Imputation Evaluation*, note de service interne du Census Bureau.

Schafer, J.L. (1990). *Comments on Autocorrelation and Imputation*, note de service interne du Census Bureau.

US Department of Commerce, Bureau of the Census (1987). *The Content Development Process for the 1990 Census of Population and Housing.*

Thomas, K.F., Harner, D.A., et Fay, R. (1984). *Preliminary Evaluation Results Memorandum No.69*, Intraclass Correlations Using a Sample of 1980 Census Data, note de service interne du Census Bureau.

Woltman, H.F., et Johnson, R. (1989). A Total Error Model for the 1990 Census, *Proceedings of the Survey Research Section*, American Statistical Association.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

COMPARAISON DE TROIS MÉTHODES BOOTSTRAP POUR DES DONNÉES D'ENQUÊTE

R.R. Sitter[1]

RÉSUMÉ

Les ouvrages statistiques proposent diverses méthodes "bootstrap" pour l'estimation de la variance et la construction d'intervalles de confiance dans les enquêtes à plan de sondage complexe où l'échantillonnage se fait sans remise. La plus ancienne de ces méthodes, et peut-être la plus intéressante à première vue, est la méthode BSR (bootstrap sans remise), proposée par Gross (1980). Malheureusement, elle ne peut s'appliquer qu'à des plans de sondage très simples. Dans un premier temps, nous allons étendre l'usage de la méthode BSR à des plans plus complexes, puis, par une étude de simulation, nous allons comparer son efficacité avec celle de deux autres méthodes, à savoir la méthode bootstrap avec transformation (Rao et Wu, 1988) et la méthode bootstrap "mirror-match" (Sitter, 1990). Ces trois méthodes résument à elles seules les diverses possibilités de bootstrap.

MOTS CLÉS: Bootstrap; jackknife; échantillonnage à deux degrés; développement d'Edgeworth.

1. INTRODUCTION

Pour la plupart des plans de sondage complexes, il est possible de calculer des estimateurs non biaisés de la variance pour des paramètres statistiques qui peuvent être exprimés comme une fonction linéaire des observations. Ce n'est généralement pas le cas pour les statistiques non linéaires et les fonctions non linéaires de statistiques. Les ouvrages de statistique proposent diverses méthodes d'estimation de la variance. Les trois plus connues sont la méthode de linéarisation (ou méthode de Taylor), la méthode du "jackknife" et la méthode BRR (balanced repeated replications). Ces méthodes présentent toutefois des lacunes. La méthode de linéarisation, par exemple, nécessite le calcul théorique et la programmation de dérivées, ce qui en rend l'application ardue. Lorsque les unités primaires d'échantillonnage sont tirées sans remise, la méthode du jackknife s'applique difficilement, ayant été développée uniquement pour l'échantillonnage stratifié. Enfin, la méthode BRR est limitée à l'échantillonnage stratifié et s'accompagne de restrictions concernant le nombre d'unités par strate. Compte tenu de ces lacunes, on a proposé diverses méthodes "bootstrap" pour l'estimation de la variance et la construction d'intervalles de confiance dans les enquêtes par sondage. Cette communication vise à comparer le rendement de trois méthodes bootstrap -- la méthode bootstrap avec transformation (Rao et Wu, 1988), la méthode bootstrap "mirror-match" (Sitter, 1990) et la méthode bootstrap sans remise (BSR) (Gross, 1980; Bickel et Freedman, 1984) -- du point de vue de l'estimation de la variance et de la construction d'intervalles de confiance dans des enquêtes à plan de sondage complexe où l'échantillon est tiré sans remise. Ces trois méthodes résument les diverses solutions bootstrap.

Dans la section 2, nous exposons les trois méthodes et formulons les algorithmes pour l'échantillonnage stratifié sans remise. Bien qu'elle soit intéressante à première vue et qu'elle semble la plus naturelle des trois, la méthode BSR ne peut s'appliquer qu'à l'échantillonnage aléatoire simple sans remise. Pour qu'elle soit

---

[1] R.R. Sitter, Assistant professor, Department of Mathematics and Statistics, Carleton University, Ottawa (Ontario), K1S 5B6. Cette communication représente une partie de la thèse de doctorat qu'a rédigée l'auteur à l'Université de Waterloo, Waterloo (Ontario), sous la direction de C.F.J. Wu.

considérée sur le même pied que les deux autres, nous proposons dans la section 3 de l'étendre à l'échantillonnage stratifié, à l'échantillonnage à deux degrés et à l'échantillonnage de Rao-Hartley-Cochran (1962). Dans la section 4, nous reproduisons les résultats d'une étude de simulation faite à l'aide de diverses populations finies. Cette étude vise à comparer l'efficacité des trois méthodes bootstrap en ce qui a trait à l'estimation de la variance et à la construction d'intervalles de confiance pour les paramètres statistiques non linéaires r (quotient), b (coefficient de régression) et c (coefficient de corrélation), de même que pour la médiane, m. Nous analysons le cas de l'échantillonnage aléatoire stratifié sans remise en faisant varier la taille des strates, le nombre de strates et la fraction de sondage à l'intérieur des strates. La méthode de linéarisation et la méthode jackknife sont incluses dans l'étude de r, b et c; de plus, l'étude de la médiane tient compte d'une méthode fondée sur les intervalles de confiance de Woodruff (1952) pour les quantiles.

2. LES TROIS MÉTHODES BOOTSTRAP

Afin d'exposer la notation nécessaire pour cette communication, nous allons formuler ici l'échantillonnage stratifié sans remise. Dans l'échantillonnage aléatoire stratifié, la population finie de N unités est répartie en L strates distinctes de $N_1, N_2, ..., N_L$ unités respectivement; par conséquent, $N_1 + N_2 + ... + N_L = N$. Un échantillon aléatoire simple est prélevé (sans remise) dans chaque strate de façon indépendante. La taille de ces échantillons est désignée par n_1, n_2, ..., n_L et la taille de l'échantillon global est $n = n_1 + n_2 + ... + n_L$. La mesure (ou plus probablement le vecteur de mesures) d'une caractéristique quelconque d'une unité est représentée par y_{hi}, où h désigne la strate et i, la ^{ième} unité de cette strate. Le paramètre de population $\theta = \theta(S)$, où $S = \{y_{hi} : h = 1,2,...,L;\ i = 1, 2, ..., N_h\}$, est estimé habituellement par $\hat{\theta} = \hat{\theta}(s)$, où $s = \{y_{hi} : h = 1, 2, ..., L; i = 1, 2, ..., n_h\}$. Le cas le plus courant est celui où $\theta = \bar{Y}$, la moyenne de population, et où l'estimateur sans biais de ce paramètre $\hat{\theta} = \bar{y} = \sum_{h=1}^{L} W_h \bar{y}_h$, où $\bar{y}_h = \sum_{i=1}^{n_h} y_{hi}/n_h$ et $W_h = N_h/N$. Un estimateur sans biais de $\mathrm{Var}(\bar{y})$ est

$$\mathrm{var}(\bar{y}) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} s_h^2, \qquad (2.1)$$

où $f_h = n_h/N_h$ et $s_h^2 = \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2/(n_h - 1)$.

2.1 Méthode bootstrap avec transformation

Rao et Wu (1988) ont proposé l'utilisation d'une méthode particulière lorsque $\hat{\theta} = g(\bar{t})$, une fonction de moyennes, où $\bar{t} = (\bar{t}_1,...,\bar{t}_s)$. Essentiellement, il s'agit de tirer (avec remise) un vecteur d'unités de l'échantillon original, d'appliquer une certaine transformation à chacune des unités rééchantillonnées et d'appliquer ensuite l'estimateur original au vecteur transformé. Les coefficients de transformation sont choisis de manière que la variance de rééchantillonnage concorde avec l'estimateur de la variance habituellement utilisé pour les fonctions linéaires. Par exemple, supposons que nous ayons au départ un échantillonnage aléatoire stratifié sans remise, où $\bar{t}_\alpha = \sum_h W_h \bar{t}_{\alpha(h)}$ et $t_{\alpha(hi)} = t_{\alpha(y_{hi})}$, une fonction de y_{hi}. Rappelons-nous que y_{hi} peut représenter un vecteur de mesures concernant l'unité i de la strate h, de sorte que de nombreuses statistiques d'usage courant peuvent être exprimées sous la forme $\hat{\theta} = g(\bar{t})$ (par ex.: quotient, coefficient de régression et coefficient de corrélation). Compte tenu de ce qui précède, la méthode bootstrap avec transformation s'énonce comme suit:

1. Prélever avec remise un échantillon aléatoire simple $\{y_{hi}^*\}_{i=1}^{n_h^*}$ $(n_h^* \geq 1)$ dans $\{y_{hi}\}_{i=1}^{n_h}$. Poser $C_h = \sqrt{n_h^*}(n_h - 1)^{-1/2}(1 - f_h)^{1/2}$, et calculer les paramètres suivants:

$$\bar{t}^*_{a(h)} = \bar{t}_{a(h)} + C_h(\bar{t}^*_{a(h)} - \bar{t}_{a(h)})$$

$$\bar{t}^*_a = \sum_{h=1}^{L} W_h \bar{t}^*_{a(h)}$$

$$\hat{\theta}^* = g(\bar{t}^*)$$

2. Répéter l'étape 1 un grand nombre de fois, B, afin d'obtenir $\hat{\theta}^*_1$, $\hat{\theta}^*_2$, ..., $\hat{\theta}^*_B$,, puis estimer $\mathrm{Var}(\hat{\theta})$ au moyen de la formule $v_r = E_*(\hat{\theta}^* - E_*\hat{\theta}^*)^2$ ou de l'approximation de Monte Carlo $v_r = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}^*_i - \hat{\theta}^*_{(.)})^2$ ($E_*\hat{\theta}^*$ et $\hat{\theta}^*_{(.)}$ peuvent être remplacés par $\hat{\theta}$).

Si $g(\bar{t}) = \bar{y}$ (unidimensionnel), alors $v_r(\bar{y}^*) = \mathrm{var}(\bar{y})$, qui est l'estimateur non biaisé habituel de $\mathrm{Var}(\bar{y})$. Pour réduire le volume de calculs, on peut appliquer la transformation à toutes les unités de l'échantillon avant de commencer la méthode bootstrap.

Rao et Wu (1988) ont élaboré des algorithmes de transformation pour un large éventail de plans d'échantillonnage, notamment l'échantillonnage stratifié, l'échantillonnage à deux degrés et l'échantillonnage avec PPT de Rao-Hartley-Cochran (1962). Ces algorithmes produisent tous les estimateurs de variance habituels pour des fonctions linéaires. Rao et Wu montrent que dans le cas de l'échantillonnage stratifié, la méthode donne des estimateurs de la variance convergents pour $\hat{\theta} = g(\bar{t})$ et que pour des fonctions linéaires, moyennant un choix judicieux de n^*_h dans chaque strate, l'histogramme bootstrap intègre le terme de deuxième ordre du développement d'Edgeworth de $\bar{y}$ lorsque $L \to \infty$.

Néanmoins, la méthode présente certaines lacunes, qui sont analysées dans Sitter (1990).

2.2 Méthode "mirror-match"

Sitter (1990) a proposé une méthode bootstrap que nous appellerons "mirror-match". De manière générale, cette méthode consiste à prélever sans remise des unités dans l'échantillon, en parfaite conformité avec le plan d'échantillonnage initial, puis à répéter l'opération, cette fois avec remise, de manière que la variance de rééchantillonnage concorde avec l'estimateur de la variance habituellement utilisé pour les fonctions linéaires. En ce qui a trait à l'échantillonnage aléatoire stratifié, la méthode "mirror-match" s'énonce comme suit:

1. Choisir une valeur $1 \le n'_h < n_h$ et prélever sans remise un sous-échantillon de taille n'_h dans la strate h pour obtenir $\underline{y}^*_h = (y^*_{h1}, y^*_{h2}, ..., y^*_{hn'_h})$.

2. Répéter l'étape (1) k_h fois ($k_h = \frac{n_h(1-f^*_h)}{n'_h(1-f_h)}$) de façon indépendante, remplaçant à chaque fois le sous-échantillon de taille n'_h dans l'échantillon, pour obtenir y^*_{h1}, y^*_{h2}, ..., $y^*_{hn^*_h}$, où $f^*_h = n'_h/n_h$ et $n^*_h = k_h n'_h$ (si k_h n'est pas un nombre entier, on procède à un arrondissement aléatoire).
3. Répéter les étapes (1) et (2) de façon indépendante pour chaque strate, et poser $\hat{\theta}^* = \hat{\theta}(\underline{y}^*_1, \underline{y}^*_2, ..., \underline{y}^*_L)$.
4. Répéter les étapes (1) à (3) un grand nombre de fois, B, afin d'obtenir $\hat{\theta}^*_1, \hat{\theta}^*_2, ..., \hat{\theta}^*_B$, puis estimer $\mathrm{Var}(\hat{\theta})$ au moyen de la formule $V_m = E_*(\hat{\theta} - E_*\hat{\theta}^*)^2$ ou de l'approximation de Monte Carlo $v_m = \sum_{i=1}^{B}(\hat{\theta}^*_i - \hat{\theta}^*_{(.)})^2 / B$ ($E_*\hat{\theta}^*$ et $\hat{\theta}^*_{(.)}$ peuvent être remplacés par $\hat{\theta}$). Si $\hat{\theta} = \bar{y}$, alors $V_m = \mathrm{var}(\bar{y})$, l'estimateur habituel de la variance. Si $f_h \ge 1/n_h$, alors le fait de choisir $n'_h = f_h n_h$ implique que la fraction

de sondage f_h^* à l'étape (1) est identique à la fraction de sondage initiale, f_h. Ce choix peut être justifié théoriquement à l'aide du développement d'Edgeworth, ce à quoi nous nous appliquons dans le paragraphe suivant.

Partant de cela, Sitter (1990) étend la méthode "mirror-match" à l'échantillonnage stratifié sans remise, à l'échantillonnage à deux degrés et à l'échantillonnage avec PPT de Rao-Hartley-Cochran (1962). Dans le cas de l'échantillonnage stratifié, Sitter montre que la méthode produit des estimateurs convergents de la variance pour les statistiques non linéaires et que, lorsque $\hat{\theta} = \bar{y}$, moyennant un choix judicieux de $n_h'(f_h^* = f_h)$, l'histogramme bootstrap intègre le terme de deuxième ordre du développement d'Edgeworth lorsque $L \to \infty$, de même que lorsque L est borné et que $n, N \to \infty$. La méthode bootstrap avec remise (BAR), proposée par McCarthy et Snowden (1985), est un cas particulier de la méthode "mirror-match" pour les plans d'échantillonnage auxquels peut être appliquée une méthode BAR (dans le cas d'un échantillonnage stratifié lorsque $n_h' = 1$).

Bien qu'elle ne soulève pas les problèmes que pose la première méthode, la méthode bootstrap "mirror-match" a aussi ses lacunes: 1) si la fraction de sondage de la strate est très petite, il n'est pas possible de choisir la taille du sous-échantillon qui permettra une correspondance avec le terme de deuxième ordre du développement d'Edgeworth de $\bar{y}$; 2) même si les n_h' sont des nombres entiers, il faut procéder à un arrondissement aléatoire pour k_h.

2.3 Méthode BSR

La méthode ci-dessous a été présentée par Bickel et Freedman (1984) comme une version de la méthode BSR (Gross, 1980) applicable à l'échantillonnage stratifié. McCarthy et Snowden (1985) s'y sont aussi intéressés. Supposons que $N_h = k_h n_h + r_h$ pour $0 \le r_h \le (n_h - 1)$, pour chaque strate, k_h et r_h étant des entiers. Construisons deux pseudo-populations pour chaque strate, la population 1 étant formée par k_h répétitions de y_h et la population 2, par $k_h + 1$ répétitions de y_h. Pour chaque strate prise individuellement, tirons sans remise n_h unités de la population 1 avec une probabilité

$$p_h = \frac{\dfrac{(1-f_h)}{(n_h-1)} - a_{h1}}{a_{h0} - a_{h1}},$$

où $a_{hj} = [k_h + j - 1] / [(k_h + j) n_h - 1]$ pour $j = 0, 1$, et le même nombre d'unités de la population 2 avec une probabilité $1 - p_h$. Si cela peut être réalisé, alors $\mathrm{Var}_*(\bar{y}^*) = \mathrm{var}(\bar{y})$. Malheureusement, il se peut que $p_h < 0$. McCarthy et Snowden (1985) en donnent une illustration pour une strate. Afin de mettre en lumière les limites de cette version de la méthode BSR, nous avons indiqué dans le tableau 1 les valeurs de p pour diverses valeurs de n et de N pour une strate dans l'échantillonnage stratifié.

Tableau 1: Valeurs de p pour la méthode BSR de Bickel et Freedman

	N					
n	25	30	40	50	75	100
2	-253.	-404.	-739.	-1174.	-2628.	-4849.
4	-12.0	-17.6	-37.7	-57.0	-134.0	-263.0
6	-1.90	-3.48	-6.79	-12.5	-30.8	-56.8
8	-0.04	-0.81	-2.07	-4.03	-11.0	-20.9
10	-0.14	0.36	-0.36	-1.33	-4.66	-9.90
15	0.17	0.89	0.07	0.16	-0.45	-2.00
20	0.59	0.32	0.92	0.28	-0.15	-0.05

3. VERSION ÉLARGIE DE LA MÉTHODE BSR

La méthode BSR pour l'échantillonnage aléatoire simple est intéressante de prime abord, n'implique aucune transformation et est facile à utiliser. Malheureusement, elle ne peut s'appliquer à des plans d'échantillonnage plus complexes. Il est donc difficile de la considérer sur le même pied que la méthode avec transformation et la méthode "mirror-match". Cela nous amène à proposer dans cette section des versions élargies de la méthode BSR, qui puissent s'appliquer à l'échantillonnage stratifié, à l'échantillonnage à deux degrés et à l'échantillonnage avec probabilités inégales de Rao-Hartley-Cochran. Les trois versions produisent les estimateurs habituels de la variance pour les fonctions linéaires et sont applicables de façon générale.

3.1 Échantillonnage stratifié

Créons la strate h de la pseudo-population par k_h répétitions de $\underline{y}_h = (y_{h1}, ..., y_{hn_h})$. Faisons de même pour chaque strate et tirons n_h' unités de la strate h sans remise. On remarquera que la méthode est la même qu'auparavant sauf qu'il n'est plus nécessaire que $N_h = k_h n_h$ et que la taille du sous-échantillon soit n_h. On choisit plutôt n_h' et k_h de manière que les équations suivantes soient satisfaites:

$$f_h^* = f_h, \text{ et } \mathrm{Var}_*(\bar{y}_h^*) = \frac{(1-f_h)}{n_h} s_h^2, \tag{3.1}$$

où $f_h^* = n_h'/k_h n_h$ est la fraction de rééchantillonnage. Si nous faisons abstraction pour l'instant de toutes les contraintes relatives aux entiers, les deux équations ci-dessus sont satisfaites si

$$n_h' = n_h - (1-f_h), \text{ et } k_h = \frac{N_h}{n_h}\left(1 - \frac{1-f_h}{n_h}\right). \tag{3.2}$$

De toute évidence, la version élargie de la méthode BSR se rapprochera sensiblement de la méthode originale dans beaucoup de cas.

Évidemment, les propos ci-dessus n'ont du sens que si n_h' et k_h sont des entiers pour tous h, et comme $0 \le f_h \le 1$, $n_h' = n_h - (1 - f_h)$ n'est pas un nombre entier à moins que $f_h = 0$, ou 1. Pour résoudre cette impasse, on procède à un arrondissement aléatoire. Comme cette technique est la même pour chaque strate, nous la décrivons ici en faisant abstraction de l'indice h.

Soit $k_1 = \lfloor k \rfloor$ et $k_2 = \lfloor k \rfloor$, où k est défini en (3.2), et soit $n_1' = n - 1$ et $n_2' = n$. De plus, posons

$$p = \frac{\frac{1-f}{n(n-1)} - a_2}{a_1 - a_2}, \tag{3.3}$$

où $a_i = [k_i(1 - n_i'/nk_i)]/[n_i'(nk_i - 1)]$ pour $i = 1, 2$. On peut montrer facilement que, pour $n \ge 2$, $0 \le p \le 1$. La méthode est donc applicable de façon générale.

Si $n \ge 2$ et que l'on utilise (k_1, n_1') avec une probabilité p, définie en (3.3), et (k_2, n_2') avec une probabilité $(1 - p)$, alors $\mathrm{Var}_*(\bar{y}^*) = \mathrm{var}(\bar{y})$. Pour vérifier cela, considérons k et n' tels qu'ils sont définis ci-dessus. Posons E_{2*} et V_{2*} comme l'espérance et la variance conditionnelles pour le rééchantillonnage, étant donné k et n'. De même, désignons par E_{1*} et V_{1*} l'espérance et la variance par rapport à la randomisation de (n', k). Alors,

$$\begin{aligned}\mathrm{Var}_*(\bar{y}^*) &= E_{1*}V_{2*}(\bar{y}^*) + V_{1*}E_{2*}(\bar{y}^*) = E_{1*}V_{2*}(\bar{y}^*)\\ &= pV_{2*}(\bar{y}^* \mid k_1, n_1') + (1-p)V_{2*}(\bar{y}^* \mid k_2, n_2')\\ &= (n-1)s^2[pa_1 + (1-p)a_2] = \frac{1-f}{n}s^2.\end{aligned}$$

3.2 Échantillonnage à deux degrés

Pour étendre la méthode BSR à l'échantillonnage à deux degrés, on peut appliquer le même principe général qu'en 3.1 à chaque degré de l'échantillonnage. En l'occurrence, $\hat{\bar{Y}} = \sum_1^n M_i\bar{y}_i / n\bar{M}_0$ et l'estimateur habituel de $\mathrm{Var}(\hat{\bar{Y}})$ est

$$\mathrm{var}(\hat{\bar{Y}}) = \frac{1-f_1}{n}s_1^2 + \sum_{i=1}^{n}\frac{f_1(1-f_{2i})}{nm_i}s_{2i}^2, \tag{3.4}$$

où $f_1 = n/N$, $f_{2i} = m_i/M_i$,

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i\bar{y}_i}{\bar{M}_0} - \hat{\bar{Y}}\right)^2, \text{ et } s_{2i}^2 = \frac{1}{n(m_i-1)}\sum_{j=1}^{m_i}\left(\frac{M_i}{\bar{M}_0}\right)^2(y_{ij} - \bar{y}_i)^2.$$

Créons une pseudo-population en répétant chaque grappe de l'échantillon k_1 fois et chaque unité de la grappe i k_{2i} fois. On forme un vecteur d'unités rééchantillonnées à partir de cette pseudo-population en prélevant n' grappes sans remise, puis m_i' unités sans remise dans chaque grappe rééchantillonnée qui est une répétition de la grappe i. Soit $\hat{\bar{Y}}^{**} = \frac{i}{n'}\sum_1^{n'}\frac{M_i}{\bar{M}_0}\bar{y}_i^{**}$, où $\bar{y}_i^{**} = \frac{i}{m_i'}\sum_1^{m_i'}y_{ij}^{**}$. Alors,

$$\mathrm{Var}_{**}(\hat{\bar{Y}}^{**}) = \frac{k_1(n-1)}{(k_1n-1)}\frac{(1-f_1^*)}{n'}s_1^2 + \sum_{i=1}^{n}\frac{k_{2i}(m_i-1)}{(k_{2i}m_i-1)}\frac{(1-f_{2i}^*)}{n'm_i'}s_{2i}^2, \tag{3.5}$$

où $f_1^* = n'/k_1n$ et $f_{2i}^* = m_i'/k_{2i}m_i$, pour n'importe quelle valeur de k_1, k_{2i}, n', et m_i'. On peut choisir les valeurs de n' et de k_1 de manière que

$$f_1^* = f_1, \text{ et } \frac{k_1(n-1)}{n'(k_1n-1)} = \frac{1}{n},$$

et les valeurs de m_i' et de k_{2i} de manière que

$$f_{2i}^* = f_{2i}, \text{ et } \frac{k_{2i}(m_i-1)}{n'm_i'(k_{2i}m_i-1)} = \frac{f_1}{nm_i}$$

pour chaque i. Si cela peut être réalisé, $\mathrm{Var}^{**}(\hat{\bar{Y}}^{**}) = \mathrm{var}(\hat{\bar{Y}})$.

Pour éviter d'obtenir des tailles de sous-échantillon qui ne sont pas des nombres entiers avec cette version de la méthode BSR, on peut procéder à un arrondissement aléatoire. Soit $k_{11} = \lfloor k_1 \rfloor$ et $k_{12} = \lfloor k_1 \rfloor$, où $k_1 = [1 - (1 - f_1)/n]/f_1$; soit $n_1' = n - 1$ et $n_2' = n$; soit $k_{2i1} = \lfloor k_{2i} \rfloor$ et $k_{2i2} = \lfloor k_{2i} \rfloor$, où

$$k_{2i} = \frac{m_i - 1 - f_1 f_{2i} n'/n}{m_i f_1 f_{2i} n'/n} ;$$

et soit $m_{i1}' = \lfloor m_i' \rfloor$ et $m_{i2}' = \lceil m_i' \rceil$, où

$$m_i' = \frac{m_i - 1 - f_1 f_{2i} n'/n}{f_1 n'/n} .$$

Dans les équations ci-dessus, n' représente une variable aléatoire qui prend les valeurs n_1' et n_2' avec les probabilités définies ci-dessous, et m_i' et k_{2i} dépendent de n'. Définissons

$$p_1 = \frac{\frac{1-f_1}{n(n-1) - a_2}}{a_1 - a_2}, \quad \text{et} \quad p_{2i} = \frac{\frac{f_1(1-f_{2i})}{nm_i(m_i-1)} - b_{i2}}{b_{i1} - b_{i2}},$$

où $a_j = [k_{1j}(1 - f_{1j})]/[n_j'(nk_{1j} - 1)]$, $b_{ij} = [k_{2ij}(1 - f_{2ij})]/[n'm_{ij}'(k_{2ij}m_i - 1)]f_{1j} = n_j'/nk_{1j}$ et $f_{2ij} = m_{ij}'/(k_{2ij}m_i - 1)$, pour $j = 1, 2$. Notons que p_{2i} dépend de la valeur de n' choisie au premier degré du rééchantillonnage. Notons de plus que k_{2ij} et m_{ij}, $j = 1, 2,$, dépendent de la valeur de n' choisie au premier degré. Il est possible de montrer que $0 \le p_1 \le 1$ et que, étant donné n', $0 \le p_{2i} \le 1$ pour chaque i si $n \ge 2$ et $m_i \ge 2$. Sitter (1989) montre que si on utilise cet algorithme, $\mathrm{Var}_{**}(\hat{\bar{Y}}^{**}) = \mathrm{var}(\hat{\bar{Y}})$.

3.3 Échantillonnage avec PPT de Rao-Hartley-Cochran (RHC)

Rao, Hartley et Cochran (1962) ont proposé une méthode simple d'échantillonnage avec probabilités inégales sans remise. Cette méthode s'énonce comme suit:

1. Diviser aléatoirement une population de N unités en n groupes $\{G_g\}_{g=1}^{n}$ de tailles respectives $\{N_g\}_{g=1}^{n}$.
2. Tirer une unité de chaque groupe avec une probabilité z_j / Z_g pour le groupe g, où $z_j = x_j/X$, $Z_g = \sum_{j \in G_g} z_j$, $x_j =$ une mesure quelconque de la taille de l'unité j, et $X = \sum_{j=1}^{N} x_j$.

Un estimateur sans biais de $\bar{Y}$, la moyenne de population, est $\hat{\bar{Y}} = \sum_{g=1}^{n} w_g y_g/n$, où $w_g = f/\pi_g$, $\pi_g = z_g/Z_g$ est la probabilité de sélection de l'unité échantillonnée du groupe g, $f = n/N$ est la probabilité de sélection suivant l'échantillonnage aléatoire simple sans remise, et y_g et z_g désignent les valeurs relatives à l'unité tirée du groupe g. Notons que, par définition, $\sum_{g=1}^{n} Z_g = 1$. Un estimateur sans biais de $\mathrm{Var}(\hat{\bar{Y}})$ est

$$\mathrm{var}(\hat{\bar{Y}}) = \frac{(\sum_{g=1}^{n} N_g^2 - N)}{(N^2 - \sum_{g=1}^{n} N_g^2)} \sum_{g=1}^{n} Z_g \left(\frac{y_g}{N_{zg}} - \hat{\bar{Y}}\right)^2 . \qquad (3.6)$$

Afin d'adapter la méthode BSR à ce type d'échantillonnage, posons $\hat{Y}_{RHC} = N\hat{\bar{Y}}_{RHC} = \sum_{g=1}^{n} Z_g y_g / z_g$. La méthode BSR s'énonce alors comme suit:

1. Répéter la paire (z_g, y_g), $k_g = Z_g / z_g$ fois, $g = 1, ..., n$, pour créer une pseudo-population.
2. Diviser aléatoirement la pseudo-population de $N^* = \sum_{g=1}^{n} k_g$ unités en n^* groupes $\{\Gamma_g^*\}_{g=1}^{n^*}$ de tailles respectives $\{N_g^*\}_{g=1}^{n^*}$.
3. Prélever aléatoirement une paire (z_i^*, y_i^*) dans chacun des n^* groupes avec une probabilité z_i^* / Z_g^*, où $Z_g^* = \sum_{i \in \Gamma_g^*} z_i^*$, et poser $\hat{\theta}^* = \hat{\theta}(\underline{z}^*, \underline{y}^*)$.
4. Répéter les étapes 1 à 3 un grand nombre de fois, B, afin d'obtenir $\hat{\theta}_1^*, ..., \hat{\theta}_B^*$, puis estimer $\mathrm{Var}(\hat{\theta})$ au moyen de la formule $v_{bwo} = E_*(\hat{\theta}^* - E_*\hat{\theta}^*)^2$ ou de l'approximation de Monte Carlo $v_{bwo} = \sum_{i=1}^{B} (\hat{\theta}_i^* - \hat{\theta}_{(\cdot)}^*)^2 / (B - 1)$ ($E_*\hat{\theta}^*$ et $\hat{\theta}_{(\cdot)}^*$ peuvent être remplacés par $\hat{\theta}$).

Si $\theta = Y$ et $\hat{\theta} = \hat{Y}_{RHC}$, nous avons

$$\mathrm{Var}_*(\hat{\theta}^*) = \left(\frac{\sum_{g=1}^{n^*} N_g^{*2} - N^*}{\sum_{g=1}^{n} N_g^2 - N} \right) \left(\frac{N^2 - \sum_{g=1}^{n} N_g^2}{N^*(N^* - 1)} \right) \mathrm{var}(\hat{\theta}). \tag{3.7}$$

Cette équation se vérifie puisque $\hat{Y}_{RHC}^* = \sum_{1}^{n^*} Z_g^* y_g^* / z_g^*$ et par conséquent

$$\mathrm{Var}_*(\hat{Y}_{RHC}^*) = \frac{\sum_{g=1}^{n^*} N_g^{*2} - N^*}{N^*(N^* - 1)} \left(\sum_{i=1}^{N^*} \frac{y_i^{*2}}{z_i^*} - Y^{*2} \right),$$

où Y^{*2} est le total de la pseudo-population. La répétition de données dans la pseudo-population nous amène directement aux conclusions.

Nous avons donc deux solutions possibles. La plus simple consiste à redresser l'estimateur de la variance par un facteur multiplicatif approprié. L'autre consiste à choisir n^* et N_g^* pour $g = 1, ..., n^*$, de manière que le produit des deux premiers facteurs de l'équation (3.7) soit voisin de un. Dans beaucoup de cas, on devrait pouvoir trouver des valeurs d'encadrement de n^* et de N_g^* et procéder à un arrondissement aléatoire, mais il ne semble pas exister de solution avec une expression analytique générale. Pour une analyse plus fouillée des solutions qui s'offrent à nous, voir Sitter (1989).

4. ÉTUDE DE SIMULATION

Dans cette section, nous décrivons une étude de simulation qui visait à comparer l'efficacité des méthodes exposées ci-dessus en ce qui regarde l'estimation de la variance et la construction d'intervalles de confiance pour un échantillonnage aléatoire stratifié sans remise. Huit populations finies sont étudiées dans les circonstances. Sitter (1990) rapporte une partie des résultats de la simulation pour les populations 3 et 7; son rapport fait abstraction des versions élargies de la BSR et de certaines tailles d'échantillon. Pour que cette étude soit plus complète, nous y avons inclus les méthodes de linéarisation et du jackknife, pour les statistiques non linéaires. Pour la médiane une méthode fondée sur les intervalles de confiance de Woodruff (1952) pour les quantiles et la jackknife ont aussi été considérées. Pour une description détaillée de ces trois méthodes, voir Kovar, Rao et Wu (1988). Pour les besoins de notre étude, de légères modifications ont été nécessaires en ce qui a trait à la fraction de sondage étant donné que Kovar, Rao et Wu étudient le cas de l'échantillonnage avec remise. Ceux-ci ont constaté très peu de différence entre les six versions jackknife qu'ils ont analysées pour l'échantillonnage

stratifié avec remise. En ce qui nous concerne, l'échantillonnage dans chaque strate se faisait sans remise, mais compte tenu des résultats de Kovar, Rao et Wu, il est peu probable que l'on observe de grandes différences d'efficacité entre ces six versions jackknife. Par conséquent, nous n'avons utilisé qu'une seule de ces versions (celle désignée par v_{J2}), que nous définissons plus bas.

Les estimateurs que nous analysons ici concernent le quotient, le coefficient de régression, le coefficient de corrélation et la médiane, définie plus loin. En reprenant la notation utilisée dans la section 2 pour l'échantillonnage stratifié, posons y_{hi} comme la caractéristique d'intérêt de l'observation i dans la strate h et x_{hi} comme une variable concomitante pertinente. Soit $t_{hi1} = x_{hi}$, $t_{hi2} = y_{hi}$, $t_{hi3} = x_{hi}y_{hi}$, $t_{hi4} = x_{hi}^2$, et $t_{hi5} = y_{hi}^2$. Les estimateurs empiriques utilisés sont les suivants (pour les paramètres de population correspondants): 1) quotient, $r = \bar{t}_2/\bar{t}_1$; 2) coefficient de régression, $b = (\bar{t}_3 - \bar{t}_1\bar{t}_2)/(\bar{t}_4 - \bar{t}_1^2)$; 3) coefficient de corrélation, $c = [\bar{t}_3 - \bar{t}_1\bar{t}_2]/[(\bar{t}_4 - \bar{t}_1^2)(\bar{t}_5 - \bar{t}_2^2)]^{1/2}$; 4) médiane, m, avec $p = 1/2$. Notons que $\bar{t}_j = \sum_{h=1}^{L} W_h\bar{t}_{hj}$ où $\bar{t}_{hj} = \sum_{i=1}^{n_h} t_{hij}/n_h$ pour $j = 1, \ldots, 5$.

4.1 Intervalles de confiance

Pour les méthodes de linéarisation et du jackknife, les intervalles de confiance sont construits selon la théorie normale et sont définis par la formule $[\hat{\theta} - z_{\alpha/2}\sqrt{v}, \hat{\theta} + z_{\alpha/2}\sqrt{v}]$, où v est l'estimation de la variance calculée à l'aide de l'une ou l'autre des méthodes et $z_{\alpha/2}$ est la limite supérieure (à $\alpha/2$ %) pour la distribution normale centrée réduite. Pour ce qui a trait aux diverses méthodes bootstrap, on peut obtenir un histogramme des B estimations bootstrap $\hat{\theta}_b^*$. En se servant de ces valeurs, on peut construire des intervalles de confiance à l'aide de la méthode des percentiles (Efron, 1982). C'est cette méthode qui est utilisée pour la médiane. En ce qui a trait aux fonctions non linéaires de moyennes, on se sert de la méthode bootstrap-t. Dans ce cas, l'histogramme bootstrap de $t_b^* = (\hat{\theta}_b^* - \hat{\theta})/\sqrt{v_J^*}$, où v_J^* est un estimateur jackknife appliqué aux observations rééchantillonnées, sert à déterminer t_U^* et t_L^*, les limites supérieure et inférieure (à $\alpha/2$ %) de t^*. L'intervalle de confiance bootstrap-t à $(1 - \alpha)$% est alors défini $[\hat{\theta} - t_U^*\sqrt{v_J}, \hat{\theta} - t_L^*\sqrt{v_J}]$, où v_J est un estimateur jackknife appliqué à tout l'échantillon. À l'aide de développements asymptotiques, Abramovitch et Singh (1985) ont démontré la supériorité des intervalles bootstrap-t unilatéraux par rapport aux intervalles de la théorie normale pour les cas de variables indépendantes et identiquement distribuées.

4.2 Formation des populations finies

Les populations finies qui ont servi à l'étude de simulation reposent sur une population hypothétique définie dans Hansen et Tepping (1985), et qu'ont approfondie Kovar, Rao et Wu (1988), et sur une autre population hypothétique définie celle-là dans Kovar, Rao et Wu (1988) et fondée elle-même sur une population décrite dans Hansen, Madow et Tepping (1983). La population de Hansen et Tepping (1985) était présentée comme une approximation des populations réelles que l'on trouvait dans la National Assessment of Educational Progress Study. Elle était divisée en $L = 32$ strates et mettait en évidence les variables aléatoires normales bidimensionnelles (x_h, y_h), distribuées suivant $N_2(\mu_{xh}, \mu_{yh}, \sigma_{xh}, \sigma_{yh}, \rho)$ pour $h = 1, \ldots, L$, ρ étant identique pour toutes les strates. La valeur de W_h varie de 0.013 à 0.042; de plus, les moyennes et les variances de strate sont telles que les coefficients de variation de x et de y sont approximativement 10 et 30% respectivement. Les valeurs des paramètres et les poids de strate figurent dans Sitter (1989). Kovar, Rao et Wu (1988) ont étendu la population hypothétique de Hansen et Tepping à des distributions Gamma en conservant la même moyenne et la même structure de covariances. Ainsi, supposons que x_h suit une distribution Gamma(α,β) avec comme paramètre de forme $\alpha = \mu_{xh}^2/\sigma_{xh}^2$ et comme paramètre d'échelle $\beta = \sigma_{xh}^2/\mu_{xh}$, de sorte que x_h a pour moyenne μ_{xh} et pour variance σ_{xh}^2. On peut déduire y_h de x_h à l'aide d'un modèle de régression linéaire, $y_h = (\sigma_{yh}\rho/\sigma_{xh})x_h + \epsilon_h$, où ϵ_h suit une distribution normale de moyenne $\mu_{yh} - \sigma_{yh}\mu_{xh}\rho/\sigma_{xh}$ et de variance $\sigma_{yh}^2(1 - \rho^2)$. Ainsi, (x_h, y_h) suit une distribution conjointe caractérisée par les moments préétablis.

Les populations finies ont été formées à partir des populations hypothétiques, puis soumises à des simulations. Les méthodes étudiées comprennent un certain nombre de méthodes bootstrap, qui sont des méthodes de calculs intensifs sur ordinateur. En outre, comme les fractions de sondage de strate peuvent jouer un rôle important dans la comparaison des diverses méthodes, il serait juste d'utiliser diverses tailles de strate pour chaque population finie. Compte tenu de ces observations, il est évident que des contraintes d'ordre budgétaire limiteront le nombre de populations pouvant être étudiées. Nous nous sommes donc inspirés des résultats généraux obtenus par Kovar, Rao et Wu (1988) pour choisir des paramètres de populations hypothétiques de manière à former des populations finies qui feraient ressortir les différences de rendement entre les diverses méthodes.

Afin de comparer le rendement des diverses méthodes en ce qui a trait à l'estimation de la variance et à la construction d'intervalles de confiance pour r, b et c, nous avons formé quatre populations finies. Deux d'entre elles ont été produites à l'aide de paramètres semblables à ceux de la population de Hansen et Tepping, sauf que σ^2_{xh} est multiplié par 20. Chacune de ces deux populations compte $N = 800$ unités réparties dans $L = 16$ strates, dont la taille, N_h, varie de 22 à 69. La population 1 a été formée à partir d'une distribution normale bidimensionnelle pour laquelle les moyennes de strate varient de 45 à 95, les écarts types de strate varient de 12 à 44 et $\rho = 0.8$ (les valeurs réelles figurent dans Sitter, 1989). La population 2 a été obtenue à l'aide d'une distribution Gamma et d'un modèle de régression linéaire, de la manière décrite plus haut, et affiche les mêmes paramètres que la population 1. Les populations 3 et 4 ont été produites à l'aide d'une distribution normale bidimensionnelle et des mêmes paramètres mais elles sont réparties dans un moins grand nombre de strates. Elles comptent, respectivement, $N = 760$ et $N = 516$ unités réparties dans $L = 6$ strates, dont la taille, N_h, varie de 80 à 200 et de 36 à 144 respectivement. Les populations 1 et 2 comptent donc un plus grand nombre de strates mais moins d'unités par strate que les populations 3 et 4.

Afin de comparer le rendement des diverses méthodes pour ce qui a trait à l'estimation de la variance et à la construction d'intervalles de confiance pour la médiane estimée, m, nous avons formé quatre populations finies, numérotées de 5 à 8. Les quatre ont été produites à l'aide de distributions normales selon le même processus que ci-dessus, sauf que seules les valeurs y_h étaient utilisées. La population 5 est semblable à la population 1; elle se compose de $N = 800$ unités réparties dans $L = 16$ strates. Les valeurs de μ_{yh} et de N_h utilisées pour la population 5 sont identiques à celles utilisées pour y_h dans la population 1, sauf que l'écart type de y_h a été divisé par un facteur de $1/\sqrt{2}$. La population 6 est semblable à la population 4; elle se compose de $n = 516$ unités réparties dans $L = 6$ strates et les valeurs de μ_{yh} et de N_h utilisées sont identiques à celles utilisées pour y_h dans la population 4, sauf que l'écart type de y_h a été divisé par un facteur de 10. Les populations 7 et 8 comprennent chacune $N = 800$ unités réparties dans $L = 32$ strates. Kovar, Rao et Wu (1988) ont analysé les valeurs de paramètres utilisées pour la population 7 pendant qu'ils comparaient le rendement de la méthode bootstrap avec transformation à celui d'autres méthodes comme la méthode de linéarisation, la méthode jackknife et la méthode de Woodruff. Ils avaient alors étudié le cas de l'échantillonnage avec remise dans une population infinie. Pour calculer les valeurs des paramètres, ils ont tout d'abord formé une population à deux variables de la manière décrite dans Hansen, Madow et Tepping (1983), où (x,y) sont fortement corrélés, puis ils ont réparti les unités y dans 32 strates en se servant des valeurs x et enfin, ils ont calculé μ_{yh}, σ_{yh} et W_h. Compte tenu des valeurs des paramètres de la population 7 et de la façon dont elles sont calculées, il est clair que les valeurs d'une population finie produite à l'aide de ces paramètres se retrouveront rarement dans plus d'une strate à la fois. Ainsi, les valeurs de W_h indiquent clairement que la médiane calculée devrait se situer vers le milieu de la strate 6. Quant à la population 8, nous l'avons formée à partir de valeurs N_h légèrement différentes de manière à obtenir une population dont la médiane se situe près d'une limite de strate. On trouvera dans Sitter (1989) les moyennes et les écarts types de strate de même que les valeurs de R, de B, de C et de la médiane, le cas échéant, pour les huit populations finies.

4.3 Mesures de rendement

Afin de simplifier notre étude, nous allons utiliser la même taille d'échantillon pour chacune des strates, $n_h = n_0$ pour $h = 1, \ldots, L$; la taille de l'échantillon global sera donc $n = n_0 L$. Pour toutes les méthodes bootstrap

étudiées, $\hat{\theta}$ est substitué à $\hat{\theta}^*_{(.)}$ dans la formule de l'approximation de Monte Carlo. De plus, toutes les méthodes sont comparées à l'EQM réelle de l'estimateur d'intérêt de la façon suivante. On a tout d'abord estimé l'EQM réelle en prélevant 3000 échantillons aléatoires stratifiés sans remise puis en utilisant la formule $\text{EQM} = \sum_{b=1}^{3000} (\hat{\theta}_b - \theta)^2 / 3000$. On a ensuite déterminé le biais relatif et la stabilité relative des estimateurs de la variance de même que les taux d'erreur pour intervalles unilatéraux et bilatéraux et la longueur normalisée des intervalles de confiance de la façon suivante.

La simulation consistait à prélever S nouveaux échantillons aléatoires stratifiés et à calculer pour chacun d'eux les estimations de la variance et des limites de confiance. On évaluait ensuite le biais relatif d'un estimateur de variance particulier v au moyen de la formule "$\text{biais} = \dfrac{\bar{v} - \text{EQM}}{\text{EQM}}$", où $\bar{v}$ est la moyenne des S estimations de la variance. La stabilité relative de l'estimateur de la variance était estimée au moyen de la formule $\text{stab} = \dfrac{\sigma_v}{\text{EQM}}$, où $\sigma_{v^2} = \sum_s (v_s - \text{EQM})^2 / S$. On comparait enfin le biais relatif et la stabilité relative des divers estimateurs de la variance.

Pour ce qui a trait à la comparaison des divers estimateurs d'intervalles de confiance, on a comparé les taux d'erreur aux extrémités des distributions avec les taux théoriques de 5 et de 10% dans chaque extrémité. On a aussi comparé entre elles les longueurs d'intervalle normalisées obtenues à l'aide des diverses méthodes. Les taux d'erreur à chaque extrémité et les longueurs normalisées ont été établis au moyen des formules suivantes:

$$U = \frac{\text{no. d'échantillons pour lesquels } \theta < \theta_{Us}}{S},$$

$$L = \frac{\text{no. d'échantillons pour lesquels } \theta < \theta_{Ls}}{S},$$

$$\text{longueur} = \frac{\sum_s (\theta_{Us} - \theta_{Ls}) / S}{2 z_{\alpha/2} \sqrt{\text{EQM}}},$$

où $(\theta_{Ls}, \theta_{Us})$ est l'intervalle de confiance estimé à partir de l'échantillon s et $z_{\alpha/2}$ est la limite supérieure (à $\alpha/2$ %) d'une distribution normale centrée réduite.

En ce qui a trait aux statistiques non linéaires r, b et c, les estimations bootstrap de la variance et des intervalles de confiance reposaient sur B = 200 échantillons bootstrap pour S = 1000 échantillons simulés. Pour des niveaux de confiance théoriques de 0.05 et de 0.1, la simulation de 1000 échantillons donnera des niveaux empiriques inférieurs à 0.01 et à 0.02 respectivement dans 95% des cas. Pour ce qui a trait à la médiane, les estimations bootstrap de la variance et des intervalles de confiance reposaient sur B = 500 échantillons bootstrap pour S = 500 échantillons simulés. Pour des niveaux de confiance théoriques de 0.05 et de 0.1, la simulation de 500 échantillons donnera des niveaux empiriques inférieurs à 0.01 et à 0.02 respectivement dans 85% des cas.

4.4 Statistiques non linéaires

Rappelons-nous d'après la section 4.2 que les populations 1 à 4 servaient à comparer des méthodes d'estimation appliquées à des fonctions non linéaires de moyennes: r (quotient), b (coefficient de régression) et c (coefficient de corrélation). Dans le cas des populations 1 et 2, on a utilisé, pour chacune, des tailles d'échantillon de 5, 6 et 7 unités par strate tandis que pour les populations 3 et 4, on s'est servi d'échantillons de 20 et 12 unités par strate respectivement.

Le tableau 2 donne l'écart absolu moyen entre le taux d'erreur observé et le taux théorique pour des intervalles de confiance unilatéraux et bilatéraux; la valeur moyenne est calculée en fonction des quatre populations et des tailles d'échantillon considérées (c.-à-d. $\tilde{L}_{5\%} = \sum_i |L_i - 5\%|$ et $\tilde{L + U}_{5\%} = \sum_i |L_i + U_i - 10\%|$). Dans

le même tableau, on trouve le biais relatif absolu moyen en pourcentage et la stabilité relative moyenne des estimateurs de la variance. Toutes les méthodes étudiées donnent de bons résultats pour ces populations. Voici quelques observations pertinentes:

1) Estimateurs de la variance: a) pour le quotient, les méthodes de linéarisation et du jackknife donnent des résultats équivalents en ce qui a trait à la stabilité et au biais relatifs; elles sont toutefois légèrement plus efficaces que les méthodes bootstrap; b) pour ce qui a trait aux coefficients de régression et de corrélation, il y a très peu de différence entre les méthodes bootstrap et la méthode jackknife; c) pour les mêmes coefficients, la méthode de linéarisation produit un biais relatif plus élevé mais une stabilité relative légèrement moindre par rapport aux autres méthodes.

2) Intervalles de confiance: a) pour le quotient, les différentes méthodes sont aussi efficaces l'une que l'autre. À mesure que s'accroît la non-linéarité ($r \to b \to c$), les méthodes bootstrap reflètent plus exactement les taux d'erreur pour les intervalles unilatéraux que ne le font les méthodes de linéarisation et du jackknife.

Tableau 2: Erreur absolue moyenne dans les extrémités des distributions, biais relatif et stabilité relative pour l'ensemble des populations étudiées (quotient, coeff. de régression et coeff. de corrélation)

	Écart absolu moyen par rapport au niveau théorique							
	5%			10%				
Méthode	$\tilde{L}$	$\tilde{U}$	$L\tilde{+}U$	$\tilde{L}$	$\tilde{U}$	$L\tilde{+}U$	|biais%|[a]	stab[b]
				Quotient				
lin.	.59	.58	.84	.39	1.05	1.19	2.86	.22
jack.	.59	.56	.83	.42	1.05	1.20	2.88	.22
bsr élarg.	.48	.64	.84	.36	.89	1.15	3.24	.25
m-m[c]1	.40	.71	.86	.44	.95	1.09	3.38	.25
m-m 2	.45	.52	.65	.34	.90	.89	3.21	.24
trans[d]1	.48	.84	1.14	.34	1.10	1.29	3.40	.24
trans 2	.50	.55	.65	.31	1.14	1.08	3.43	.24
				Coeff de régression				
lin.	.84	1.33	1.96	1.24	1.14	1.68	5.05	.30
jack.	.70	.91	1.04	1.10	.74	1.19	3.61	.34
bsr élarg.	.89	.60	.59	1.24	.67	1.21	3.41	.33
m-m 1	.81	.41	.70	1.24	.80	1.56	3.31	.32
m-m 2	.84	.41	.82	1.14	.70	1.16	3.34	.32
trans 1	1.06	.32	.99	1.15	.84	1.49	3.11	.32
trans 2	.84	.30	.71	1.31	.79	1.25	3.25	.32
				Coeff. de corrélation				
lin.	2.44	1.11	1.58	3.10	1.51	2.01	4.66	.38
jack.	1.78	1.33	.60	2.23	1.79	1.04	3.41	.42
bsr élarg.	1.15	1.18	.55	1.46	1.21	.85	3.40	.42
m-m 1	.96	1.07	.44	1.54	1.50	.64	2.90	.42
m-m 2	.83	1.30	.60	1.40	1.37	.85	2.48	.40
trans 1	1.01	1.03	.76	1.75	1.43	1.32	2.34	.41
trans 2	.95	1.05	.80	1.50	1.10	1.00	2.58	.42

[a] moyenne (|biais rel. en %|) calculée pour les populations et les tailles d'échantillon considérées.
[b] moyenne (stabilité rel.).
[c] m-m 1: $n_h' = f_h n_h$, m-m 2: $n_h' = 1$
[d] trans 1: taille du sous-échantillon choisie de manière qu'elle concorde avec des troisièmes moments; trans 2: taille du sous-échantillon = $n_h - 1$

Kovar, Rao et Wu (1988) ont observé la même tendance dans leur étude; b) à mesure que s'accroît la non-linéarité, il paraît de plus en plus avantageux, dans la méthode mirror-match, de choisir la taille du sous-échantillon de manière qu'elle concorde avec le terme de deuxième ordre du développement d'Edgeworth de la moyenne. Bien qu'on ne puisse en dire autant de la méthode bootstrap avec transformation, la façon de choisir la taille du sous-échantillon dans cette méthode donne de bien meilleurs résultats que ce qu'avait montré l'étude de Kovar, Rao et Wu (1988). Dans leur étude on avait observé des résultats pitoyables après avoir choisi la taille du sous-échantillon de manière qu'elle s'accorde avec le développement d'Edgeworth de la moyenne pour la méthode Bootstrap avec transformation, ce qui est contraire à ce à quoi on pouvait normalement s'attendre.

4.5 Médiane

Les quatre populations que nous avons étudiées par rapport à la médiane, et qui sont décrites dans la section 4.2, forment naturellement deux ensembles: populations 5 et 6 et populations 7 et 8. Les deux dernières représentent le cas où la stratification se fait à l'aide d'une variable concomitante fortement corrélée avec la caractéristique d'intérêt. Pour les populations 5 et 6, on a utilisé des tailles d'échantillon, dans le premier cas, de 5 et 6 unités par strate et dans le second cas, de 12 unités. En ce qui a trait aux populations 7 et 8, on s'est servi de tailles d'echantillon de 4 à 7 unités par strate dans les deux cas.

Le tableau 3 contient les résultats pertinents pour la population 6 (les résultats obtenus avec la population 5 étaient qualitativement semblables). Il donne le taux d'erreur observé pour des intervalles unilatéraux et bilatéraux de même que la longueur d'intervalle normalisée pour des niveaux théoriques de 5 et 10%, en plus du biais relatif en pourcentage et de la stabilité relative des estimateurs de la variance. Conformément aux résultats théoriques, la méthode jackknife laisse beaucoup à désirer. L'efficacité de l'estimateur fondé sur la méthode de Woodruff dépend de la valeur de α, mais en ce qui concerne la population 6, cet estimateur est plus efficient que les estimateurs bootstrap pour un grand nombre de valeurs de α. En ce qui regarde les intervalles de confiance, il est difficile de départager les méthodes. Les méthodes bootstrap exagèrent les niveaux de confiance et produisent des intervalles qui ont longueur normalisée moindre que celle des intervalles de Woodruff, lesquels minimisent les niveaux de confiance. Notons que les méthodes bootstrap reflètent mal les taux d'erreur pour intervalles unilatéraux.

Tableau 3: Population 6: variances et I.C. pour la médiane

	Taux d'erreur théorique									
	5%			10%			Long. norm.			
Méthode	L	U	L+U	L	U	L+U	5%	10%	% bias	stab
wood	3.8	4.8	8.6	8.2	7.4	15.6	0.96	1.02		
α = .01									22.7	0.64
α = .025									7.4	0.59
α = .05									6.6	0.63
α = .1									2.7	0.68
α = .2									17.8	0.88
jack	13.8	12.4	26.2	18.2	15.2	33.4	1.39	1.39	234.8	6.02
est. élarg.	11.4	3.4	14.8	16.0	6.6	22.6	0.92	0.92	17.2	0.85
m-m*1	11.0	2.8	13.8	15.6	6.6	22.2	0.96	0.92	22.2	0.85
m-m 2	7.2	4.8	12.0	14.4	8.2	22.6	0.96	0.93	30.0	0.95
trans[b]1	10.2	4.0	14.2	15.0	6.6	21.6	0.94	0.94	25.0	0.87
trans 2	10.6	3.8	14.4	15.6	7.2	22.8	0.96	0.99	38.7	1.00

* m-m 1: $n_h' = f_h n_h$, m-m 2: $n_h' = 1$

[b] trans 1: taille du sous-échantillon choisie de manière qu'elle concorde avec les troisièmes moments;
trans 2: taille du sous-échantillon = $n_h - 1$

Tableau 4: Variances pour la médiane

	Population 7				Population 8			
n_o =	5		7		5		7	
Méthode	bias[a]	stab	bias	stab	bias	stab	bias	stab
wood								
α = .01	220.5	2.46	42.7	0.65	612.4	6.75	900.7	9.69
α = .025	-9.2	0.44	-28.5	0.46	570.7	6.27	550.7	7.92
α = .05	18.9	0.59	-6.4	0.47	409.9	5.58	751.8	10.58
α = .1	68.8	1.05	32.9	0.74	623.9	8.23	1523.2	21.18
α = .2	-27.5	0.70	-41.7	0.67	1095.6	14.10	1523.2	21.18
jack	83.3	3.08	77.6	2.79	12.2	2.13	20.5	1.62
bsr élarg.	9.1	0.86	16.0	0.87	2.5	1.32	19.1	1.17
m-m[b]1	14.2	0.84	8.5	0.81	18.6	1.39	5.9	1.01
m-m 2	23.2	0.95	18.2	0.79	3.0	1.26	-35.7	0.76
trans[c]1	38.1	1.19	960.6	13.15	26.3	0.83	large	large
trans 2	8.2	0.87	1.1	0.68	144.2	2.49	large	large

[a] biais en %

[b] m-m 1: $n_h' = f_h n_h$, m-m 2: $n_h' = 1$

[c] trans 1: taille du sous-échantillon choisie de manière qu'elle concorde avec des troisièmes moments; trans 2: taille du sous-échantillon = $n_h - 1$

Tableau 5: Population 7: I.C. pour la médiane

	Taux d'erreur théorique à chaque extrémité						Long. norm.	
	5%			10%			1 - α	
Méthode	L	U	L+U	L	U	L+U	0.8	0.9
				n_o = 5				
wood.	6.0	2.2	8.2	5.8	17.6	23.4	1.26	0.77
jack.	14.8	6.2	19.0	14.8	13.4	28.2	1.09	1.09
bsr élarg.	5.8	7.0	12.8	5.8	17.6	23.4	0.76	0.77
m-m[a]1	5.8	2.2	8.0	5.8	15.4	21.2	0.83	1.04
m-m 2	5.8	2.2	8.0	5.8	13.6	19.4	0.83	0.85
trans[b]1	7.2	17.6	24.8	7.2	17.6	24.8	0.81	1.00
trans 2	9.0	6.2	15.2	9.2	6.2	15.2	0.69	0.89
				n_o = 7				
wood.	0.2	2.6	2.8	6.4	19.0	25.4	1.11	0.69
jack.	18.6	8.0	26.6	21.6	8.0	29.6	1.06	1.06
bsr élarg.	9.0	2.0	11.0	9.0	2.0	11.0	0.77	0.98
m-m 1	9.0	2.0	11.0	9.0	12.8	21.8	0.76	0.96
m-m 2	1.8	2.2	4.0	9.0	15.8	23.8	1.02	0.73
trans 1	2.4	63.0	65.4	5.2	91.0	97.2	1.02	0.20
trans 2	2.6	10.2	12.8	11.2	11.8	23.0	0.83	0.74

[a] m-m 1: $n_h' = f_h n_h$, m-m 2: $n_h' = 1$

[b] trans 1: taille du sous-échantillon choisie de manière qu'elle concorde avec les troisièmes moments; trans 2: taille du sous-échantillon = $n_h - 1$

Le tableau 4 contient les résultats relatifs aux estimateurs de la variance pour les populations 7 et 8. On voit que la BSR élargie de la section 2.3 et la mirror-match sont plus robustes que la méthode Woodruff et la méthode avec transformation en ce qui regarde la stratification; là encore, la méthode jackknife laisse à désirer. Il convient de souligner que même si Kovar, Rao et Wu (1988) ont inclus la médiane dans leur étude de simulation, la méthode bootstrap avec transformation, telle que la présentent Rao et Wu (1988), ne peut s'appliquer qu'à des fonctions de moyennes.

Le tableau 5 contient les résultats relatifs aux intervalles de confiance pour la population 7 avec comme taille d'échantillon n_0 = 5 et 7. Dans ce cas, c'est la méthode mirror-match qui donne les meilleurs résultats.

5. CONCLUSIONS

Dans l'ensemble, parmi les situations considérées dans l'étude, ce sont les méthodes mirror-match et BSR qui donnent les meilleurs résultats. Il n'est pas étonnant de constater que la méthode bootstrap avec transformation produit de bons résultats dans le cas de l'échantillonnage aléatoire stratifié avec les fonctions non linéaires de moyennes étudiées. Dans une situation aussi simple, une transformation appliquée au rééchantillonnage i.i.d. donne un plan qui est proche de l'échantillonnage sans remise. Il est permis de croire qu'à mesure que les estimateurs et le plan d'échantillonnage deviennent plus complexes, les méthodes qui reproduisent le mieux l'échantillonnage original donneront les meilleurs résultats. À cet égard, disons que la méthode mirror-match et, à plus forte raison peut-être, la méthode BSR élargie sont le plus susceptibles de répondre à ces attentes. Les résultats relatifs à la médiane tendent à confirmer cette affirmation, surtout lorsque la stratification se fait au moyen d'une variable concomitante fortement corrélée. De toute évidence, une étude théorique et empirique de ces méthodes s'impose pour des plans d'échantillonnage plus complexes et c'est ce à quoi nous allons nous intéresser dans un proche avenir.

BIBLIOGRAPHIE

Abramovitch, L., et Singh, K. (1985). Edgeworth corrected pivotal Statistics and the bootstrap. *The Annals of Statistics*, 13, 116-132.

Bickel, P.J., et Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia, Society for Industrial and Applied Mathematics.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 181-184.

Hansen, M.H., Madow, W.G., et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-807.

Hansen, M.H., et Tepping, B.J. (1985). Estimation of variance in NAEP. Non publié.

Kovar, J.G., Rao, J.N.K., et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, Supplément, 25-45.

McCarthy, P.J., et Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics* (Ser. 2, No. 95), Public Health Service Publication 85-1369, Washington, DC: U.S. Government Printing Office.

Rao, J.N.K., Hartley, H.O., et Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. Roy. Statistical Social Series*, B, 24, 482-491.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association,* 83, 231-241.

Sitter, R. (1989). Resampling procedures for complex survey data. Ph.D. Thesis, Department of Statistics and Actuarial Science, University of Waterloo.

Sitter, R. (1990). A resampling procedures for complex survey data. Technical Report 149, Department of Mathematics and Statistics, Carleton University.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

CORRECTION SIMPLE POUR LE BIAIS DES ESTIMATEURS DE VARIANCE TRADITIONNELS

Y. Leblond[1]

RÉSUMÉ

Les estimateurs de variance traditionnels d'estimateurs non linéaires de caractéristiques d'une population sont généralement biaisés lorsque la taille d'échantillon est relativement petite. Par exemple, la méthode de linéarisation de Taylor et la méthode dite du jackknife donnent lieu à des estimateurs qui ont respectivement tendance à sous-estimer et surestimer la variance des estimateurs habituellement utilisés en échantillonnage. Nous suggérons une méthode simple permettant de réduire le biais des estimateurs de variance habituellement utilisés. Cette méthode consiste à obtenir un estimateur de variance modifié en multipliant l'estimateur de variance traditionnel par un facteur de correction qui ne dépend que de la taille d'échantillon n et d'information auxiliaire. Ce facteur de correction tend vers 1 lorsque n augmente et ainsi l'estimateur de variance modifié obtenu est asymptotiquement équivalent à l'estimateur traditionnel. Ce facteur est relativement simple à justifier et simple à calculer. Il est obtenu en utilisant une méthode assistée d'un modèle. L'application de cette méthode à l'estimateur de variance de l'estimateur par le quotient est présenté de même que les résultats d'une étude par simulations qui atteste la robustesse de la méthode.

MOTS CLÉS: Estimateur quotient; biais; variance; Taylor; jackknife; modèle de superpopulation.

1. INTRODUCTION

Dans la littérature récente concernant l'estimation de variance en échantillonnage, plusieurs auteurs se sont penchés sur le problème consistant à obtenir un estimateur de variance dont l'erreur quadratique moyenne (EQM) soit minimale dans une certaine classe d'estimateurs. On note surtout les travaux de Wu (1982, 1985), Wu et Deng (1983) et Deng et Wu (1987) concernant l'estimation de variance des estimateurs quotient et régression sous le plan aléatoire simple et stratifié simple. Ces auteurs ont fait une comparaison de différents estimateurs de variance en considérant uniquement la variabilité introduite par la sélection aléatoire de l'échantillon. Ils ont également comparé le biais de différents estimateurs en supposant un modèle de superpopulation. Royall et Cumberland (1978, 1981, 1981a) et Valliant (1987) ont, quant à eux, considéré le problème d'estimation de variance en étudiant les propriétés de différents estimateurs sous certains modèles de superpopulation.

Il ressort de l'ensemble de ces travaux qu'il est difficile de déterminer le "meilleur" estimateur de variance sans faire certaines hypothèses modélistes concernant la population étudiée. Toutefois ces hypothèses sont rarement vérifiées en pratique; d'où la nécessité de trouver de nouvelles méthodes d'estimation de variance plus robustes que celles traditionnellement utilisées.

Le problème ici considéré est celui qui consiste à corriger un estimateur de variance donné afin de diminuer son biais sous le plan d'échantillonnage. Notre but ici n'est pas de proposer un estimateur de variance qui possède

[1] Y. Leblond, Division des méthodes enquêtes-entreprises, 11e étage, édifice R.H. Coats, Statistique Canada, Ottawa, (Ontario), K1A 0T6, Canada.

un biais inférieur à ses concurrents ou qui possède un biais minimum dans une certaine classe d'estimateurs de variance, mais plutôt de définir une méthode de réduire le biais de l'estimateur utilisé, quel qu'il soit.

Pour cela on cherche à déterminer un estimateur de variance modifié $\hat{V}_M$ d'un estimateur de variance traditionnel $\hat{V}$ ayant les propriétés suivantes:

1- le bias de $\hat{V}_M$ soit inférieur en valeur absolu à celui de $\hat{V}$,

2- $\hat{V}_M$ soit asymptotiquement égal à $\hat{V}$,

3- la forme de $\hat{V}_M$ soit simple, facile à calculer et à justifier.

Nous exposerons dans la section 2 la méthode utilisée afin d'obtenir l'estimateur de variance modifié de trois estimateurs de variance traditionnels de l'estimateur par le quotient. Nous présenterons à la section 3 quelques résultats de simulations illustrant le bien fondé de notre méthode de réduction du biais.

2. ESTIMATEUR PAR LE QUOTIENT

Considérons un plan d'échantillonnage aléatoire simple (SI) de n unités parmi N. Posons $f=n/N$ et définissons l'estimateur par le quotient de la moyenne de population $\bar{Y}$
par

$$\bar{y}_Q = \bar{X}\,\frac{\bar{x}}{\bar{y}}$$

où $\bar{X} = \sum_{i=1}^{N} \frac{x_i}{N}$, $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$ et $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$.

On considère l'estimation de l'erreur quadratique moyenne (EQM), $\mathrm{EQM}(\bar{y}_Q) = E(\bar{y}_Q - \bar{Y})^2$ où $E(\cdot)$ désigne l'espérance par rapport au plan SI. Considérons trois estimateurs de variance de $\bar{y}_Q$ couramment utilisés: l'estimateur obtenu par linéarisation de Taylor ($\hat{V}_T$), l'estimateur de Taylor avec correction de Royall ($\hat{V}_{T2}$) et l'estimateur jackknife ($\hat{V}_J$) définis respectivement par

$$\hat{V}_T = \frac{1-f}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{R}x_i)^2}{n-1},$$

$$\hat{V}_{T2} = \left(\frac{\bar{X}}{\bar{x}}\right)^2 \hat{V}_T ,$$

$$\hat{V}_J = (1-f)\,\frac{n-1}{n} \sum_{j=1}^{n} (\hat{R}_{(j)} - \hat{R}_{(\cdot)})^2$$

où $\hat{R} = \bar{y}/\bar{x}$, $\hat{R}_{(j)} = \sum_{i\neq j}^{n} y_i \,/\, \sum_{i\neq j}^{n} x_i$ et $\hat{R}_{(\cdot)} = \sum_{j=1}^{n} \hat{R}_{(j)}/n$

Soit $\hat{V}$ un de ces trois estimateurs de variance. Leblond (1989) démontre que, sous certaines conditions de régularité, le biais de $\hat{V}$ est alors

$$\begin{aligned}\mathrm{Biais}(\hat{V}) &= E\,\{\hat{V} - EQM(\bar{y}_Q)\} \\ &= B_2 + O(n^{-3})\end{aligned}$$

où $B_2 = O(n^{-2})$ est une expression complexe dépendant des y_k et où $O(\cdot)$ désigne que l'argument est borné. De plus Leblond (1989) démontre que pour $\hat{V}_T$, B_2 est toujours négatif et que pour $\hat{V}_J$, B_2 est toujours positif.

Nous proposons une méthode permettant de corriger $\hat{V}$ afin de réduire l'ordre du biais à un terme $O(n^{-3})$. Cette méthode consiste à obtenir un estimateur $\hat{B}_2$ et B_2 puis à définir l'estimateur de variance modifié $\hat{V}_M$ de la façon suivante:

$$\hat{V}_M = \hat{V} - \hat{B}_2 \tag{1}$$

Toutefois, de par sa forme complexe, B_2 est difficile à estimer. Afin de simplifier l'estimation de B_2, nous considérons une approche assistée d'un modèle. Le modèle utilisé ici est celui généralement considéré comme sous-jacent à l'estimateur par le quotient:

$$E_\xi(y_i) = \beta x_i,\ V_\xi(y_i) = \sigma^2 x_i \quad \text{et} \quad \text{cov}_\xi(y_i, y_j) = 0,\ i \neq j$$

Sous cette approche modéliste, nous considérons le biais anticipé de $\hat{V}$, soit

$$E_\xi[\text{Biais}(\hat{V})] = E_\xi(B_2) + O(n^{-3})$$

L'avantage d'utiliser la notion de biais anticipé réside dans la forme simple de $E_\xi(B_2)$. En effet $E_\xi(B_2) = K_x\sigma^2$ où $K_x = O(n^{-2})$ ne dépend que des valeurs de x (voir tableau 1).

Tableau 1: Biais asymptotique (jusqu'au terme $O(n^{-3})$) des estimateurs de variance.

Biais	$\hat{V}_T$	$\hat{V}_{T2}$	$\hat{V}_J$	Estimation
B_2	<0	?	>0	complexe
$E_\xi(B_2)$	$-2M_x$	$-M_x(1+f)$	$M_x(1-f)$	simple

où $M_x = \dfrac{(1-f)}{n(n-1)} \dfrac{S^2x}{\bar{X}} \sigma^2$.

Ainsi on obtient $\hat{V}_M$ en posant dans (1) $\hat{B}_2$ égal à l'estimateur de $E_\xi(B_2)$. Pour obtenir cet estimateur il suffit d'estimer σ^2. On exigera d'estimer σ^2 par $\hat{\sigma}^2$ de telle sorte que

$$\hat{\sigma}^2 = C_x\hat{V} \tag{2}$$

et

$$E_\xi(\hat{\sigma}^2) = \sigma^2 + O_p(n^{-1}), \tag{3}$$

où C_x est un terme ne dépendant que des valeurs de x et O_p désigne que l'argument est borné en probabilité par rapport au plan SI. La condition (2) permet d'obtenir un estimateur de variance modifié, $\hat{V}_M$, de forme simple puisqu'égal à un facteur multiplicatif x estimateur traditionnel. Afin de déterminer C_x pour chacun des trois estimateurs considérés, on peut démontrer le résultat suivant:

$$E_\xi \left[\left(\frac{\bar{X}}{\bar{x}} \right)^g \hat{V}_T \right] = \frac{\sigma^2}{C_{x(g)}} + O_p(n^{-2}),$$

où

$$C_{x(g)} = \frac{(n-1)}{\bar{x}(1-f)} \left(\frac{\bar{x}}{\bar{X}} \right)^g.$$

Il s'en suit que pour $\hat{V} = \left(\frac{\bar{X}}{\bar{x}} \right)^g \hat{V}_T$,, on obtient alors $C_x = C_{x(g)}$ permettant d'obtenir $\hat{\sigma}^2$ tel qu'exigé par (2) et vérifiant (3).

Le terme C_x est alors obtenu pour $\hat{V}_T$ et $\hat{V}_{T2}$ en posant respectivement $g=O$ et $g=2$ dans $C_{x(g)}$. La valeur de C_x pour $\hat{V}_J$ est la même que pour $\hat{V}_{T2}$ car

$$\hat{V}_{T2} = \hat{V}_J + O_p(n^{-2})$$

Posant $\hat{B}_2 = C_x K_x \hat{V}$ dans (1), alors on obtient

$$\hat{V}_M = \hat{V} - K_x C_x \hat{V} = A_x \hat{V},$$

où $A_x = 1 - K_x C_x$ est appelé facteur de correction. On a que A_x converge en probabilité vers 1. Ainsi asymptotiquement l'estimateur modifié converge vers sa version originale. La forme explicite des estimateurs de variance modifié des trois estimateurs ici considérés est alors obtenue en substituant les valeurs appropriées de K_x et C_x:

$$\hat{V}_{MT} = \left(1 + \frac{2}{n} (CV_x)^2 \frac{\bar{X}}{\bar{x}} \right) \hat{V}_T$$

$$\hat{V}_{MT2} = \left(1 + \frac{1+f}{n} (CV_x)^2 \frac{\bar{x}}{\bar{X}} \right) \hat{V}_{T2}$$

$$\hat{V}_{MJ} = \left(1 - \frac{1-f}{n} (CV_x)^2 \frac{\bar{x}}{\bar{X}} \right) \hat{V}_J$$

où $CV_x = S_x/\bar{X}$ est le coefficient de variation de x dans la population. Remarquons que pour $\hat{V}_{MT}$ le facteur de correction est supérieur à 1 et que pour $\hat{V}_{MJ}$ ce facteur est inférieur à 1, ce qui devrait permettre, conformément aux résultats présentés dans le tableau 1, une réduction dans le bon sens du biais.

De plus, dans le cas où $x_i = 1$ pour tout i, les trois estimateurs de variance modifiés sont alors égaux à

$$\hat{V}(\bar{y}) = \frac{1-f}{n} \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}.$$

Ainsi pour une telle situation l'estimateur de variance modifié est égal à l'estimateur sans biais habituellement utilisé. Ceci est un signe du bien fondé de la correction apportée.

Remarquons finalement que le biais d'ordre B_2 d'ordre n^{-2} de $\hat{V}$ n'est pas complètement éliminé par la correction apportée. En effet cette correction s'effectue en soustrayant à $\hat{V}$ l'estimateur asymptotiquement sans biais de $E_\xi(B_2)$ et B_2. Or pour une population quelconque $E_\xi(B_2)$ et B_2 ne sont pas nécessairement égaux (sauf si la population est totalement conforme au modèle ξ utilisé). Toutefois les résultats de simulation présentés dans la section suivante indiquent que dans la plupart des cas étudiés le facteur de correction permet une réduction substantielle du biais des estimateurs de variance traditionnels.

3. RÉSULTATS DE SIMULATIONS

Le but de l'étude par simulations était de comparer les estimateurs de variance modifiés et traditionnels présentés à la section 2. Le biais relatif et l'erreur quadratique moyenne (EQM) sous certains modèles de superpopulation ont servi de critères de comparaison. Pour ce faire, nous avons considéré des populations générées à partir du modèle général de superpopulation suivant:

$$E_\xi(y_i) = \alpha + \beta x_i + \gamma x_i^2,\ V_\xi(y_i) = \sigma^2 x_i^g \text{ et } \mathrm{cov}_\xi(y_i, y_j) = 0,\ i \neq j \qquad (4)$$

Chaque détermination du paramètre $\underline{\theta} = (\alpha, \beta, \gamma, \sigma^2, g)$ donne lieu à un modèle de superpopulation. Pour chacune des superpopulations considérées, 30 populations finies de taille $N = 100$ ont été générées à l'aide d'une fonction de densité gamma. L'utilisation d'une densité gamma est appropriée parce qu'elle est de distribution asymétrique tout comme la majorité des populations étudiées dans les enquêtes économiques. Pour chacune de ces 30 populations de chaque superpopulation nous avons généré 2 000 échantillons selon un tirage aléatoire simple sans remise de taille fixe, où pour chaque échantillon, l'estimateur par le quotient de même que divers estimateurs de variance étaient calculés. Nous avons considéré les tailles d'échantillon $n = 5, 10, 15, 20$.

Soient $\bar{y}_{Qrst}$ et $\hat{V}_{rst}$ respectivement l'estimateur par le quotient et un estimateur de variance calculés à partir du r^e échantillon de la s^e population de la t^e superpopulation. Soit également $\bar{Y}_{st}$ la moyenne de la s^e population de la t^e superpopulation.

L'EQM de l'estimateur par le quotient, le biais et l'EQM de $\hat{V}$ pour la s^e population de la t^e superpopulation sont respectivement définies par

$$\mathrm{EQM}_{st}(\bar{y}_Q) = \sum_{r=1}^{2000} \frac{(\bar{y}_{Qrst} - \bar{Y}_{st})^2}{2000},$$

$$\mathrm{Biais}_{st}(\hat{V}) = \sum_{r=1}^{2000} [\hat{V}_{rst} - \mathrm{EQM}_{st}(\bar{y}_Q)]/2000$$

$$\mathrm{Var}_{st}(\hat{V}) = \sum_{r=1}^{2000} \left[\hat{V}_{rst} - \left(\sum_{r=1}^{2000} \hat{V}_{rst}/2000\right)\right]^2 /1999.$$

Ensuite pour chaque superpopulation t on détermine le biais relatif et l'EQM de $\hat{V}$.

$$\text{Brel}_t = \sum_{s=1}^{30} \frac{\text{Biais}_{st}(\hat{V})/EQM_{st}(\bar{y}_Q)}{30}$$

$$\text{EQM}_t(\hat{V}) = \sum_{s=1}^{30} \frac{(\text{Biais}_{st}(\hat{V}))^2 + \text{Var}_{st}(\hat{V})}{30}$$

Ces dernières quantités représentent une approximation de l'espérance sous le modèle de superpopulation t du biais relatif et de l'EQM des estimateurs de variance considérés. Elles servent de base de comparaison entre les différents estimateurs.

Les populations générées à partir du modèle de superpopulation défini en (4) avec $\alpha=0$, $\gamma=0$ et $g=1$ ($\beta>0$, $\sigma^2>0$) correspondent à ce que nous appelons des populations favorables aux estimateurs de variance modifiés du fait que ces derniers sont obtenus à partir de ce même modèle. Ainsi pour de telles populations, on s'attend à ce que la correction apportée aux estimateurs de variance élimine le biais B_2 d'ordre n^{-2}. Par opposition, nous appelons populations non favorables toutes populations issues de (4) avec au moins une des valeurs de α, γ, g dans $\underline{\theta}$ différentes de celles utilisées pour générer les populations favorables (voir tableau 2). Ces populations non favorables permettent de vérifier la robustesse des estimateurs de variance modifiés par rapport à la réduction du biais. Les modèles de superpopulation utilisés, au nombre de six, pour générer les populations non favorables sont les mêmes que ceux utilisés par Valliant (1990).

Tableau 2: Paramètres utilisés pour générer les populations non favorables

Modèle	α	β	γ	σ^2	g
1	100	1.5	0.	0.25	1.5
2	100	1.5	0.	0.25	2.0
3	100	1.8	-0.0008	0.25	1.5
4	100	1.8	-0.0008	0.25	2.0
5	100	-0.3	0.0009	0.25	1.5
6	100	-0.3	0.0009	0.25	2.0

L'étude de la performance des différents estimateurs de variance consistait à comparer pour chaque type d'estimateur (Taylor, Royall et jackknife) donné, le biais relatif de $\hat{V}$ et $\hat{V}_M$ tant pour des populations favorables que non favorables. Il n'était pas question de comparer les types d'estimateurs entre eux: les études de Wu et Deng (1983) ont déjà montré qu'il est difficile d'identifier le "meilleur" estimateur de variance, la performance de ceux-ci dépendant du modèle utilisé. Aussi nous voulons plutôt par les études par simulation répondre à la question suivante: Est-ce que la modification d'un estimateur de variance permet une réduction sensible du biais et ce même pour une population non conforme au modèle sous-jacent à cette modification? Les résultats obtenus nous permettent de répondre par l'affirmative à cette question.

Les résultats de simulation obtenus pour les populations favorables sont conformes à la théorie présentée à la section 2 (voir figure 1, 2, 3). Les trois estimateurs de variance modifiés possèdent un biais d'ordre inférieur à leur version traditionnelle. Ceci se vérifie par le fait que le biais de $\hat{V}$ est supérieur en valeur absolue à celui de $\hat{V}_M$ et que le biais de $\hat{V}_M$ tend plus rapidement vers 0 que celui de $\hat{V}$ lorsque η augmente. De plus les trois estimateurs de variances modifiés possèdent un biais relatif semblable, à peu près égal à 0, dû à l'élimination du terme d'ordre $O(n^{-2})$. La différence entre le biais relatif des estimateurs traditionnels est conforme aux résultats présentés dans le Tableau 1 concernant $E_\xi(B_2)$: le biais de $\hat{V}_T$ est, en valeur absolue, environ deux fois plus grand que celui de $\hat{V}_{T2}$ et $\hat{V}_J$.

Pour les populations non favorables, la correction permet une réduction significative du biais uniquement pour $\hat{V}_T$ et $\hat{V}_J$ (voir figure 1 et 3). L'estimateur $\hat{V}_{T2}$ ne partage pas de façon aussi marquée cette propriété de robustesse: pour plusieurs populations non favorables la correction a peu d'impact (voir figure 2) et, pour certains modèles

de superpopulation différents de ceux décrits par le tableau 2, elle a même tendance à faire augmenter la valeur absolue du biais. Ceci peut s'expliquer par le fait que la forme même de cet estimateur est au départ liée au modèle de superpopulation favorable. En effet, $\hat{V}_{T2}$ est approximativement sans biais sous ce modèle. Le cas non favorable des figures 1 à 3 réfère aux résultats obtenus pour les populations générées selon le modèle 4 et réflète la tendance générale des résultats obtenus pour les autres modèles.

Figure 1: Biais relatif (%) de l'estimateur $\hat{V}_T$ et de sa modification.

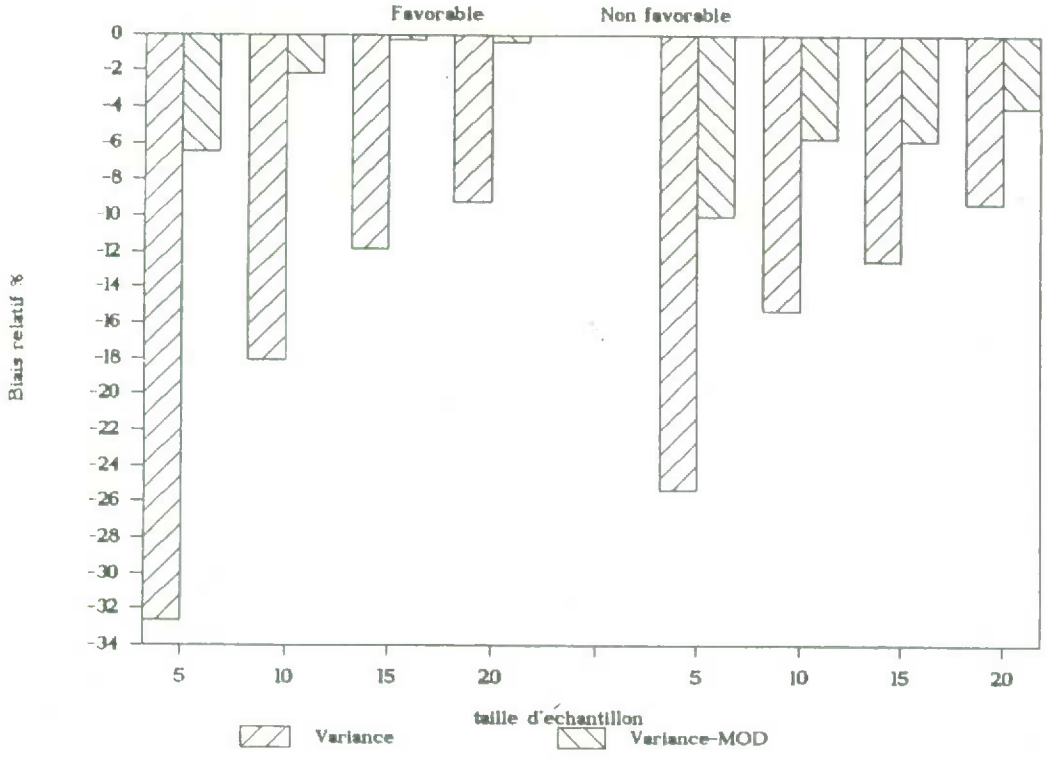


Figure 2: Biais relatif (%) de l'estimateur $\hat{V}_{T2}$ et de sa modification.

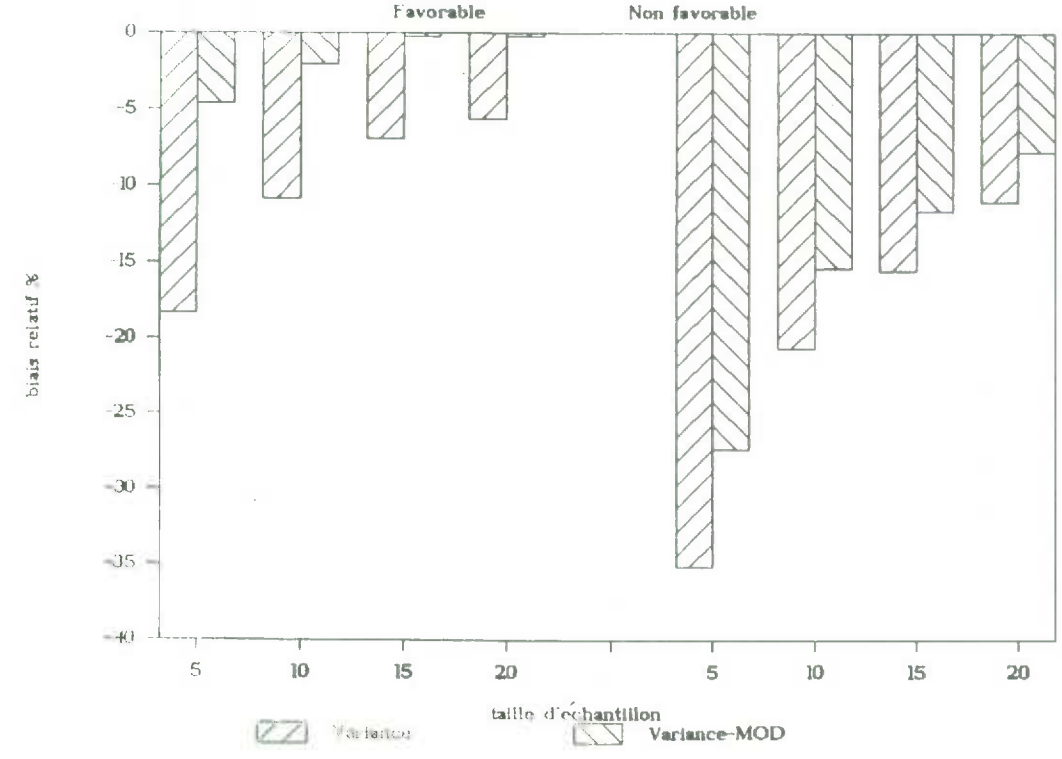
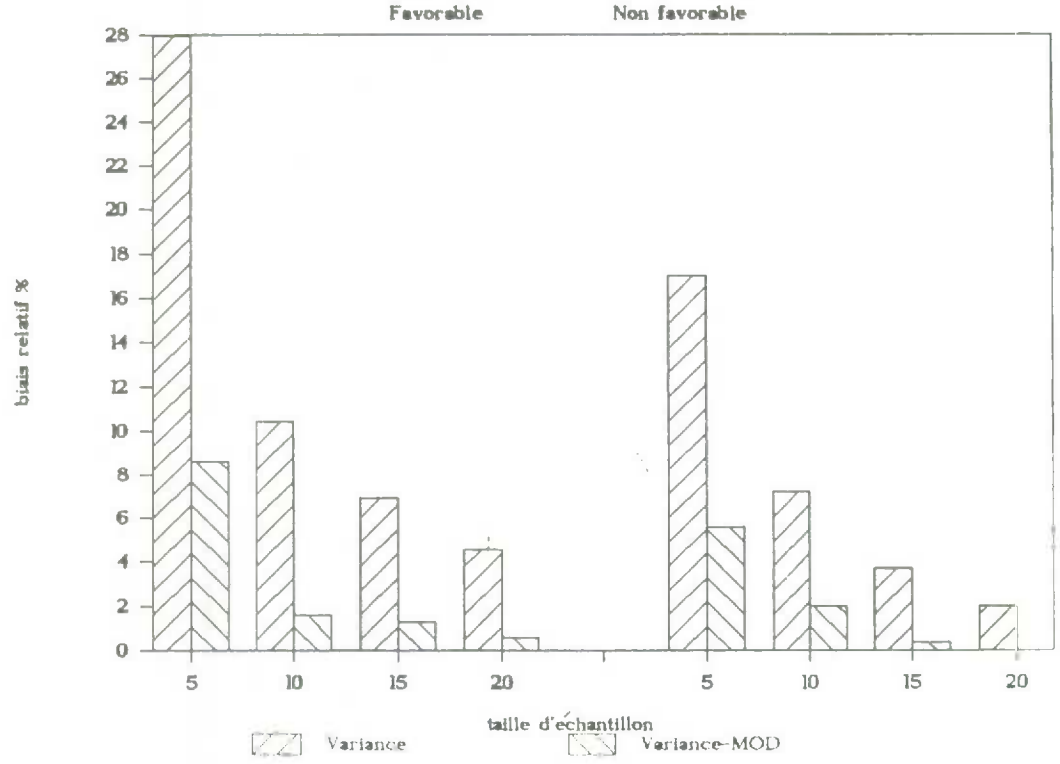


Figure 3: Biais relatif (%) de l'estimateur $\hat{V}_J$ et de sa modification



Nous avons également comparé par simulation l'EQM des estimateurs de variance pour les six superpopulations considérées (voir tableau 3). Les estimateurs de variance modifiés $\hat{V}_{MT}$ et $\hat{V}_{MT2}$ possèdent, règle générale, une EQM supérieure à leur version traditionnelle. Pour ces deux estimateurs l'augmentation de leur EQM est le prix à payer pour obtenir un biais relatif moindre. Ce "prix" semble toutefois être plus élevé pour $\hat{V}_{MT}$. Il en va autrement pour l'estimateur jackknife: $\hat{V}_{MJ}$ possède une EQM inférieure à celle de $\hat{V}_{MT}$. Ces constatations sont valides pour la grande majorité des populations générées bien que l'on présente au tableau 3 uniquement les résultats obtenus pour le modèle 5.

Tableau 3: **Augmentation (diminution) de l'EQM dûe à la modification des estimateurs de variance traditionnels sous le modèle 5**

$$(\Delta\ \mathrm{EQM}(\hat{V}_M) = 100\left(\frac{\mathrm{EQM}(\hat{V}_M) - \mathrm{EQM}(\hat{V})}{\mathrm{EQM}(\hat{V})}\right).$$

n	Δ EQM $(\hat{V}_{MT})$	Δ EQM $(\hat{V}_{MT2})$	Δ EQM $(\hat{V}_{MJ})$
5	28.2	7.3	-5.0
10	15.7	6.3	-9.4
15	8.9	3.5	-6.1
20	6.8	3.0	-4.0

En résumé, comme estimateur de variance fiable et robuste (c'est à dire avec biais relativement peu élevé, EQM faible et ce, le plus indépendamment possible de la distribution de la population considérée) nous

recommandons, à la lumière des résultats obtenus, l'utilisation de $\hat{V}_{MT}$. D'autre part, si on désire utiliser un estimateur jackknife, alors on a tout à gagner à utiliser $\hat{V}_{MJ}$, car celui-ci possède à la fois un biais et une EQM inférieur à $\hat{V}_J$.

4. CONCLUSION

Nous avons présenté une méthode visant à réduire le biais des estimateurs de variance traditionnellement utilisés pour estimer la variance de l'estimateur par le quotient. Cette méthode consiste à multiplier ces estimateurs de variance par un facteur de correction permettant ainsi une réduction du biais. Les résultats de simulation démontrent que cette correction peut s'avérer efficace et robuste lorsqu'appliquée aux estimateurs de variance de Taylor et jackknife. Pour ce dernier estimateur, la correction a pour effet de diminuer son EQM.

Il serait intéressant de généraliser la méthode présentée ici l'estimation de la variance d'autres types d'estimateur de régression: combiné, quotient séparé, régression généralisé, etc..

BIBLIOGRAPHIE

Deng. L.Y., et Wu, C.F.J. (1987). Estimation of variance of the regression estimator, *Journal of the American Statistical Association*, 82, 568-576.

Leblond, Y. (1989). *Contributions à la théorie d'estimation pour sous-populations*, Thèse de doctorat, Université de Montréal.

Royall, R.M., et Cumberland, W.G. (1978). Variance estimation in finite population sampling, *Journal of the American Statistical Association*, 73, 351-358.

Royall, R.M., et Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance, *Journal of the American Statistical Association*, 76, 66-67.

Royall, R.M., et Cumberland, W.G. (1981b). The finite population linear regression estimator and estimators of its variance: an empirical study, *Journal of the American Statistical Association*, 76, 98-102.

Valliant, R. (1987). Conditionnal properties of some estimators in stratified sampling, *Journal of the American Statistical Association*, 82, 509-519.

Valliant. R. (1990). Comparison of variance estimators in stratified random and systematic sampling, *Journal of Official Statistics*, 6, 115-131.

Wu, C.F.J. (1982). Estimation of variance of ratio estimator, *Biometrika*, 69, 183-189.

Wu, C.F.J. (1985). Variance estimation of the combined ratio and combined regression estimators, *Journal of the Royal Statistical Society*, Ser. B, 47, 147-157.

Wu, C.F.J., et Deng, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. Dans *Scientific Inference, Data Analysis and Robustness*, Ed. G.E.P. Box *et al.*, 245-277. New York: Academic Press.

SESSION 5

L'amélioration de la collecte des données

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

AMÉLIORATION DE LA QUALITÉ DES ESTIMATIONS ASSUJETTIES À DES CONTRAINTES DE TEMPS À L'AIDE D'UN MODE DE COLLECTE MIXTE ITAO/AIAO

G.S. Werking et R.L. Clayton[1]

RÉSUMÉ

L'établissement d'estimations assujetties à des contraintes de temps pose un problème constant, qui est celui des limitations importantes que comportent les méthodes de collecte des données par la poste. Pour surmonter cette difficulté, le Bureau of Labor Statistics a effectué pendant 7 ans d'importants travaux de recherche sur deux méthodes d'introduction des données par reconnaissance de la parole et au moyen d'un clavier à boutons : l'interview téléphonique assistée par ordinateur (ITAO) et l'auto-interview assistée par ordinateur (AIAO). La présente communication donne un aperçu de certains des principaux résultats de ces travaux ayant trait au rendement et au coût de ces méthodes. L'exposé s'achève par des observations sur un programme d'application à grande échelle de ces techniques à un échantillon mensuel de 350,000 entreprises.

MOTS CLÉS: Statistiques sur l'emploi; révisions; collecte des données au moyen d'un clavier à boutons; collecte des données par reconnaissance de la parole; analyse des coûts.

1. INTRODUCTION

1.1 Les statistiques sur l'emploi aux États-Unis

Le premier vendredi de chaque mois, le Bureau of Labor Statistics (BLS) diffuse des données sur la situation de l'emploi aux États-Unis pour le mois précédent. Le jour de la publication des données, le Commissioner of Labor Statistics se présente devant le Joint Economic Committee du Congrès pour faire une analyse détaillée des données et des tendances relatives au mois en cours; au même moment, les données sont mises à la disposition des médias d'information ainsi que de la communauté financière et du monde des affaires. Cet ensemble de statistiques, qui est suivi de près par les observateurs, constitue le premier indicateur de l'activité économique pour le mois précédent et est utilisé comme un des principaux étalons de mesure de la santé de l'économie américaine. Les statistiques publiées comprennent des données sur l'emploi, les gains et les heures par industrie, recueillies au moyen de l'enquête mensuelle réalisée par le Bureau auprès de 350,000 entreprises - la Current Employment Statistics Survey (CESS) - ainsi que des données sur la population active et le chômage, recueillies au moyen d'une enquête réalisée par le Bureau auprès d'un échantillon de 60,000 ménages - la Current Population Survey (CPS).

Les données de l'enquête sur les entreprises trouvent nombre d'utilisations importantes sur le plan économique. Étant donné l'envergure et l'actualité de la CESS ainsi que l'importance des statistiques de base sur la paie qu'elle permet de recueillir, les estimations mensuelles de l'enquête sont non seulement utilisées comme indicateurs économiques principaux, mais elles servent également à l'élaboration de nombreux indicateurs importants de l'activité économique du pays : le revenu personnel pour le calcul du produit national brut, l'indice des indicateurs avancés d'activité, l'indice des indicateurs instantanés d'activité, l'indice de la production

---

[1] G.S. Werking et R.L. Clayton, Monthly Industry Employment Statistics Division, Bureau of Labor Statistics, Washington, D.C., USA, 20212.

industrielle, certaines mesures des gains réels et certaines mesures de la productivité. Bien que l'actualité et l'exactitude des statistiques de la CESS soient des caractéristiques essentielles pour l'analyse des conditions économiques qui prévalent aux États-Unis, les données sont recueillies par la poste depuis la mise en oeuvre initiale de l'enquête au début du XX^{e} siècle. L'utilisation de cette méthode de collecte se traduit par la publication initiale, pour les agrégats de niveau élevé, d'estimations «provisoires» portant seulement sur un échantillon des questionnaires reçus, puis par la diffusion, deux mois plus tard, d'estimations «définitives» portant sur l'ensemble de l'échantillon. L'établissement d'estimations provisoires et définitives pour un mois donné nécessite qu'on procède périodiquement à une révision substantielle des estimations mensuelles. Ces révisions ont une incidence non seulement sur les statistiques de base de la CESS, mais aussi sur les autres statistiques utilisant les estimations de la CESS comme intrant. Ce sont ces révisions des estimations mensuelles de la CESS qui ont amené le Bureau à mettre en oeuvre un programme de recherche sur les techniques de collecte téléphonique automatisée des données.

La présente communication donne un aperçu des travaux de recherche effectués par le Bureau pendant 7 ans sur les techniques de collecte téléphonique des données et résume certains des principaux résultats de ces travaux. Vous trouverez aux sections suivantes une description du processus de mise en oeuvre de la CESS, une analyse du programme de recherche évaluant les méthodes de collecte des données par interview téléphonique assistée par ordinateur (ITAO), par introduction des données au moyen d'un clavier à boutons (IDCB) et par reconnaissance de la parole (RP), un exposé détaillé de certains des principaux résultats des travaux de recherche ayant trait au rendement et au coût de ces méthodes, enfin, pour conclure, une analyse du programme d'application à grande échelle de ces techniques dans le cadre de la CESS.

1.2 Current Employment Statistics Survey

Avec son échantillon de 350,000 unités, la CESS est la plus importante enquête-échantillon mensuelle effectuée aux États-Unis. Cette enquête est réalisée par le Bureau dans le cadre d'un programme de coopération entre le fédéral et les États. Aux termes de ce programme, il revient au Bureau de préciser le plan de sondage et les procédures opérationnelles de l'enquête, tandis qu'il incombe aux États de mettre en oeuvre les activités de collecte des données ainsi que les activités de contrôle et de conciliation. Le Bureau produit et publie, pour l'ensemble du pays, des estimations mensuelles complètes pour toutes les industries dont le code comporte 2, 3 ou 4 chiffres, tandis que les États produisent des estimations mensuelles à l'échelle de l'État et de la région (270 régions statistiques métropolitaines).

On s'entend pour reconnaître que les estimations de la CESS constituent des statistiques économiques très précises. Une fois l'an, on obtient à partir des dossiers d'impôt de l'assurance-chômage les chiffres de l'emploi pour l'ensemble de l'univers pour l'année précédente et on utilise ces chiffres pour procéder à un étalonnage (réalignement) annuel des estimations de la CESS en fonction des chiffres obtenus pour l'ensemble de l'univers. Cet étalonnage annuel permet d'obtenir une estimation plus précise pour le mois en cours ainsi qu'une estimation de l'erreur globale affectant l'enquête. Au cours des cinq dernières années, l'écart moyen entre l'estimation définitive fondée sur l'échantillon de la CESS et le chiffre obtenu pour l'ensemble de l'univers a été inférieur à .2%, et il est resté pratiquement nul pendant 4 ans au cours des années quatre-vingt. Alors qu'on considère que les estimations mensuelles définitives de la CESS correspondent très étroitement aux chiffres obtenus pour l'ensemble de l'univers, les estimations mensuelles provisoires, fondées sur environ 50% des questionnaires de l'échantillon reçus par la poste, ont fait périodiquement l'objet d'importantes révisions après avoir été comparées aux estimations définitives publiées deux mois plus tard. Bien qu'on ait pu, au fil des ans, apporter des améliorations permettant de réduire l'ampleur des révisions mensuelles, on considère l'exécution d'importantes révisions périodiques comme un sous-produit de la mise en oeuvre d'une enquête décentralisée de grande envergure utilisant les méthodes de collecte de données par la poste.

Un certain nombre de changements survenus au cours de la dernière décennie devaient avoir une incidence considérable sur le programme de la CESS, en influant tant sur l'urgence de résoudre la question des révisions mensuelles que sur les options susceptibles d'être choisies à cet effet. Au cours des années quatre-vingt, les utilisateurs sont devenus beaucoup plus conscients de l'importance de la qualité des données et, sans que la qualité des produits de la CESS ne se soit nécessairement détériorée, leurs attentes sur le plan de la qualité et de l'utilisabilité des données se sont considérablement élevées. L'émergence de cette nouvelle façon de voir est dans une large mesure attribuable aux travaux de Deming, Juran et d'autres auteurs portant sur la gestion de

la qualité. Les années quatre-vingt ont également fait ressortir l'importance de l'utilisation des statistiques sur la paie de la CESS pour l'établissement d'un diagnostic actuel de l'état de santé de l'économie américaine; toutefois, ce gain de visibilité des statistiques de la CESS s'est accompagné d'un accroissement correspondant de la frustration manifestée par les utilisateurs à l'égard des révisions mensuelles. Nous avons également assisté au cours des années quatre-vingt à de spectaculaires percées technologiques et, en particulier, à l'avènement de l'ère des micro-ordinateurs. Cette nouvelle technologie a offert aux organismes statistiques de nombreuses occasions d'améliorer le contrôle et la qualité de la collecte des données, notamment : l'ITAO, l'introduction des données au moyen du clavier à boutons, la reconnaissance par la parole, l'auto-interview assistée par ordinateur (AIAO) et le télécopieur. Plusieurs de ces techniques allaient ultimement permettre d'améliorer de façon sensible l'actualité et la qualité des données tout en maintenant les coûts de mise en oeuvre au même niveau, voire même en les réduisant.

Après avoir réalisé, au cours des années quatre-vingt, des travaux de recherche expérimentale dans le cadre de la CESS et avoir soumis à des essais réels certaines des techniques les plus évoluées de collecte automatisée des données, le Bureau prévoit procéder à une application à grande échelle de ces techniques en 1991.

2. PROGRAMME DE RECHERCHE SUR LA CESS

2.1 Objectifs des travaux de recherche

Au début des années quatre-vingt, le Bureau a effectué pendant sept ans d'importants travaux de recherche sur les causes des réponses tardives et sur les autres méthodes de collecte susceptibles de permettre un accroissement important des taux de réponse en vue de l'établissement des estimations provisoires. Ces travaux visaient à obtenir des réponses aux trois questions de base suivantes :

- L'entreprise dispose-t-elle des données à temps pour pouvoir les transmettre avant la date limite de diffusion des estimations préliminaires?

- Existe-t-il des méthodes de collecte des données pouvant permettre l'obtention d'un taux de réponse de 80 à 90% en respectant des contraintes de temps aussi serrées?

- Est-il possible de maintenir le coût de mise en oeuvre de ces méthodes de collecte des données à un niveau à peu près équivalent au coût de mise en oeuvre des actuelles méthodes de collecte par la poste?

Le programme de recherche a permis de mettre au point une méthode de collecte mixte ITAO/IDCB assurant l'obtention des taux de réponse désirés et le respect des contraintes financières imposées. On trouve aux sections suivantes une brève description de ces méthodes de collecte assistée par ordinateur personnel (OP) et des essais réalisés au cours des travaux ainsi qu'un exposé des taux de réponse obtenus et une analyse des coûts. Vous trouverez de plus amples renseignements sur ces essais dans les mémoires de recherche dont il est fait état dans la bibliographie.

2.2 Méthodes de collecte des données

Compte tenu de la date de diffusion des estimations provisoires, les délais de collecte des données de la CESS sont très serrés. Comme la période de référence de la CESS est la période de paie englobant le 12e jour du mois, on ne dispose que de 2½ semaines pour recueillir les données, perforer et contrôler les questionnaires, puis totaliser, valider et publier les données. Afin de respecter ces contraintes de temps, il faut disposer d'une méthode de collecte permettant d'obtenir les données requises dès que l'entreprise en dispose. Les principales méthodes de collecte étudiées sont décrites ci-après.

Poste - Le questionnaire de la CESS est un questionnaire d'une page prévoyant l'espace nécessaire pour permettre à l'employeur d'inscrire les données relatives à 12 mois. Après avoir reçu le questionnaire par la poste vers le 12e jour du mois (c.-à-d. la date de référence de l'enquête), l'employeur y inscrit les éléments d'information requis sur la ligne correspondant au mois en cours. Les cinq éléments d'information de base recueillis sont les suivants : nombre total d'employés, nombre d'employés de sexe féminin, nombre d'employés

de production (ou de membres non cadres), heures et gains. Une fois le questionnaire rempli, l'employeur le renvoie par la poste à l'organisme d'État, où il est perforé à l'aide d'un clavier puis mis de côté en vue de son envoi par la poste le mois suivant. Comme nous l'avons mentionné ci-devant, ce processus permet actuellement d'obtenir un taux de réponse de 50% au cours des 2½ semaines dont on dispose avant la diffusion des estimations provisoires.

Interview téléphonique assistée par ordinateur - Suivant la méthode ITAO, l'employeur reçoit le questionnaire de la CESS par la poste au début de l'année et le conserve pour y inscrire les données mensuelles. Chaque mois, lorsqu'il dispose des données sur la paie, l'employeur inscrit sur le questionnaire les éléments d'information relatifs à ce mois et attend l'appel ITAO de l'organisme d'État. Lorsque l'organisme téléphone, les données sont recueillies par ITAO et contrôlées, puis on fixe la date et l'heure de l'appel en vue de la collecte des données du mois suivant.

Introduction des données au moyen du clavier à boutons - Suivant la méthode IDCB, l'employeur suit les mêmes étapes que pour la méthode ITAO, mais, plutôt que d'attendre l'appel ITAO de l'organisme d'État, il compose lui-même un numéro 800 raccordé à l'OP à clavier de cet organisme, puis se sert du clavier à boutons de son poste pour introduire les éléments d'information à la suite des messages de guidage appropriés prévus dans l'interview automatisée. Après avoir été introduit, chaque élément d'information est relu pour permettre au répondant de le vérifier.

Reconnaissance de la parole - La déclaration des données par reconnaissance de la parole et la collecte au moyen du clavier à boutons s'effectuent de la même façon, à la différence près que la première méthode ne nécessite pas l'utilisation d'un clavier à boutons. Il suffit en effet à l'employeur de lire les données inscrites sur la formule pour que l'OP à entrée vocale les traduise et les lui relise pour lui permettre de les vérifier. Le système RV est un système multilocuteur qui permet la reconnaissance de la parole continue et peut reconnaître les chiffres de 0 à 9 ainsi que les termes «yes» et «no».

2.3 Essais réalisés dans le cadre du programme de recherche

En 1983, le Bureau a entrepris l'élaboration d'un système ITAO articulé sur OP destiné à être utilisé dans le cadre d'un essai portant sur deux États, dont la mise en oeuvre était prévue pour 1984 (figure 1 - toutes les figures sont en appendice). C'est le système ITAO élaboré par l'Université de la Californie à Berkeley qui a été retenu aux fins de cet essai et utilisé subséquemment pendant toute la durée des travaux de recherche. Après avoir initialement sélectionné un échantillon aléatoire de 200 unités dans chaque État, le Bureau a graduellement affiné les procédures et les systèmes de collecte au cours des 7 années suivantes. Les essais ayant permis d'obtenir des taux de réponse très élevés, ils ont été étendus à 9 États en 1986, puis à un total de 14 États en 1988. La composition de l'échantillon d'essai a également été modifiée en 1986. Plutôt que de porter sur des échantillons aléatoires sélectionnés au sein de l'échantillon global de la CESS, les essais subséquents ont utilisé uniquement des échantillons aléatoires d'entreprises donnant habituellement une réponse tardive (c.-à-d. d'unités affichant un taux de réponse inférieur à 20% au moment de la date limite de diffusion des estimations provisoires). Ainsi, le critère choisi pour évaluer l'efficacité des nouvelles méthodes de collecte par ITAO et par IDCB a été leur capacité de faire passer à un taux permanent de 80 à 90% le taux de réponse des échantillons d'unités ayant affiché un taux de réponse de 0 à 20% au moment de la diffusion des estimations provisoires. En 1990, à la conclusion des travaux de recherche sur l'ITAO, le Bureau utilisait la méthode de l'ITAO pour recueillir des données auprès de 5,000 unités chaque mois et avait réalisé au total plus d'un quart de million d'interviews téléphoniques assistées par ordinateur.

Tandis que l'ITAO se révélait très efficace pour accroître les taux de réponse, il est également devenu évident, en 1985, que la collecte par ITAO serait plus coûteuse que l'ancienne méthode de collecte par la poste. Dès cet instant, de nouveaux travaux ont été entrepris pour chercher à réduire le coût de mise en oeuvre de la méthode ITAO tout en maintenant le taux de réponse mensuel au niveau élevé atteint jusqu'alors. Bien qu'on ait par la suite réussi à réduire la durée de la période nécessaire pour réaliser une interview téléphonique assistée par ordinateur, c'est l'utilisation d'une nouvelle méthode de déclaration téléphonique articulée sur OP qui devait permettre d'assurer une réduction spectaculaire des coûts de collecte par ITAO.

En 1985, de nombreuses banques américaines utilisaient, aux fins de l'encaissement des chèques aux guichets-autos, un système de vérification par introduction des données au moyen d'un clavier à boutons. Après avoir choisi un système de déclaration au moyen d'un clavier à boutons articulé sur OP se prêtant à la réalisation d'essais dans le cadre d'une enquête, le Bureau a effectué en 1986 un essai de collecte de données à l'aide de cette technique dans deux États. Il convient de noter que l'IDCB n'était alors considérée ni comme une technique destinée à remplacer directement la collecte par la poste ni comme une solution de rechange pour l'ITAO. L'ITAO avait pour objet, par le biais d'un contact personnel et d'un processus de sensibilisation, d'amener les entreprises donnant d'ordinaire une réponse tardive à donner leur réponse dans les délais prévus, tandis que l'EDCB devait être utilisé auprès des entreprises répondant dans les délais prévus afin de maintenir leur taux de réponse au même niveau élevé tout en permettant une réduction substantielle du coût de collecte par unité. Au cours des 5 années où elle a été utilisée, l'IDCB s'est également révélée être une méthode de collecte téléphonique des données très efficace et très fiable. Au moment où les travaux de recherche sur l'IDCB en arrivent également à leur conclusion, plus de 5,000 unités réparties entre 14 États déclarent chaque mois leurs données au moyen de cette technique et le Bureau a recueilli au total plus de 100,000 questionnaires à l'aide de cette nouvelle méthode de déclaration automatisée.

Dans la foulée des travaux sur l'IDCB, le Bureau effectue actuellement plusieurs essais à petite échelle à l'aide d'un nouveau système de déclaration par reconnaissance de la parole. Les résultats préliminaires de ces essais indiquent que le nouveau système permet d'obtenir des taux de réponse mensuels aussi élevés que l'IDCB, mais que la déclaration par RP a pour avantage de paraître plus naturelle aux répondants et que ces derniers la préfèrent à l'IDCB. Actuellement, le coût d'acquisition du matériel de RP est environ 15 fois supérieur à celui du matériel d'IDCB; toutefois, d'ici quelques années, ce coût diminuera et la déclaration par RP constituera une solution de rechange rentable pour l'IDCB.

2.4 Résultats des travaux de recherche

Au cours des 7 dernières années, le Bureau a été en mesure de déterminer que les données sur la paie sont accessibles dans la plupart des entreprises avant la date limite de diffusion des estimations provisoires et que la méthode de collecte par ITAO permet, à l'intérieur d'un délai de 6 mois, d'amener les entreprises ayant traditionnellement donné des réponses tardives (c.-à-d. affichant un taux de réponse de 0 à 20% pour les estimations provisoires) à transmettre leurs réponses dans les délais prévus et à afficher un taux de réponse de 82 à 84% (figure 1). Les taux de réponse obtenus sont demeurés remarquablement stables au fil des ans, au fur et à mesure que la taille de l'échantillon ITAO s'élargissait, pour passer de 400 à 5,000 unités, et que le nombre d'États participants s'accroissait, pour passer de 2 à 14. Selon les résultats de la recherche, la majorité des entreprises disposent des données à temps pour respecter la date limite de diffusion, et l'utilisation de la méthode ITAO permet d'augmenter dans une mesure de 60 à 80% le taux de réponse des entreprises donnant une réponse tardive, ainsi que de maintenir ce taux à l'intérieur de la plage visée de 80 à 90% sur de longues périodes de temps. On a découvert que le principal facteur limitant la capacité du répondant à respecter la date limite de diffusion était la longueur de la période de paie de l'entreprise (figure 3). Cette période est généralement d'une semaine, de deux semaines, de quinze jours ou d'un mois. Les données sur les périodes de paie hebdomadaires et bimensuelles peuvent presque toujours être recueillies à temps pour être diffusées, tandis que les données sur les périodes de paie à la quatorzaine sont la plupart du temps accessibles dans les délais fixés; toutefois, la majorité des périodes de paie mensuelles se terminent longtemps après la date limite de diffusion. L'existence de périodes de paie mensuelles a été un des principaux facteurs limitant les taux de réponse obtenus à l'aide de la méthode ITAO à une plage maximale de 82 à 84%.

Les travaux de recherche sur l'ITAO ont permis d'obtenir plusieurs autres résultats importants. Lorsqu'on utilise cette méthode, environ 60% des répondants sont prêts à transmettre leurs données à la date fixée pour le premier appel, tandis que les autres 40% utilisent le premier appel comme un appel de sollicitation. Ce taux est demeuré relativement stable tant d'un État à l'autre que d'une année d'essai à l'autre. On prévoit réaliser un essai à petite échelle afin de déterminer s'il est possible de réduire de façon significative le nombre de rappels nécessaires en faisant parvenir une carte postale de préavis au répondant, quelques jours avant la date fixée pour l'ITAO.

La durée moyenne de la période nécessaire pour réaliser une interview téléphonique assistée par ordinateur est fonction du nombre d'éléments d'information à recueillir, du rendement temporel de l'appareil utilisé et de

l'expérience de la personne recueillant les données. Ainsi, on a pu réduire d'un tiers la durée moyenne des appels ITAO (figure 1) en modernisant le matériel utilisé et en donnant une meilleure formation aux intervieweurs. Une autre question très importante prise en considération au cours des essais a été l'incidence de la méthode ITAO sur la perte d'effectifs de l'échantillon. On craignait que les employeurs cessent de participer au programme pour éviter d'être constamment ennuyés par des appels téléphoniques. Toutefois, il s'est avéré que le taux de perte d'effectifs de l'échantillon était environ trois fois moins élevé avec la méthode ITAO qu'avec la méthode de collecte par la poste, et que la première méthode permettait de réduire presque à zéro le nombre de répondants d'importance sortant de l'échantillon. Bref, il semble que l'ITAO a reçu un accueil favorable auprès de la majeure partie des répondants et a permis d'optimaliser le taux de réponse pouvant être obtenu pour les estimations provisoires.

Par suite de l'accroissement des coûts suscité par l'utilisation de la méthode ITAO, le Bureau a entrepris des travaux de recherche sur la collecte des données au moyen du clavier à boutons. Pendant les 4 années qu'ont duré les essais, il s'est révélé possible d'amener les répondants rapides affichant un taux de réponse de 82 à 84% avec la méthode ITAO à afficher des taux de réponse aussi élevés en ayant recours à une méthode de déclaration par IDCB complètement automatisée (figure 1). L'importance de ce résultat s'explique par les économies que l'IDCB permet de réaliser par rapport à l'ITAO. Un des principaux sujets de préoccupation relatifs à la collecte par IDCB était que, contrairement au phénomène observé pour la collecte par ITAO, où les appels s'échelonnent sur toute la journée, les répondants n'aient tendance à concentrer leurs appels au cours de la même période, entraînant du même coup la transmission de signaux d'occupation et nécessitant l'utilisation d'un nombre excessif d'OP à clavier pour traiter les appels reçus en période de pointe. Heureusement, ces inquiétudes se sont avérées sans fondement et, bien que les OP à clavier soient en service 24 heures par jour, la majeure partie des appels sont répartis assez uniformément entre 8 h et 17 h (figure 4). Par ailleurs, le pourcentage de répondants nécessitant un appel de sollicitation a tendance à s'établir au même niveau (environ 40%), qu'on utilise la méthode IDCB ou la méthode ITAO. On met actuellement à l'essai des méthodes visant à réduire le nombre d'appels de sollicitation nécessités par la méthode IDCB. Un des principaux avantages que présente cette méthode pour le répondant est qu'elle nécessite une durée d'interview deux fois moins longue que la méthode ITAO, la durée moyenne d'interview IDCB s'établissant à seulement 1 minute 45 secondes. En outre, la plupart des entreprises sont déjà pourvues de postes à clavier et les estimations courantes indiquent que plus de 80% des employeurs sont en mesure de déclarer les données au moyen de la méthode IDCB. Bien que la méthode IDCB offre de nombreux avantages à l'organisme statistique, sa caractéristique la plus attrayante est la faveur dont elle jouit auprès des répondants. Sa rapidité et sa commodité font que ces derniers la préfèrent à la collecte des données par la poste et à la méthode ITAO.

Mentionnons à titre d'observation générale relative à l'élaboration d'un programme de recherche sur l'ITAO que la nature du matériel ou du logiciel utilisé au cours des travaux de recherche n'est pas particulièrement importante, pour autant que ce matériel ou ce logiciel soit assez souple pour être modifié. Il est possible que les résultats finaux des essais indiquent que la mise en oeuvre du système ITAO pose des exigences très différentes de celles posées initialement pour le programme de recherche. À cet égard, l'activité la plus importante et celle à laquelle il faut consacrer le plus de temps est l'élaboration et l'affinement des méthodes et procédures de communication avec les répondants. Une fois qu'on a élaboré des méthodes et procédures efficaces, les exigences posées par la mise en oeuvre du «bon» système deviennent beaucoup plus claires.

2.5 Analyse des coûts

Après que les essais eurent démontré que tant la méthode ITAO que la méthode IDCB offraient un rendement élevé et étaient bien acceptées par les répondants, la dernière phase des travaux de recherche a porté sur une analyse des coûts transitoires afférents à la méthode ITAO et des coûts courants de mise en oeuvre de la méthode IDCB.

À cette fin, nous avons étudié les principales catégories de dépenses «salariales» et «non salariales» relatives aux méthodes de collecte par la poste, par ITAO et par IDCB (figure 2). L'étude a porté non seulement sur des estimations des dépenses actuelles, mais aussi sur les dépenses prévues au cours des 10 prochaines années, compte tenu du taux actuel d'accroissement des principaux postes de dépense. Comme la méthode ITAO ne devait être utilisée que de façon transitoire pendant une période de 6 mois (c.-à-d. le temps nécessaire pour amener les unités ayant donné une réponse tardive suivant la méthode de collecte par la poste à répondre dans

les délais fixés suivant la méthode ITAO) avant que l'on applique pour de bon la méthode IDCB, l'analyse des coûts a surtout porté sur une comparaison des coûts afférents à la mise en oeuvre de la méthode de collecte par la poste et des coûts afférents à la mise en oeuvre de la méthode IDCB.

Sur le plan des catégories de dépenses salariales, l'utilisation de l'IDCB a permis de remplacer les opérations mensuelles d'envoi postal, de retour par la poste, d'enregistrement et de contrôle des formules par une seule opération annuelle d'envoi postal, et donc d'éliminer l'exécution d'une importante opération de bureau mensuelle dans les États. De même, les opérations de perforation par lots, de validation des perforations et de contrôle des formules nécessitées par la méthode de collecte par la poste ont été complètement éliminées grâce à l'IDCB, selon laquelle le répondant introduit au clavier à boutons les données relatives à son entreprise, puis valide lui-même chacune des entrées. Autre caractéristique d'importance, l'IDCB possède, sur le plan de rentabilité, un avantage certain par rapport à la collecte par la poste du fait qu'elle facilite le suivi téléphonique à l'égard des cas de non-réponse. Grâce à l'IDCB, il est possible de générer une liste exacte et à jour des répondants n'ayant pas encore transmis leurs données par téléphone, puis d'utiliser cette liste pour effectuer de brefs appels de sollicitation. Lorsque les données étaient recueillies par la poste, il était peu commode d'assurer le suivi téléphonique des cas apparents de non-réponse puisque les fonctionnaires de l'État ne savaient pas si la formule du répondant avait été remplie ou non, si elle avait été postée, ou si elle en était au stade de l'enregistrement ou à celui de la perforation; en outre, les répondants ayant déjà posté leur formule avaient tendance à s'offenser de recevoir un autre appel au sujet d'une activité qu'ils considéraient comme terminée. Comme la participation à l'enquête était facultative et qu'on ignorait si le répondant avait rempli son questionnaire ou non, on n'effectuait des appels de sollicitation qu'auprès des unités critiques (gros employeurs).

Comme le nombre de rejets au contrôle s'est établi à peu près au même niveau pour la collecte par IDCB et pour la collecte par la poste, l'utilisation de la première méthode n'a permis de réaliser aucune économie sensible au chapitre de la conciliation de contrôle. Il en a été de même pour les cartes postales de rappel, le nombre des cartes envoyées aux établissements donnant une réponse tardive suivant la méthode de collecte par la poste étant à peu près égal au nombre de cartes de préavis envoyées aux répondants pour leur rappeler d'introduire leurs données au clavier à la date prévue suivant la méthode de collecte par IDCB.

Sur le plan des catégories de dépenses non salariales, les frais d'affranchissement (actuellement 50 cents par unité) sont remplacés par la somme des frais d'appel et de la fraction non amortie du coût de l'appareil d'IDCB (actuellement 46 cents par unité). Par ailleurs, on assiste à une hausse annuelle des frais d'affranchissement de l'ordre de 5% environ (figure 2). Cette hausse est causée par une augmentation annuelle des frais de main-d'oeuvre (+ 5.7%) et par l'augmentation générale du prix de l'essence, les frais de main-d'oeuvre intervenant pour plus de 80% des frais totaux d'affranchissement. Par contraste, on a assisté au cours des dernières années à une diminution des frais d'appel (- 1.7%) ainsi que du prix d'achat des micro-ordinateurs (- 19.5%).

Sauf sur le plan de la nouvelle nécessité d'effectuer des appels de sollicitation à l'égard de tous les cas de non-réponse, on peut donc démontrer que le passage de la collecte par la poste à la collecte par IDCB permet de réaliser des économies. Toutefois, il est peut-être encore plus important de noter que, selon une projection sur une période de 10 ans des coûts afférents à la mise en oeuvre de ces deux méthodes (figure 5), ces économies connaîtront une progression substantielle. Par ailleurs, on tentera d'utiliser les économies réalisées grâce à l'utilisation de l'IDCB afin de compenser les frais relatifs à l'exécution d'appels de sollicitation à l'égard de tous les cas de non-réponse.

On peut tirer plusieurs conclusions importantes de cet examen du rendement et de cette analyse des coûts. La collecte des données par la poste ne constitue plus la méthode de collecte la moins coûteuse dont disposent les organismes statistiques. Non seulement les percées technologiques importantes des années quatre-vingt sur le plan de la collecte téléphonique automatisée permettent-elles d'abaisser les coûts d'utilisation de cette technique au-dessous de ceux de la collecte par la poste, mais elles permettent aussi de réduire les délais de collecte et d'exercer un meilleur contrôle sur le processus de collecte. En outre, au cours des 5 à 10 prochaines années, l'accroissement des frais de main-d'oeuvre et d'affranchissement relatifs à la collecte des données par la poste rendra cette dernière encore moins concurrentielle par rapport aux nouvelles méthodes à haute technologie et à faible intensité de main-d'oeuvre. Selon les interviews de suivi réalisées par le Bureau auprès des répondants s'étant convertis à l'IDCB, ces derniers éprouvent très peu de difficulté à s'adapter à cette nouvelle méthode de déclaration. Virtuellement, tous les répondants ont affiché une préférence marquée pour la déclaration par

IDCB par rapport à la déclaration par la poste ou par ITAO et trouvé que la nouvelle technique constituait un prolongement naturel d'autres applications similaires de la bureautique. De plus, on peut considérer que l'ITCB constitue une méthode de remplacement fiable pour la collecte des données d'enquête. Au cours des quatre dernières années de collecte, on n'a enregistré aucune défaillance majeure du matériel ni aucune interruption du processus de collecte. Au besoin, les défaillances mineures du matériel ont pu facilement être résolues à l'aide d'un OP de secours. En outre, pour plus de sécurité, le processus de collecte par IDCB des États prévoira une option de renvoi automatique des appels, destinée à réacheminer les appels à un central en cas de défaillance majeure du système utilisé à l'échelle de l'État.

3. MISE EN OEUVRE

3.1 Principales difficultés

À la fin de 1989, le Bureau avait terminé un programme de recherche couronné de succès lui ayant permis d'assurer le maintien d'un rendement élevé tout au long des 7 années qu'avaient duré les travaux. Toutefois, il y a un pas important à franchir entre la réalisation d'un programme de recherche réussi et la mise en oeuvre de nouvelles méthodes à l'échelle réelle. Bien qu'on ait recueilli des données auprès de plus de 10,000 unités au cours de la mise à l'essai de ces nouvelles techniques, ces unités représentaient moins de 3% de l'échantillon de la CESS. L'apport des modifications proposées pour la collecte mensuelle, auprès d'un échantillon de 350,000 unités, de données ayant été recueillies pendant plus d'un demi-siècle à l'aide d'un système décentralisé de collecte par la poste nécessite non seulement l'expression d'un besoin manifeste de changement de la part des utilisateurs, mais aussi l'existence d'un soutien très large à l'échelle du pays, des régions et des États.

En l'occurrence, les besoins des utilisateurs avaient commencé à évoluer dès le début des années quatre-vingt. Au cours de cette décennie, l'économie des États-Unis a connu la plus longue période de croissance soutenue en temps de paix de son histoire, le nombre des emplois créés s'établissant à 19 millions et les taux de chômage atteignant leur plus bas niveau depuis le début des années soixante-dix. Au milieu des années quatre-vingt, la politique économique était clairement axée sur l'établissement d'une croissance économique non inflationniste. Les observateurs examinaient d'un oeil attentif les données mensuelles de la CESS sur la croissance de l'emploi et sur les salaires afin d'y détecter tout signe de pression inflationniste par les salaires résultant de l'existence d'une forte croissance de l'emploi pendant une période de faible chômage. Cette utilisation et cette visibilité accrues des données mensuelles se sont accompagnées d'un accroissement correspondant de la frustration manifestée par les utilisateurs à l'égard des importantes révisions dont les estimations provisoires faisaient périodiquement l'objet. Bien que les estimations provisoires de la CESS aient toujours fait l'objet de révisions mensuelles et que l'importance de ces révisions ait diminué au fil des ans, l'apport de révisions de l'ordre de 100,000 unités aux estimations provisoires de l'évolution de l'emploi d'un mois à l'autre était maintenant considéré comme inacceptable. La demande des utilisateurs pour des estimations provisoires plus précises devait amener le Bureau à élaborer des propositions en vue de l'application des méthodes de collecte automatisée par ITAO et par IDCB dans le cadre de la plus importante enquête mensuelle réalisée par l'administration des États-Unis.

Bien que la demande des utilisateurs constitue un critère de première importance, des modifications de cet ordre ne sauraient être apportées sans le soutien inconditionnel des États chargés de la collecte des données. Tout au long du programme de recherche, les chercheurs devaient garder à l'esprit que le système de collecte étudié était celui des États et qu'il devait donc être conçu pour s'intégrer facilement à leurs mécanismes d'enquête et avoir l'incidence organisationnelle la plus faible possible. À cette fin, il a été possible d'apporter des modifications aux systèmes ITAO et IDCB pendant toute la durée des travaux. Chaque nouvelle version des systèmes tenait compte du plus grand nombre possible de suggestions et d'exigences des États. On peut dans une large mesure attribuer la réussite des travaux d'élaboration à l'ouverture d'esprit et à la persévérance dont ont fait preuve les 14 États participant aux travaux en ne cessant de formuler de nouvelles recommandations visant à améliorer les systèmes et les procédures. À la fin, la méthode ITAO s'était mutée d'une simulation malhabile d'interview d'enquête-ménage type en une technique rapide et efficace à «écrans» et «fenêtres» convenant parfaitement à la saisie et au contrôle de données économiques longitudinales. Ainsi, à la conclusion des travaux de recherche, les systèmes et procédures avaient fait l'objet de nombreux essais et affinements dans un large éventail d'États. Cette façon d'aborder les essais a permis aux États de développer une confiance solide

dans les méthodes et les systèmes proposés. Cette caractéristique se révélera essentielle pour le respect du calendrier de mise en oeuvre rapide de ces méthodes de collecte d'avant-garde proposé par le Bureau.

3.2 Approche et incidence

L'objectif principal de la proposition de mise en oeuvre était de maintenir au-dessous d'un seuil limite l'ampleur des révisions apportées aux estimations provisoires. Au cours des 5 dernières années, environ 40% des révisions étaient supérieures à 50,000 unités et 13% d'entre elles étaient supérieures à 100,000 unités (figure 6). L'étude de mise en oeuvre avait pour objectif de repérer un ensemble minimal d'entreprises donnant des réponses tardives qui, si on pouvait obtenir leurs réponses avant la date limite de diffusion, permettraient de maintenir l'ampleur des révisions au-dessous d'un seuil jugé acceptable (modification de moins de 50,000 unités du chiffre des fluctuations mensuelles). Bien sûr, il aurait été possible de demander à l'ensemble des 175,000 entreprises donnant des réponses tardives de se convertir aux nouvelles méthodes de collecte, mais on a jugé que la mise en oeuvre d'une telle solution serait longue et coûteuse. Bien qu'il ait été nécessaire de maintenir l'ampleur des révisions au-dessous d'un seuil acceptable (c.-à-d. que l'objectif visé n'était pas d'éliminer toutes les révisions), il fallait également s'assurer d'atteindre cet objectif dans les meilleurs délais (c.-à-d. en convertissant le plus petit nombre possible d'entreprises pour permettre le maintien des révisions au-dessous du seuil de 50,000 unités).

Contrairement aux enquêtes-ménages, les enquêtes auprès des entreprises utilisent en général une pondération différentielle pour les diverses unités de collecte, les unités très importantes constituant des unités à tirage complet dans le plan de sondage. Selon le plan de sondage de la CESS, les unités comptant 100 employés ou plus ne forment que 20% de l'échantillon (c.-à-d. qu'elles sont au nombre de 75,000), mais interviennent pour plus de 83% de l'emploi dans l'échantillon non pondéré. Comme ces unités ont tendance à afficher un taux de réponse beaucoup moins élevé pour les estimations provisoires, toute divergence entre la tendance de l'emploi chez les unités donnant une réponse tardive et la même tendance chez les unités répondant dans les délais fixés est susceptible de nécessiter une révision substantielle des estimations établies à partir de l'échantillon. Des études ont été réalisées afin d'évaluer l'incidence des unités comptant 100 employés ou plus sur les estimations provisoires. À cette fin, celles de ces unités donnant des réponses tardives ont été incluses dans l'échantillon initial utilisé pour établir les estimations provisoires et les estimations ont été recalculées. On a ensuite comparé ces estimations aux estimations provisoires initiales afin de déterminer l'incidence des unités comptant 100 employés ou plus sur les révisions. Les résultats ont indiqué que d'une demie à deux tiers de la révision était attribuable à ces unités. On a ensuite obtenu des résultats similaires en effectuant les mêmes études pendant plusieurs mois. En appliquant ces taux projetés de réduction de l'ampleur des révisions aux révisions effectuées au cours des 5 dernières années, on a découvert que 97% des révisions se situaient alors au-dessous du seuil de 50,000 unités comparativement à seulement 60% actuellement. Cette réduction considérable de la taille de l'échantillon visé par la conversion à la méthode ITAO/IDCB a rendu possible l'adoption d'un calendrier de mise en oeuvre accélérée permettant de maintenir les coûts de conversion au niveau minimal nécessaire pour éliminer les révisions d'importance des estimations provisoires. Le Bureau entreprendra la mise en oeuvre des méthodes de collecte ITAO/IDCB dans 25 États en 1991 et prévoit étendre l'utilisation de ces méthodes à tous les États en 1992. Grâce à ces nouvelles méthodes de collecte, le Bureau sera en mesure de contribuer à la solution d'un des problèmes de qualité les plus épineux et les plus apparents pour les utilisateurs des données de la CESS.

4. RÉSUMÉ

Les années quatre-vingt ont été une période de changements multiples pour les organismes statistiques. Certains de ces changements ont été le résultat des réalisations de ces organismes, tandis que d'autres, plus subtils, sont venus modifier le cadre général dans lequel nous réalisons nos enquêtes.

Au cours des années quatre-vingt, les utilisateurs sont devenus beaucoup plus conscients de l'importance de la qualité des données et prompts à cerner et à signaler les limites de nos produits. Sans que la qualité de nos produits ne se soit nécessairement détériorée, les attentes des utilisateurs sur le plan de la qualité et de l'utilisabilité se sont considérablement élevées. À titre d'organismes statistiques, nous nous devons de relever ce défi afin d'être en mesure de maintenir notre crédibilité auprès des utilisateurs. Nous avons également assisté au cours de la dernière décennie à de spectaculaires percées technologiques et, en particulier, à l'avènement de

l'ère des micro-ordinateurs. Cette nouvelle technologie a offert aux organismes statistiques de nombreuses occasions d'améliorer la qualité et le contrôle de la collecte des données : l'ITAO, l'AIAO, l'IDCB, la RP et le télécopieur. Certaines de ces techniques permettent d'améliorer la qualité et le contrôle tout en réduisant les frais de mise en oeuvre. Il est d'ailleurs fort possible que la prochaine décennie nous offre des occasions encore meilleures d'utiliser les nouvelles technologies pour améliorer la qualité et réduire les délais de collecte des données tout en sabrant dans les dépenses.

Lorsque nous examinons nos programmes statistiques, nous découvrons fréquemment un cadre de mise en oeuvre très rigide. Souvent, les méthodes de collecte des données utilisées pour nos enquêtes n'ont jamais été modifiées depuis la mise en oeuvre initiale de ces dernières. Les hypothèses que nous utilisons au sujet des coûts de collecte des données et nos analyses des coûts sont d'ordinaire nettement périmées et s'appuient sur des approches souvent simplistes. Comme la collecte des données intervient généralement pour la majeure partie des coûts de mise en oeuvre d'une enquête, elle est d'ordinaire bien ancrée dans la structure organisationnelle de l'organisme et il peut être assez difficile de la restructurer afin de permettre l'apport de modifications d'envergure. C'est à l'intérieur de ce cadre que nous devrons relever les défis et saisir les occasions que nous offriront les années quatre-vingt-dix.

Au cours de la prochaine décennie, les organismes statistiques devront relever le triple défi :

- d'être sensibles à l'évolution des besoins des utilisateurs en matière de qualité des données;
- de tout faire pour que leurs travaux de recherche tiennent compte de l'évolution rapide des technologies et des méthodes de collecte automatisée des données; enfin, peut-être par dessus tout,
- de trouver des façons d'appliquer les résultats des travaux de recherche couronnés de succès aux programmes courants.

Ces critères seront déterminants pour la future position concurrentielle de nos programmes et de nos organismes en termes de coûts et de qualité.

BIBLIOGRAPHIE

Clayton, R., et Harrell, L. (1989). Developing a Cost Model of Alternative Data Collection Methods: MAIL, CATI and TDE, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Clayton, R., et Winter, D. (1990). Speech Data Entry: Results of the First Test of Voice Recognition for Data Collection, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Groves, R.M.J. editor et coll. (1988). *Telephone Survey Methodology*, New York: John Wiley and Sons.

Phipps, P.A., et Tupek, A.R. (1990). Assessing Measurement Errors in a Touchtone Recognition Survey, présenté à Measurement Error Conference, Tucson, Arizona.

Ponikowski, C., et Meily, S. (1988). Use of Touchtone Recognition Technology in Establishment Survey Data Collection, présenté à la première Annual Field Technologies Conference, St. Petersburg, Floride.

Statistical Policy Working Paper 15 (1988). Quality in Establishment Surveys, Office of Management and Budget.

Statistical Policy Working Paper 19 (1990). Computer Assisted Survey Information Collection, Office of Management and Budget.

Werkings, G.S., Tupek, A.R., Ponikowski, C., et Rosen, R. (1986). A CATI Feasibility Study for a Monthly Establishment Survey, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 639-643.

Werking, G.S., et Tupek, A.R. (1987). Modernizing the Current Employment Statistics Program, *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 122-130.

Werking, G., Tupek, A. et Clayton, R. (1988). CATI and Touchtone Self-response Applications for Establishment Surveys, *Journal of Official Statistics*, 4, 4, 349-362.

Figure 1. **Résultats sommaires de la recherche**

		1984	1985	1986	1987	1988	1989	1990
Poste	Taux de réponse	47%	47%	48%	49%	49%	51%	52%
	Unités	400	400	2000	3000	5000	5000	5000
ITAO	Taux de réponse	83%	84%	82%	84%	83%	84%	82%
	% de rappels	44%	42%	40%	41%	42%	41%	41%
	Durée moyenne en minutes	5.6	5.6	5.0	4.8	4.4	3.5	3.8
IDCB et RP	Unités				400	600	2000	5000
	Taux de réponse				78%	80%	84%	82%
	% de rappels				45%	45%	43%	40%
	Durée moyenne en minutes				1.8	1.8	1.7	1.7

Figure 2. Coûts de collecte des données
(les flèches indiquent la direction des récentes variations de prix)

Catégorie de dépenses	Poste	ITAO	Auto-interview (IDCB et RP)
DÉPENSES SALARIALES			
Envoi postal	→		
Retour par la poste	→		
Entrée des données	→	→	
Contrôle et conciliation	→	→	→
Suivi des cas de non-réponse			→
DÉPENSES NON SALARIALES			
Frais d'affranchissement	→		→
Frais d'appel		→	→
Micro-ordinateurs		→	→

Facteurs récents de variation annuelle des prix

Main-d'oeuvre	+ 5.7%	ISE, Administrations des États et administrations régionales
Affranchissement	+ 5.0%	Service des postes des É.-U.
Frais d'appel	- 1.7%	IPC-É.U., appels interurbains à l'intérieur d'un même État
Micro-ordinateurs	- 19.5%	IPA Indices de prix expérimentaux (ordinateur de 16 bits)

Figure 3. CESS, rendement de l'ITAO à la première date limite, selon la durée de la période de paie
Septembre 1989 à août 1990

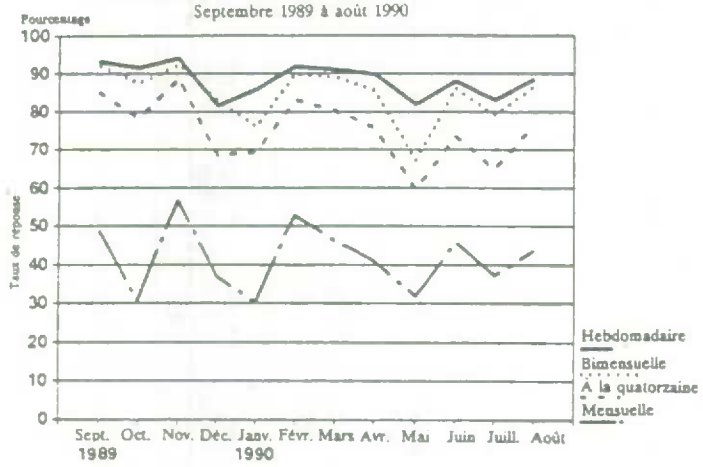


Figure 4. Introduction des données au moyen d'un clavier à boutons
Répartition des appels IDCB selon l'heure
Mai 1988 à février 1989

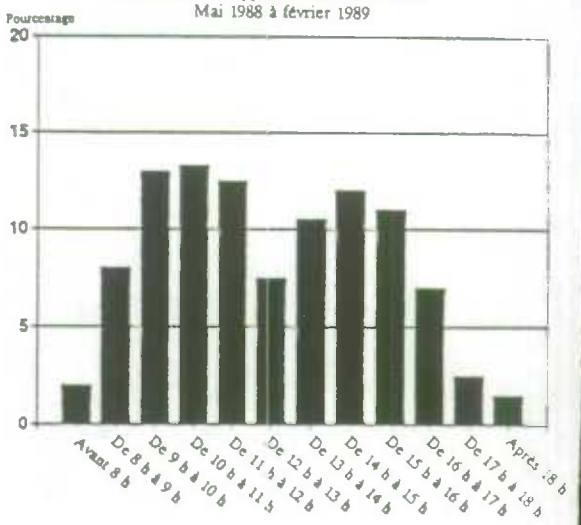
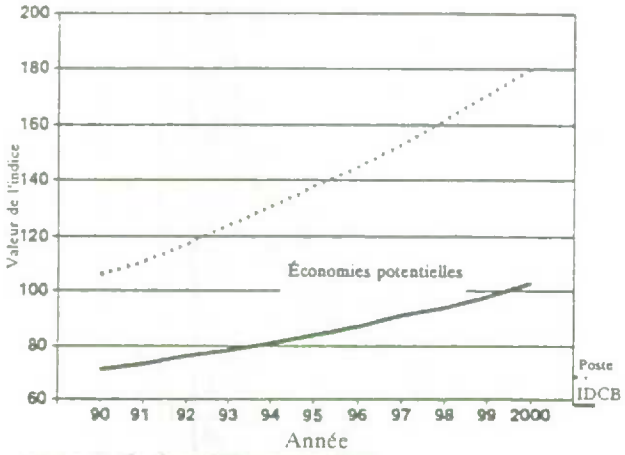
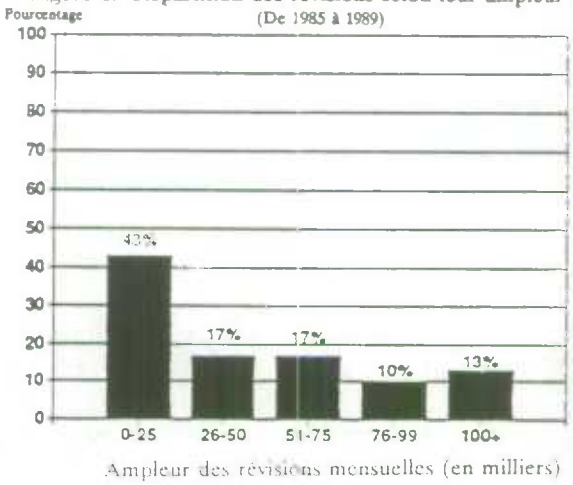


Figure 5. Coûts estimatifs par unité, selon le mode : de 1990 à 2000



Nota : Projections officieuses du BLS

Figure 6. Répartition des révisions selon leur ampleur
(De 1985 à 1989)



Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

STRATÉGIE DE SUIVI POUR LES ENQUÊTES ÉCONOMIQUES

J.-M. Berthelot et M. Latouche[1]

RÉSUMÉ

Lors du développement de la fonction générique de collecte et de saisie des données, une attention particulière a été portée à l'optimisation de la stratégie de suivi. L'objectif est de minimiser la quantité de ressources utilisées sans affecter, de façon significative, la qualité des données. Les ressources allouées au recontact sont concentrées sur les unités suspectes pouvant avoir un impact significatif sur les estimations. Cet article présente la stratégie globale pour les suivis dans les enquêtes économiques, ainsi que les résultats d'une étude empirique utilisant la stratégie développée.

MOTS CLÉS: Collecte et saisie; recontact; fonction de caractérisation.

1. INTRODUCTION

Une des préoccupations importantes de Statistique Canada depuis quelques années est d'essayer d'améliorer l'efficacité de ses opérations. On veut réduire le fardeau de réponse, améliorer l'efficacité du procédé de production ainsi que la qualité du produit fini. Pour répondre efficacement à ces objectifs, il est devenu impérieux de reconsidérer la conception de certains de nos procédés de production. Ceci a mené à la création du Projet de Développement des Fonctions Génériques d'Enquêtes (DFGE). Ce projet consiste en un développement d'outils génériques d'enquêtes de telle sorte que les méthodes, les opérations et le logiciel informatique en découlant respectent les exigences de la majorité des enquêtes (Colledge, 1987).

La Fonction Générique de Collecte et de Saisie (FGCS) est une des fonctions d'enquête développée dans le cadre du DFGE. Cette fonction comprend l'ensemble des activités nécessaires pour l'acquisition, la validation et la conversion en ordinolingue des données d'enquêtes (GSFD, 1989). La fonction générique de collecte et de saisie a pour but de fournir une base méthodologique solide sur laquelle l'ensemble des enquêtes pourra s'appuyer. L'objectif global visé est de maximiser l'efficience de la fonction générique de collecte et de saisie en termes de coûts, de temps, de ressources et surtout en termes de qualité des données (Bilocq, 1988).

La collecte et la saisie des données a depuis toujours été une opération exigeante au niveau des ressources humaines et des délais. En particulier, une des tâches qui peut être très coûteuse est celle du suivi. Le suivi consiste en un contact avec le répondant, subséquent au contact initial, dans le but d'obtenir de l'information manquante, de vérifier, et si nécessaire, de corriger l'information suspecte. Le suivi est utilisé pour la non-réponse, totale ou partielle, pour la correction des erreurs et pour la confirmation des données aberrantes.

Par conséquent, lors du développement de la fonction générique de collecte et de saisie des données, une attention particulière a été portée à l'optimisation de la stratégie de suivi. L'objectif est de minimiser la quantité de ressources utilisées sans affecter, de façon significative, la qualité des données. Les ressources allouées au recontact sont concentrées sur les unités suspectes pouvant avoir un impact significatif sur les estimations. Cet article présente la stratégie globale pour les suivis dans les enquêtes économiques, le développement d'une fonction de caractérisation pour la classification des unités suspectes selon leur impact potentiel au niveau de l'estimation, ainsi que l'application de la stratégie aux données de l'Enquête annuelle sur le commerce de détail.

---

[1] J.-M. Berthelot et M. Latouche, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11e étage, Section "A", Immeuble R.H. Coats, Ottawa, (Ontario), Canada, K1A 0T6. FAX (613) 951-1462.

2. LA SITUATION ACTUELLE

Jusqu'à présent, la stratégie de suivi a consisté à effectuer un recontact auprès de tous les répondants qui ne retournaient pas leur questionnaire, retournaient un questionnaire incomplet, ou retournaient un questionnaire comportant des données suspectes. L'objectif de cette stratégie est de s'assurer que toutes les erreurs sont identifiées et corrigées, peu importe leur impact au niveau de l'estimation.

Afin de s'assurer de l'identification de toutes les unités suspectes, une vérification extensive des données est effectuée. Un nombre important de règles est utilisé pour la vérification des données. Les listes de rejets produites par l'application de ces règles doivent être vérifiées manuellement, résultant en une charge de travail considérable. L'utilisation de cette approche résulte souvent en une sur-vérification des données. On produit des listes de rejets qui comportent un faible pourcentage d'unités réellement en erreur. À long terme une perte de crédibilité des règles de vérification en découle, pouvant affecter l'efficacité du procédé de vérification.

Différentes règles de vérification peuvent être appliquées à des étapes distinctes du traitement d'une enquête. Par conséquent, il est parfois nécessaire de recontacter un répondant à plusieurs reprises afin de pouvoir vérifier toutes les erreurs potentielles présentes sur un questionnaire. Un fardeau de réponse élevé peut découler de cette situation.

En raison des contraintes de budget et de temps, il est pratiquement impossible de recontacter toutes les unités comportant de la non-réponse ou des données suspectes. Une décision est prise généralement de recontacter uniquement certains répondants. Cette décision est effectué à un certain moment lors du processus de collecte et de saisie, dépendant des contraintes de temps et de budget. Il est peu fréquent que les limites imposées sur les efforts de recontacts soient basées sur un plan d'échantillonnage pré-établi. L'impact de ces limites sur la qualité des données ne peut par conséquent être évalué et inclus dans l'estimation de l'erreur de l'enquête.

En résumé, la stratégie actuelle est très dispendieuse, prend beaucoup de temps et résulte en un fardeau de réponse considérable. Le procédé est peu efficace en raison du grand nombre de règles de vérification utilisées. L'erreur découlant des limites imposées par les contraintes de temps et de budget ne peut pas être mesurée et incluse dans les estimations. De plus, il n'y a pas d'évidence prouvant qu'il est nécessaire d'identifier et de corriger toutes les erreurs pour assurer un niveau acceptable de qualité des données.

3. LA STRATÉGIE PROPOSÉE

Dans un contexte de restrictions budgétaires, il est pratiquement impossible d'appliquer la stratégie actuelle. Conséquemment, une nouvelle stratégie qui tient compte des contraintes opérationnelles doit être développée. Une attention particulière doit être portée à certains facteurs lors du développement de cette stratégie de suivi pour les enquêtes économiques. Il faut essayer de diminuer les ressources utilisées pour le suivi, améliorer le procédé de vérification, automatiser le processus, limiter les recontacts et, mesurer et évaluer l'erreur engendrée par le procédé de suivi.

La stratégie proposée est relativement simple et repose sur un grand principe: le développement d'une approche intégrée pour toutes les étapes du traitement des données. Ce principe a toutefois des répercussions sur chacune des étapes du traitement. Le reste de cette section traite de ces répercussions.

3.1 La saisie

La saisie consiste à convertir en ordinolingue l'information présente sur un questionnaire. À ce niveau, il faut s'assurer que le système informatique soit conçu dans l'objectif de transférer, sans modifications, l'information du questionnaire sur un médium électronique, peu importe l'information présente sur le questionnaire. De cette façon, on évite tout traitement manuel relié aux contraintes imposées par le système informatique. Il n'est plus nécessaire, par exemple, de vérifier les questionnaires manuellement afin de s'assurer que des champs définis comme obligatoire par le système informatique soient présents sur tous les questionnaires. Les données présentes sur un questionnaire deviennent des sénatus-consultes; des contraintes qu'il faut respecter. Il faut donc s'assurer que le système informatique soit suffisamment flexible pour effectuer la saisie de toute l'information

qui est présente sur un questionnaire sans obstruction due à l'information manquante. Les recontacts engendrés par les erreurs de saisie avec l'ancienne approche sont par le fait même éliminés.

3.2 La vérification

L'objectif de la vérification est de permettre l'identification des erreurs, des données suspectes et des incohérences présentes dans les données. Pour que la vérification soit faite de façon efficace, il est nécessaire de structurer la vérification des données en tenant compte de ces objectifs. Dans la stratégie proposée, la vérification a été séparée en trois étapes distinctes: la vérification de la saisie, la vérification pour les données suspectes et la vérification de cohérence. Les paragraphes suivants donnent un bref aperçu du contenu de ces trois étapes.

La vérification de la saisie des données a pour but de maximiser la concordance entre l'information effectivement saisie et l'information présente sur un questionnaire. Elle est effectuée en utilisant des règles de syntaxe, des vérifications d'étendue ainsi que des règles de cohérence de base. Elle est faite de façon interactive lors de la saisie. Lorsqu'un rejet est identifié par la procédure de vérification, l'opérateur doit simplement s'assurer que l'information qui a été saisie corresponde à l'information présente sur le questionnaire. Si tel est le cas, l'opérateur doit simplement accuser réception du message et continuer la saisie du reste du questionnaire. Dans le cas contraire, l'opérateur doit corriger l'information saisie pour qu'elle corresponde à celle présente sur le questionnaire. De cette façon, un questionnaire peut être saisi d'un seul coup, sans autres interventions manuelles.

La vérification pour les données suspectes a pour objectif d'identifier les données quantitatives aberrantes ou en erreur. Seule la non-réponse et les unités identifiées par cette procédure seront sujettes à un suivi. De cette façon, on limite la quantité potentielle de suivis. L'identification des données aberrantes ou en erreur est effectuée par l'intermédiaire de méthodes statistiques de détection de données aberrantes faisant appel à la distribution des données telle qu'observée lors d'un ou plusieurs cycles précédents d'une enquête (Hidiroglou et Berthelot, 1986).

La vérification de cohérence a pour objectif d'assurer que l'ensemble des variables quantitatives d'un questionnaire soient cohérentes entre elles. L'identification des incohérences est faite par l'utilisation d'un ensemble de règles linéaires imposées sur un groupe de variables. Si une incohérence est rencontrée elle doit être corrigée par le biais d'une imputation automatique des variables responsables de celle-ci. De cette façon, tous les suivis reliés à la présence d'incohérences dans un questionnaire sont éliminés. La vérification de cohérence doit être développée conjointement avec la vérification pour les données suspectes car leur rôle est complémentaire.

3.3 Le recontact

Le recontact est l'activité qui permet de retourner au répondant pour obtenir de l'information manquante et de confirmer ou corriger de l'information considérée comme suspecte par les règles de vérification. L'utilisation d'un recontact sélectif est une des caractéristiques principale de la stratégie. Les efforts de recontact sont concentrés à deux niveaux.

Premièrement, il est nécessaire d'effectuer un suivi pour la non-réponse totale. Une non-réponse peut aussi bien correspondre à une unité hors-enquête qu'à un mauvais répondant faisant partie du champ de l'enquête. Ce suivi est indispensable pour assurer qu'une imputation appropriée sera effectuée lors d'une étape ultérieure du traitement.

Deuxièmement, il est indispensable de concentrer le reste des ressources sur les unités ayant un impact significatif au niveau de l'estimation. Pour ce faire, une fonction de caractérisation peut être utilisée. Son objectif est d'évaluer l'impact au niveau de l'estimation des questionnaires comportant des données suspectes. Un plan de sondage des questionnaires qui utilise la valeur de la fonction de caractérisation comme variable d'intérêt permet d'optimiser l'allocation des ressources disponibles pour le recontact.

3.4 L'imputation

Puisque la stratégie fait la promotion d'un recontact sélectif, des données incohérentes ou manquantes seront présentes sur le fichier après le recontact. Afin d'obtenir un fichier complet et cohérent, il est nécessaire d'imputer des valeurs pour ces dossiers. L'imputation de ces données doit être faite de façon automatique. Le système générique de vérification et d'imputation (SGVI) développé à Statistique Canada est un candidat idéal pour cette tâche. Il utilise un ensemble de règles linéaires pour identifier les incohérences, localiser les champs à imputer et imputer des valeurs respectant les règles de cohérences. De plus, ce système garantit la cohérence au niveau des micro-données (Whitridge *et al.* 1988).

3.5 Mesure de l'erreur

Puisque la nouvelle stratégie promouvant la restriction des recontacts, il est nécessaire de pouvoir mesurer l'erreur engendrée par le fait de ne suivre qu'une partie des erreurs potentielles. Pour ce faire, un plan de sondage pour fin de recontact des erreurs potentielles doit être établi avant le début de la collecte. La conception du plan de sondage est effectuée en utilisant les résultats d'un ou plusieurs cycles précédents d'une enquête. Le plan de sondage doit couvrir l'ensemble de l'échantillon; c'est-à-dire que toutes les unités de l'échantillon doivent avoir une probabilité non nulle d'être sélectionnées pour le recontact.

Pour chacun des répondants, l'erreur peut être évaluée pour chacune des variables à l'aide des différences observées entre la valeur après recontact et la valeur avant recontact (incluant les valeurs imputées). Pour chacune des variables, l'évaluation de l'erreur au niveau des estimations est obtenue par l'entremise du plan de sondage pour les recontacts. Le principe mis de l'avant est similaire à celui utilisé dans un contrôle de la qualité. On utilise un sous-échantillon selon un plan pré-établi dans le but d'évaluer l'erreur au niveau de l'échantillon total.

La stratégie proposée fait la promotion de l'utilisation d'une procédure de recontact sélectif. Un effort systématique de recontact est effectué auprès de la non-réponse totale de façon à s'assurer que l'état du répondant (actif, inactif ou hors-enquête) est obtenu pour garantir si nécessaire une imputation appropriée lors d'une étape subséquente. Par la suite, les ressources reliées au recontact sont concentrées sur les unités suspectes qui peuvent avoir un impact significatif au niveau des estimations. La cohérence interne de tous les répondants est garantie par l'utilisation du SGVI. L'utilisation d'un recontact sélectif combiné avec le SGVI devrait réduire les coût opérationnels et les délais sans affecter de façon importante la qualité des données.

4. FONCTION DE CARACTÉRISATION

Un des éléments essentiels de la stratégie proposée est l'utilisation d'une fonction de caractérisation. Une fonction de caractérisation est une formule mathématique qui associe un score relatif à chacun des répondants. La fonction utilise comme paramètres d'entrée les caractéristiques du répondant qui sont reliées à son impact potentiel sur les estimations. Un score est calculé pour chacune des variables d'un questionnaire et par la suite, sommé au niveau de du questionnaire, afin d'obtenir un score global pour le répondant. Les répondants avec les scores les plus élevés sont considérés comme étant ceux qui ont le plus d'impact sur les estimations. Quatre critères ont guidé la recherche d'une fonction de caractérisation: la taille du répondant, la taille et le nombre d'erreurs potentielles, et des considérations d'ordre pratique.

La taille d'un répondant selon une variable d'intérêt est un bon indicateur de l'influence que celui-ci peut avoir sur les estimations et devrait être considéré dans l'élaboration d'une fonction de caractérisation. Une telle indication peut habituellement être obtenue soit par les données rapportées au cycle courant, par les données d'un cycle précédent ou bien par des données administratives.

Une mesure de l'erreur que pourrait engendrer une donnée suspecte est un facteur important pour discriminer les répondants. S'il était possible de recontacter en ordre décroissant les vraies erreurs, on atteindrait à un certain moment un point à partir duquel il n'est plus nécessaire de recontacter des répondants, l'erreur résiduelle étant négligeable. La taille de l'erreur peut être évaluée en utilisant des différences ou des tendances entre les

valeurs brutes et les données historiques. Dans le contexte d'une enquête par échantillonnage, le poids de sélection des unités doit aussi être utilisé lors de l'évaluation de l'erreur.

Dans la même ligne de pensée, le nombre de données suspectes dans un questionnaire est un autre facteur important. Parmi les unités ayant relativement le même impact, celle qui a le plus grand nombre de données suspectes devrait être recontactée en priorité de façon à vérifier le plus d'information possible.

Les considérations d'ordre pratique sont nécessaires afin de respecter les besoins des différents intervenants. Certaines variables peuvent être considérées comme essentielles par les spécialistes du sujet et par le fait même nécessitent une attention particulière. De plus, les méthodologistes peuvent constater qu'il est difficile de bien imputer certaines variables. Dans ces deux cas, les variables en cause devraient avoir un poids plus important dans le calcul du score global pour un répondant. Le taux de réponse au niveau des cellules doit aussi être considéré. Un faible taux de réponse dans une cellule peut faire augmenter les besoins de recontact pour celle-ci. La distribution des valeurs de la fonction devrait être similaire d'une cellule de publication à une autre. De cette façon, les paramètres de la fonction pourraient être déterminés à un niveau d'agrégation plus élevé que la cellule de publication et ainsi augmenter la stabilité de ceux-ci.

L'utilisation d'une fonction de caractérisation doit se faire à l'intérieur des limites opérationnelles. La fonction doit être facile à mettre en place et suffisamment flexible pour pouvoir s'adapter à différentes enquêtes. Son utilisation ne doit pas perturber le traitement des données. Une fonction de caractérisation doit utiliser uniquement l'information du répondant qui fait l'objet du traitement, conjointement avec des paramètres d'entrée.

Le développement d'une fonction de caractérisation doit tenir compte de ces critères. L'objectif est de développer une formulation générale qui permet un équilibre entre ces différents facteurs.

5. APPLICATION DE LA STRATÉGIE

De façon à évaluer la pertinence de la stratégie proposée, une simulation a été appliquée aux données de l'Enquête annuelle du commerce de détail de 1987. Les objectifs principaux étaient de vérifier l'applicabilité de certains des principes mis de l'avant dans la stratégie et d'évaluer l'utilisation d'une fonction de caractérisation avec des données d'enquête.

5.1 Données utilisées

L'Enquête annuelle du commerce de détail (EACD) est une enquête sous la responsabilité de la Division de l'industrie de Statistique Canada. C'est un recensement des détaillants canadiens ayant des ventes et recettes totales d'au moins un million de dollars par année. La population est séparée en chaînes et en indépendants. Les chaînes correspondent habituellement aux grandes entreprises ayant plus de quatre locaux d'affaires dans le même genre de commerce, les autres entreprises sont considérés comme des indépendants. Pour la simulation, uniquement les indépendants furent utilisés.

Pour les fins de l'étude, 12 variables économiques quantitatives ont été utilisées. Une partie de la population comprenant 2 genres de commerce pour toutes les provinces et deux provinces pour tous les genre de commerce a été utilisée. Cette partie comprend un total de 2053 questionnaires. Les genres de commerce proviennent des secteurs de l'industrie de l'alimentation (genre de commerce 20) et de l'industrie des pièces et accessoires pour automobiles (genre de commerce 120). Les deux provinces retenues sont l'Ile-du-Prince-Édouard et l'Alberta.

La stratégie a été appliquée aux données brutes de 1987, telles que fournies par les répondants. Les données finales de 1987, telles que produites par la Division de l'industrie, ont été utilisées comme données de contrôle pour évaluer l'efficacité des résultats de la simulation. Les données finales de 1985 et 1986 ont été utilisées pour le calcul des paramètres pour la vérification et pour la fonction de caractérisation.

5.2 Méthodologie

La méthodologie utilisée pour la simulation respecte le plus possible la stratégie de suivi proposée. Elle a toutefois due être adaptée de façon à tenir compte des contraintes opérationnelles.

5.2.1 Saisie et vérification

Pour s'assurer que les données brutes correspondaient bien aux données présentes sur un questionnaire, les 2053 questionnaires ont été re-saisis. Lors de la re-saisie, aucune modification n'a été apportée aux données. Afin d'assurer la qualité de la saisie, une double saisie a été effectuée.

L'établissement d'un plan de vérification pour les données suspectes et potentiellement en erreur a été fait en plusieurs étapes. Des règles de validation de base ont été développées en utilisant la structure du questionnaire et la connaissance des spécialistes de la Division de l'industrie. Ceci nous a permis d'identifier la non-réponse partielle ou totale ainsi que les unités hors enquête.

Par la suite, les variables ont été groupées de façon à former des sous-ensembles homogènes des variables en utilisant la méthode décrite dans Bilocq et Berthelot (1990). Le lien entre les variables de chacun des sous-ensembles a été modélisé. Ces deux étapes ont été effectuées en utilisant les données de 1985 et 1986 pour l'ensemble de la population.

Les données suspectes ou en erreur ont été identifiées en utilisant les règles de base et les modèles des liens entre les variables appliqués aux données brutes de 1987. Le résultat de ce plan de vérification (succès ou échec) a été conservé pour chacune des variables et pour chacun des questionnaires.

5.2.2 Recontact

Pour la simulation, uniquement les unités comportant de la non-réponse, des erreurs ou des données suspectes étaient sujettes au recontact. Un recontact a été simulé en remplaçant les données brutes d'un questionnaire par ses données finales. Cette façon de procéder correspond à faire l'hypothèse qu'un seul recontact est suffisant pour corriger tous les problèmes présents sur un questionnaire. Cette hypothèse est requise pour effectuer la simulation. En effet, il était hors de question de recontacter des répondants plus de 3 ans après la fin de la période de référence de l'enquête.

Les recontacts ont été séparés en deux catégories mutuellement exclusives: les automatiques et les sélectifs. Un recontact automatique a été utilisé pour la non-réponse totale, les unités potentiellement hors-enquête ainsi que pour les unités suspectes faisant partie d'une petite cellule de publication. Les petites cellules ont été définies comme tout croisement province par genre de commerce qui contenait moins de 10 unités rapportantes.

Par contre, un recontact sélectif a été effectué en utilisant une fonction de caractérisation. La fonction de caractérisation a été évaluée pour chacun des questionnaires ayant échoué la vérification pour les données suspectes. Le score obtenu donne une indication de l'impact au niveau des estimations pour un questionnaire donné. Plus le score est élevé, plus l'impact potentiel est important. Pour les fins de la simulation, uniquement les unités ayant un score supérieur à une borne pré-déterminé ont été recontactés. Ceci correspond à effectuer un recontact avec probabilité 1 pour les unités au-dessus de la borne et un recontact avec probabilité 0 pour les autres. Cette façon de procéder diffère sensiblement de la stratégie qui propose un plan d'échantillonnage avec probabilités différentes de 0 pour toutes les unités.

Lors du développement d'une fonction de caractérisation pour l'EACD, trois formulations différentes ont été explorées. Chacune d'entre-elles met l'accent sur un aspect spécifique ou sur une contrainte opérationnelle particulière. Après différentes analyses une formulation a été retenue. C'est un compromis entre la simplicité, la précision et la production d'une distribution similaire des valeurs de la fonction pour chacune des cellules de publication. Les formulations considérées ainsi que les analyses s'y rattachant sont documentées dans Latouche et Berthelot (1990).

La formulation générale retenue met l'accent sur la différence absolue entre la valeur courante brute et la valeur finale du cycle précédent. Elle est définie au niveau de chacune des cellules de publication.

Pour une cellule p donnée on a:

$y_{k,i,t}$ la valeur brute pour le répondant k de la cellule p pour la variable i au temps t;

$y_{k,i,t-1}$ la valeur finale correspondante au temps $t-1$;

$w_{k,i,t}$ le poids de sélection au temps t;

$\hat{Y}_{.,i,t-1}$ le total pour de la cellule p pour la variable i au temps $t-1$;

Pour un répondant k, la fonction est alors définie par:

$$\text{SCORE}_k = \sum_{i=1}^{I} \frac{W_{k,i,t} \times |Y_{k,i,t} - Y_{k,i,t-1}| \times Z_{k,i,t} \times v_{.,i,t}}{\hat{Y}_{.,i,t-1}}$$

où

$$Z_{k,i,t} = \begin{cases} 0 \text{ si } y_{k,i,t} \text{ est accepté par la vérification} \\ 1 \text{ si } y_{k,i,t} \text{ est suspect} \end{cases} \quad \text{et}$$

$v_{.,i,t}$ le poids indiquant l'importance de la variable i au temps t.

Pour la simulation, un poids uniforme a été utilisé pour chacune des variables; c'est-à-dire qu'aucune variable n'a été considérée comme étant plus importante qu'une autre. L'EACD étant un recensement, un poids de sélection égal à un a été utilisé. Le poids de sélection n'a pas été considéré.

Pour évaluer le comportement et l'effet de la fonction de caractérisation, différents taux de recontact ont été utilisés pour les recontacts sélectifs. Les taux de recontact considérés sont zéro, dix-sept, trente-quatre, cinquante et cent pour cent. Zéro pour cent correspond à effectuer uniquement les recontacts automatiques. Cent pour cent correspond à faire un recontact de toutes les unités comportant des données suspectes, tandis que cinquante pour cent représente un point milieu. Dix-sept et trente-quatre pour cent ont été utilisés comme compromis entre zéro et cinquante pour cent.

Ces différents taux de recontact ont été utilisés dans le but d'identifier le point à partir duquel un recontact additionnel n'améliore pas de façon perceptible la qualité des données. À partir de ce point, le coût supplémentaire d'un recontact n'est plus justifié et l'utilisation des ressources pour le recontact n'est plus optimale.

5.2.3 Cohérence et imputation

Le système générique de vérification et d'imputation de Statistique Canada a été utilisé pour s'assurer que tous les dossiers étaient complets et cohérents. Un ensemble de règles linéaires a été imposé sur les données après la simulation des recontacts. Cet ensemble est utilisé pour définir un dossier cohérent. Ceci revient à déterminer un espace multidimensionnel à l'intérieur duquel un dossier est considéré comme valide. L'emploi de ce type de règles permet l'utilisation de la programmation linéaire pour identifier et corriger automatiquement les variables causant l'incohérence d'un dossier (Schiopu-Kratina et Kovar, 1989).

Le SGVI a été exécuté pour chacun des cinq taux de recontact. Cinq fichiers contenant les résultats finaux de la simulation ont été créés suite à l'application du SGVI. Ces fichiers ont été analysés afin d'évaluer la pertinence de la stratégie proposée.

5.3 Résultats

Au niveau de la vérification des données, 1163 des 2053 questionnaires ont échoués au moins une des règles. Parmi ces 1163 échecs, 584 font partie des recontacts automatiques. Ils sont répartis en 245 non-réponses totales, 322 unités potentiellement hors-enquête et 17 unités suspectes à l'intérieur de petites cellules (croisement genre de commerce par province).

Il reste donc un total de 579 questionnaires comportant des données suspectes qui sont éligibles pour le recontact sélectif. Les cinq taux de recontact sélectif appliqués aux 579 unités ont déterminé le nombre total de recontacts effectués dans chacun des cas. Ces pourcentages se traduisent en 0(0%), 97(17%), 194(34%), 285(50%) et 579 (100%) recontacts sélectifs. Ces taux de recontact sont approximativement les mêmes pour chacune des cellules de publication.

Pour évaluer l'erreur engendrée par la stratégie, l'hypothèse a été faite que la valeur finale correspondait à la vraie valeur. Ceci revient à supposer que les données finales ne comportent aucune erreur et représentent la vérité. L'efficacité de la stratégie a été évaluée pour chacune des variables en comparant les estimations obtenues par la stratégie avec celles obtenues en utilisant les données finales de 1987, et ce à différents niveaux. La mesure utilisée a été nommée "pseudo-biais". Pour un niveau d'agrégation donné a et pour une variable i, le pseudo-biais est défini de la façon suivante:

$$\text{pseudo-biais}_{a,i} = \frac{\hat{Y}_{a,i,87} - Y'_{a,i,87}}{Y'_{a,i,87}} \times 100$$

Ou $\hat{Y}_{a,i,87}$ correspond au total selon la simulation et $Y'_{a,i,87}$, correspond au total selon les données finales.

Au niveau de la partie étudiée lors de la simulation, les résultats sont encourageants. La figure I, ci-dessous, montre pour trois niveaux d'agrégation différents le pseudo-biais pour la variable ventes et recettes totales pour les cinq taux de recontact sélectif. Pour chacun des taux de recontact sélectif le pseudo-biais est relativement petit pour cette variable. De façon générale, le pseudo-biais diminue lorsque le taux de recontact sélectif augmente. Le point majeur à souligner est que la réduction du pseudo-biais est plus importante lorsque le taux de recontact sélectif passe de 17% à 34% que lorsqu'il passe de 34% à 50%. Avec un taux de 34%, le pseudo-biais est seulement de -0.18% et il est peu amélioré avec un taux de 50%, passant à -0.14%. Le fait de faire 91 recontacts supplémentaires entre ces deux taux améliore l'estimation mais pas de façon substantielle. Il semble à première vue qu'un taux de recontact sélectif de 34% soit suffisant. Un patron similaire a été observé pour les autres variables étudiées.

De façon à s'assurer que l'on choisit le taux de recontact sélectif le plus approprié pour l'EACD, des comparaisons ont été effectuées au niveau des cellules de publication. La figure II ci-dessous présente la distribution du pseudo-biais des cellules de publication pour les cinq taux de recontacts. La moyenne du pseudo-biais par cellule avec un intervalle de confiance de plus ou moins un écart-type pour les cinq taux de recontact sélectif y sont présentés. La dispersion du pseudo-biais diminue de façon substantielle lorsque le taux de recontact passe de 0% à 17% et puis à 34%. La moyenne et l'écart-type du pseudo-biais sont relativement constant dès qu'un taux de recontact sélectif de 34% est utilisé.

FIGURE I--PSEUDO-BIAIS

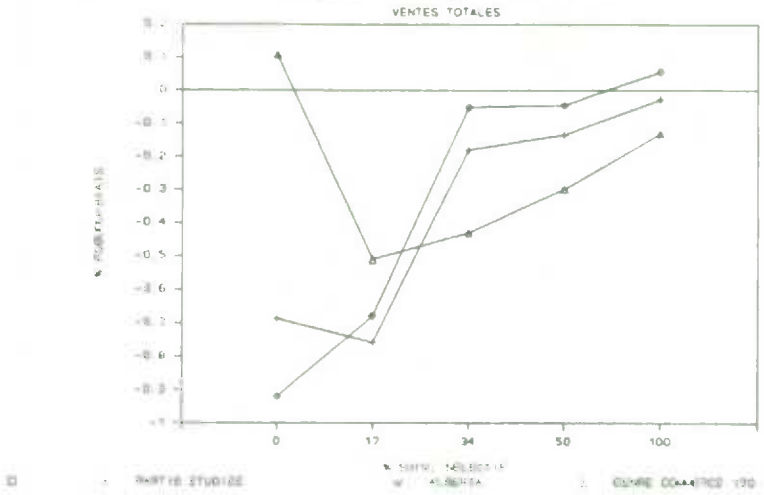
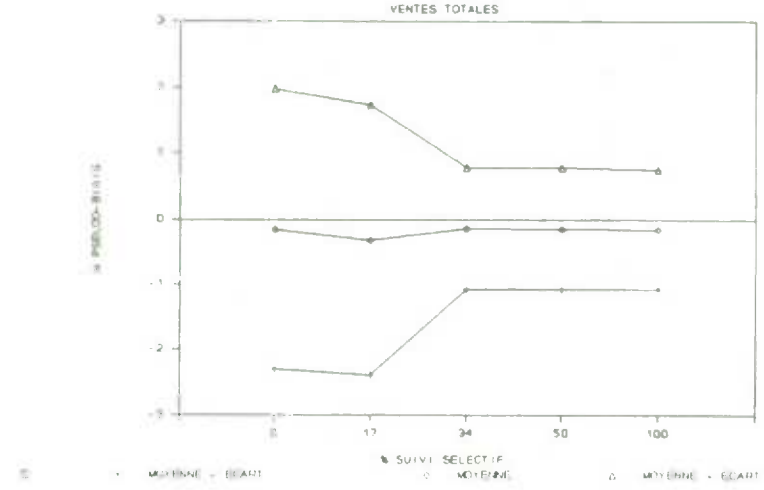


FIGURE II--MOYENNE +/- ECART



Un fait intéressant à noter à partir des figures I et II est que le pseudo-biais n'est pas égal à zéro lorsque 100% des recontacts sélectifs sont effectués. Différentes raisons peuvent expliquer ce fait. Premièrement, des erreurs peuvent avoir été identifiées lors de l'analyse, subséquente au traitement des données, effectuée par les économistes de la Division de l'industrie. Deuxièmement, le plan de vérification qui a été utilisé dans la simulation est différent de celui de l'enquête. Troisièmement, la méthode d'imputation utilisée dans la simulation est différente de celle utilisée dans l'enquête. Ces trois facteurs sont suffisants pour expliquer la différence entre la simulation et l'enquête lorsque 100% des recontacts sélectifs sont effectués.

Les résultats présentés dans les figures I et II semblent indiquer que les dépenses supplémentaires requises pour recontacter plus de 34% des données suspectes ne sont pas justifiées par la faible amélioration de la qualité qui en résulte. En tenant compte de l'objectif principal de la stratégie qui est d'optimiser l'utilisation des ressources sans trop affecter la qualité des données, un taux de recontact sélectif de 34% est jugé adéquat.

Ces résultats ont été obtenus à partir de l'analyse de la variable ventes et recettes totales uniquement. Afin de s'assurer que cette conclusion est valide de façon générale, le pseudo-biais pour un taux de recontact sélectif de 34% a été étudié pour les autres variables.

La table I présente le pseudo-biais pour les variables présentes chez plus de 95% des répondants pour les 2053 questionnaires utilisés pour la simulation.

Tableau I: Pleuso-biais, variables souvent rapportées

VARIABLES	TOTAL ENQUÊTE EN $1000	PSEUDO-BIAIS %
VENTES	8,515,601	0.28
VENTES TOTALES	9,015,559	-0.28
STOCK OUVERTURE	1,405,293	-0.34
STOCK FERMETURE	1,480,133	0.34
ACHATS	6,847,548	0.33
SALAIRES	937,354	0.78

Le pseudo-biais pour l'ensemble des variables souvent rapportées est relativement petit. La valeur la plus grande correspond à la variable salaires pour laquelle la simulation produit une sur-estimation de 0.78%. Un taux de recontact de 34% semble un compromis acceptable pour ces variables.

La table II ci-dessous présente le pseudo-biais pour les autres variables. Celles-ci moins rapportées, non pas en raison d'une non-réponse partielle mais plutôt par le fait que les caractéristiques qu'elles mesurent ne s'appliquent pas à tous les répondants.

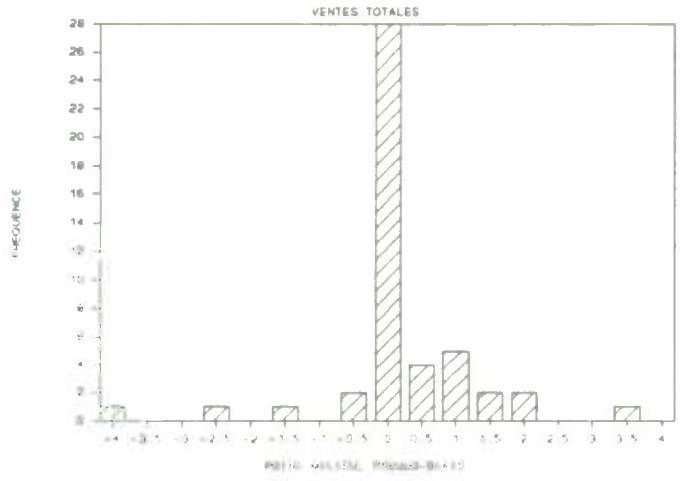
Tableau II: Pseudo-biais, variables sporadiquement rapportées

VARIABLE	TOTAL ENQUÊTE EN $1000	PSEUDO-BIAIS %	PRÉSENCE &
COMMISSIONS BRUTES	14,232	8.33	5
RÉPARATIONS	398,542	-3.20	46
LOCATION	41,344	0.09	10
RESTAURATION	9,325	-35.18	2
AUTRES EXPLOITATIONS	36,513	-37.55	15
HORS EXPLOITATION	16,349	-1.46	22

Le pseudo-biais est acceptable pour trois des six variables sporadiquement rapportées. Il est toutefois démesuré pour les *commissions brutes, les services de restauration* et *les autres revenus d'exploitation*. Pour les variables *commissions brutes* et *services de restauration*, on explique ce fait par leur très faible présence. Une aussi faible présence rend la tâche de la vérification et de l'imputation particulièrement difficile. Pour remédier à cette situation, on devrait donner un poids plus important à ces deux variables dans la fonction de caractérisation et appliquer un plan de vérification plus serré. La variable autres revenus d'exploitation est un cas spécial car elle est très peu reliée aux variables principales et peut varier beaucoup d'un répondant à un autre. Conséquemment, il est très difficile d'établir un plan de vérification adéquat pour cette variable. Même si le problème est différent, la solution est sensiblement la même: augmenter le poids de cette variable dans la fonction et développer un plan de vérification serré qui tient compte des caractéristiques spécifiques à cette variable.

La figure III présente l'histogramme de la répartition du pseudo-biais au niveau des cellules de publication pour la variable ventes totales pour un taux de recontact sélectif de 34%.

FIGURE III : PSEUDO-BIAIS PAR CELLULE



Dans cet histogramme, on note que les résultats de la simulation ne sont jamais à plus de 4% des résultats de l'enquête. Pour la plupart des cellules de publication, le pseudo-biais est à moins de 1.5% et la distribution est centrée à zéro. Les résultats de la simulation pour la variable ventes totales sont très encourageants. La distribution du pseudo-biais au niveau des cellules de publication est centrée à zéro et n'a pas une grande variance. Un patron similaire a été observé pour les variables souvent rapportées.

6. CONCLUSION

Les résultats obtenus à partir de la simulation de la stratégie proposée de suivi pour les enquêtes économiques sont prometteurs. Il est clair que le recontact des unités ayant un impact important au niveau des estimations est nécessaire pour l'obtention d'un niveau de qualité acceptable. La simulation a aussi démontré qu'il n'est pas nécessaire de recontacter toutes les unités suspectes. Le recontact d'un nombre limité d'unités suspectes est suffisant pour obtenir des estimations d'une qualité acceptable.

Le recontact sélectif des données suspectes combiné avec un système automatique de vérification et d'imputation permet d'obtenir des estimations dont le niveau de qualité est acceptable. L'utilisation de cette approche en production pourrait permettre d'économiser des ressources au niveau du traitement des données. Les ressources ainsi économisées pourraient être ré-affectées à d'autres tâches de l'assurance de la qualité et à l'analyse des données.

Même si les résultats de la simulation de la stratégie proposée sont intéressants, une faiblesse importante a été identifiée. Les résultats ne sont pas concluants pour les variables sporadiquement rapportées. On croît toutefois que ce problème peut être résolu en utilisant un plan de vérification plus serré et en augmentant l'importance relative de ces variables dans le calcul de la fonction de caractérisation.

La simulation présentée dans ce document représente un premier essai pour rendre opérationnel la stratégie de suivi proposée. Même si les résultats sont prometteurs, il reste beaucoup de travail de recherche à effectuer. On doit par exemple, considérer la possibilité d'effectuer un suivi sélectif pour la non-réponse totale et pour les unités hors-enquête, développer un plan de sondage pour pouvoir évaluer l'erreur engendrée par la stratégie, modifier la formulation de la fonction de caractérisation pour tenir compte de la non-réponse et simuler l'utilisation de la stratégie pour une enquête par échantillonnage. L'utilisation de l'approche proposée par Särndal (1990) pour l'évaluation de la précision des estimations lorsqu'on est en présence de valeurs imputées

mérite d'être étudiée. La possibilité d'étendre la stratégie développée pour les enquêtes économiques aux enquêtes sociales doit aussi être considérée.

BIBLIOGRAPHIE

Bilocq, F. (1988). La fonction générale de collecte et de saisie, présenté au colloque annuel sur les statistiques de l'ACFAS à Moncton, N.-B., mai 1988.

Bilocq, F., et Berthelot, J.-M. (1990). *Analyse sur le groupement des variables et sur la détection des données suspectes*, cahier de travail BSMD-90-005E/F, Statistique Canada, mars 1990.

Colledge, M. (1987). The Business Survey Redesign Project - Implementation of a New Strategy at Statistics Canada, présenté au US Bureau of the Census Third Annual Research Conference, Washington, mars 1987.

GSFD, Generalized Survey Function Development Team (1989). *Methodological and Operational Concepts in the Collection and Capture Module*, rapport technique, Statistique Canada, 1989.

Hidiroglou, M.A., et Berthelot, J.-M. (1986). Contrôle statistique et imputation dans les enquêtes-entreprises périodiques, *Techniques d'enquête*, 12, 1, 79-89.

Latouche, M., et Berthelot, J.-M. (1990). Use of a score function for error correction in business surveys at Statistics Canada, présenté à International Conference on Measurement Errors in Surveys, Tucson, Arizona, novembre 1990.

Särndal, C.E. (1990). *Estimation of precision in the generalized estimation system when imputation is used*, rapport technique, Statistique Canada, mars 1990.

Schiopu-Kratina, I., et Kovar, J.G. (1989). *Use of Chernikova's algorithm in the generalized edit and imputation system*, cahier de travail BSMD-89-001E, Statistique Canada, janvier 1989.

Whitridge, P., Kovar, J., et MacMillan, J. (1988). Système généralisé de vérification et d'imputation pour les enquêtes économiques à Statistique Canada, présenté au colloque annuel sur les statistiques de l'ACFAS à Moncton, N.-B., mai 1988.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

LA QUALITÉ DES DONNÉES DANS LES ENQUÊTES SUR LE RENDEMENT DES CULTURES

R. Fecso[1]

RÉSUMÉ

Les enquêtes sur le rendement des cultures effectuées dans de nombreux pays comportent plusieurs étapes de sous-échantillonnage qui sont déterminées par l'agent de dénombrement et nécessitent souvent plusieurs visites dans les exploitations choisies au dernier stade de l'échantillonnage. Ces techniques, de même que la prise de mesures souvent difficile associée au travail sur le terrain, peuvent soulever beaucoup de problèmes de qualité des données. Cette communication présente certains problèmes caractéristiques relevés lors d'enquêtes menées aux É.-U. Elle expose aussi une stratégie de gestion pour l'amélioration des données de même que les résultats d'études récentes sur la mesure de la qualité.

MOTS CLÉS: Gestion de la qualité totale, évaluation du rendement des cultures, biais.

1. INTRODUCTION

L'agriculture est l'activité économique la plus importante dans de nombreux pays et le rendement des cultures est souvent la composante de la production agricole la plus variable d'une année à une autre. Par conséquent, les enquêtes sur le rendement des cultures et les enquêtes connexes suscitent de l'intérêt depuis un certain temps déjà. Par exemple, Francis Bacon a étudié les liens entre le temps et les fluctuations agricoles au 17e siècle. Jusqu'au milieu de notre siècle, la recherche sur la mesure du rendement des cultures s'est développée. Bien qu'elles aient été modifiées souvent, les méthodes élaborées il y a plus de 30 ans pour mesurer le rendement des cultures demeurent semblables à celles qui sont utilisées de nos jours dans de nombreux pays.

La présente communication porte sur une méthode d'évaluation du rendement des cultures que l'on appelle la mesure objective ou récolte des cultures, telle qu'appliquée par le National Agricultural Statistics Service (NASS) du U.S. Department of Agriculture. Notre objectif principal consiste à examiner des études méthodologiques qui ne sont pas biens connues et à commencer à établir la documentation connexe. Cette documentation revêt de l'importance pour nous du fait que ces enquêtes n'ont pas fait l'objet d'une grande activité de recherche depuis plusieurs années. Étant donné que nous sommes en train de constituer un nouveau groupe de recherche pour l'étude de la mesure du rendement et que la plupart des membres du groupe sont nouveaux, la documentation de base des travaux existants les aidera à centrer leur attention sur les questions importantes. Nous espérons aussi que notre recherche encouragera le dialogue avec d'autres pays concernant leurs expériences avec la mesure du rendement.

1.1 Focalisation sur la qualité

Les organismes statistiques, tout comme le secteur privé et les organismes de services, s'intéressent aux questions relatives à la qualité. Il y a environ cinq ans, le NASS a mis sur pied un projet pilote de gestion de la qualité totale (GQT) de portée limitée à l'aide des données d'enquêtes sur le rendement objectif. Le projet a été bien accueilli par la direction, mais celle-ci a déterminé qu'une autre série de données qui présentait plus d'intérêt constituerait la première application officielle. Il est à espérer que les travaux relatifs à la GQT seront repris

[1] R. Fecso, Research and Applications Division, National Agricultural Statistics Service, U.S.D.A., Washington, DC 20250-2000, U.S.A.

en même temps que le programme de recherche sur le rendement. Ainsi, certains des résultats obtenus dans le cadre de la GQT seront traités dans le présent article. Fondamentalement, il s'agit de déterminer comment fournir les renseignements appropriés qui permettront de préciser de quelle façon la qualité des données peut être améliorée à partir d'un plan de sondage complexe comportant un estimateur complexe, compte tenu du fait que les ressources allouées pour la recherche sont limitées.

2. ENQUÊTES SUR LES CULTURES

La présente section fournit de brefs renseignements généraux relatifs aux enquêtes sur le rendement des cultures. Les données sur le rendement sont importantes car elles, ont une incidence sur les prix auxquels les marchandises se transigent sur les marchés boursiers, la répartition des services de transport (wagons porte-rails, péniches, etc.), le crédit agricole et divers programmes commerciaux et gouvernementaux. Il existe deux types fondamentaux de collecte de données sur le rendement: les données déclarées et les données recueillies objectivement (dénombrement sur le terrain).

Les données sur le rendement déclarées par les producteurs et autres particuliers bien informés au sujet des conditions agricoles locales remontent à 1862 aux Etats-Unis. Les enquêtes sur le rendement des cultures sont habituellement menées à l'aide de questionnaires envoyés par la poste qui demandent aux répondants d'évaluer l'état observé de leurs cultures ou le rendement prévu. De plus amples renseignements sur ces enquêtes peuvent être obtenues en consultant Fecso (1990). La présente discussion aura principalement trait aux enquêtes sur le rendement objectif (RO).

2.1 Enquêtes sur le rendement objectif

La collecte de données des terres en culture forme la base des enquêtes sur le RO. Le NASS mène ces enquêtes pour obtenir des données sur de nombreuses grandes cultures (maïs, soja, coton, blé, riz et pommes de terre) et des cultures spécialisées (arbres fruitiers, arbres à noix et raisins). Une vaste généralisation du plan de sondage, qui varie selon la culture, sera donnée ici. Pour plus de détails, voir Francisco, Fuller et Fecso (1987).

L'enquête de base est une enquête annuelle à base aréolaire. Un échantillon stratifié d'environ 15,000 segments aréolaires d'approximativement un mille carré en strates agricoles est utilisé pour repérer des champs des cultures visées. La superficie en acres est estimée à l'aide de cette enquête et les champs font l'objet d'un sous-échantillonnage en vue de l'évaluation du RO. Le plan de sous-échantillonnage donne lieu à un échantillon autopondéré. De façon générale, il y a moins de 2,000 champs choisis par culture, le choix étant limité aux principaux états producteurs (habituellement, il s'agit de 10 états qui représentent plus de 80 pour cent de la production d'une culture).

Dans chaque champ choisi, deux parcelles appelées unités sont choisies à l'aide d'une méthode de sélection aléatoire des rangées et (ou) des pas. La taille d'une unité dépend de la culture. Dans le cas des cultures dont les plants sont peu espacés comme le blé, les unités sont d'environ une verge carrée tandis que pour les cultures comme le maïs, les unités sont des rangées de 15 pieds de longueur. Chaque agent de dénombrement se voit attribuer environ 15 échantillons qui peuvent être répartis parmi plusieurs cultures à rendement objectif dans leur secteur. Le petit nombre d'échantillons qui est nécessaire pour produire surer des données d'enquête du temps opportun, rend difficile la mesure des effets reliés à l'agent de dénombrement.

Les mesures prises sur les unités varient selon la culture et le mois d'enquête. Par exemple, le rendement de fin de saison pour le maïs est déterminé par l'utilisation suivante des données recueillies:

Estimation du rendement brut par acre de maïs pour le champ

$$= \frac{\text{(43,560 pieds carrés dans un acre}}{\text{(longueur de toutes les rangées de l'échantillon [4 x 15 = 60]) (largeur de 8 rangées/8)}}$$

$$X \quad \text{[(Nombre d'épis dans l'unité 1) + (nombre d'épis dans l'unité 2)]}$$

X $\frac{\text{Poids des épis épluchés dans le champ}}{\text{Nombre d'épis épluchés dans le champ}}$

X $\frac{\text{Poids des grains détachés de l'épi}}{\text{Poids des épis égrenés}}$

X $\frac{\text{(Poids des grains analysés pour leur teneur en eau) (1-[\% d'humidité} \div 100])}{\text{Poids des épis dans le sac - poids du sac) (fraction d'égrenage)}}$

X (Ajustement du poids en fonction des boisseaux)

Les mesures détaillées de fin de saison sont prises pour faciliter les prévisions. Avant cette période, les plants n'ont pas encore pleinement développé les caractéristiques mesurées à la maturité. Les caractéristiques observables à ce stade, qui présentent la plus forte corrélation avec les mesures finales et sont raisonnablement faciles à recueillir, sont mesurées sur place et utilisées pour établir des prévisions relatives à la culture. Par exemple, durant les premiers mois, le nombre de tiges est utilisé pour prévoir le nombre futur d'épis. Pour plus de détails sur les modèles, voir Reiser, Fecso et Chua (1989) et Reiser, Fecso et Taylor (1987).

Les échantillons ci-dessus fournissent des données servant à évaluer le rendement brut ou biologique. Le rendement économique, l'information qui nous intéresse, doit faire l'objet d'un ajustement pour tenir compte de la perte attribuable à la récolte. Dans un demi-échantillon des champs, deux unités additionnelles sont repérées après la récolte. Les grains qui restent dans l'unité sont glanés et leur nombre est calculé pour un acre.

3. ÉTUDES MÉTHODOLOGIQUES

L'étude méthodologique de la mesure du rendement objectif ne date pas d'hier (voir Mahalanobis (1946) et Zarkovich (1966)). La présente section vise à présenter certaines des constatations faites par le NASS qui ne sont pas susceptibles d'être connues en dehors de ce service. La présentation sera structurée selon les composantes de l'équation d'estimation du rendement. Pour chaque composante, une brève description des problèmes ou des méthodes élaborées sera fournie. Notre intention n'est pas d'établir un profil complet de toutes les erreurs, bien qu'il s'agisse là d'un objectif à long terme, mais de donner un aperçu de la complexité de la question du contrôle de la qualité des données de ces enquêtes.

3.1 Arrangement des unités

Les deux unités sont repérées dans un champ à l'aide d'une méthode aléatoire visant à donner à tous les plants une probabilité presque égale de sélection. Le NASS détermine ses unités en choisissant une rangée au hasard et en précisant ensuite un nombre aléatoire de pas dans le champ le long de la rangée choisie. (Lorsque les rangées ne sont pas visibles, on a plutôt recours à un nombre aléatoire de pas le long de la bordure du champ). Le nombre aléatoire de rangées et de pas est déterminé selon la superficie en acres du champ. Un algorithme détermine les nombres aléatoires lorsque les champs sont choisis. La sélection est limitée au quart du champ qui est le plus accessible. On n'a pas constaté que cette méthode causait un biais lorsqu'elle a été étudiée. Les plants le long de la bordure et dans les sillons de virage du tracteur constituent un problème. L'algorithme a été changé plusieurs fois dans les années 50 et 60 afin d'assurer une distribution raisonnablement uniforme des unités près des bordures et dans les sillons de virage.

Dans le cas des arbres fruitiers, un autre niveau de randomisation est nécessaire. Les agents de dénombrement choisissent des nombres aléatoires pour déterminer un chemin qui leur permet de suivre une branche jusqu'à son extrémité pour compter les fruits qui s'y trouvent. Il faut appliquer des méthodes rigoureuses pour le marquage des arbres et la sélection des nombres aléatoires afin d'éviter la présélection de branches plus faciles à atteindre.

Le fait de marcher à pas mesurés le long d'une rangée peut porter l'agent de dénombrement à causer un biais. Celui-ci voyant l'état du champ devant lui peut être portée à raccourcir ou à allonger son pas pour atteindre un endroit visuellement attrayant. Un espace tampon est utilisé pour aider à réduire le biais possible dans un tel

cas. Lorsque le dernier pas a été fait, l'agent de dénombrement mesure ou place une perche à la pointe de son pied. La perche mesure habituellement cinq pieds. (Pour certaines cultures, un ruban à mesurer est utilisé pour délimiter l'espace tampon de cinq pieds.) La limite de l'espace tampon indique le point où commence l'unité à échantillonner. Une expérience contrôlée s'appliquant au blé a montré que les chiffres obtenus quand on emploie un espace tampon sont inférieurs à ceux qui sont obtenus lorsque l'on n'a pas recours à cette technique. Des différences ont été observées en moyenne entre les chiffres des unités un et deux, ce qui constitue une autre indication que la méthode des pas mesurés peut entraîner des biais subtils.

3.2 Mesure de la superficie en acres

Le rendement par acre d'un champ est déterminé en convertissant le rendement par unité à un rendement par acre. La taille des unités varie selon la culture. Les unités des champs où les rangées sont évidentes ont des longueurs fixes, mais la largeur de leurs rangées varie. Dans les champs où les rangées ne sont pas évidentes, une unité rectangulaire de taille fixe est créée.

Le fait de ramener l'unité de surface à un acre donne lieu à des erreurs tant pour ce qui est de la longueur que de la largeur de l'unité. Selon une nouvelle mesure des parcelles faites à l'aide d'un ruban pour mesurer quinze pieds après l'espace tampon de cinq pieds, plusieurs parcelles avaient une longueur de 20 pieds. L'utilisation de cadres métalliques fixes d'environ trois pieds de longueur peut causer, aux extrémités de ces cadres, une inclusion ou une exclusion de plants, ce qui change effectivement la longueur d'une parcelle.

Les largeurs des rangées sont variables. Pour contrôler cette variation, les agents de dénombrement prennent deux mesures, en travers des espaces d'une et de quatre rangées, pour les unités servant à déterminer le rendement brut. Pour déterminer les pertes relatives à la récolte, les mesures sont prises en travers des espaces d'une et de cinq rangées dans des unités additionnelles. Le rapport de la mesure de quatre rangées à une rangée présentait une distribution intéressante. Il y avait une importante pointe à quatre avec une erreur normale de faible variance aux environs de quatre. Il y avait aussi une pointe à cinq. (On a constaté l'inverse dans le cas des données relatives au rapport de cinq à un établies après la récolte. On s'est interrogé pour savoir si les instructions sur la façon de mesurer entre cinq rangées (quatre espaces) auraient pu être mal comprises ou si les instructions relatives aux mesures à prendre avant et après la récolte auraient pu être interverties. Un an plus tard, ce phénomène a été mentionné durant les cours de formation. La pointe additionnelle est disparue sans qu'aucun autre changement n'ait été apporté au contenu de la formation.

3.3 Chiffres relatifs à l'unité

Différentes caractéristiques agronomiques sont mesurées ou observées dans l'unité selon la culture et le mois. La fatigue et d'autres difficultés reliées à la tâche font que ces données sont sujettes à des erreurs. Les chiffres tendent à diminuer en moyenne à mesure que le dénombrement se poursuit dans les champs, ce qui est attribuable à la fatigue découlant du comptage répété d'un grand nombre d'éléments. D'autres tâches sont difficiles. La détermination de la phase de maturité peut poser des problèmes dans les cas limites d'une phase de maturité. Une erreur dans la détermination de la phase de maturité entraîne une erreur de mesure au niveau des prévisions. Il est également difficile de compter les noeuds et les tiges latérales des plantes de soja, de mesurer la longueur des épis de maïs ou des rangs de grains, de trouver toutes les capsules de coton ou de déterminer la formation des graines.

3.4 Détermination des poids

Il faut déterminer le poids par corps fructifère. Pour obtenir cette donnée, un sous-échantillon de corps fructifères tiré des unités dans le champ est envoyé à un bureau régional où il fera l'objet d'autres mesures. Le sous-échantillonnage peut causer des erreurs. Par exemple, on a constaté que le poids moyen des épis de maïs envoyés au bureau régional était supérieur à celui de tous les épis de l'unité. Les chiffres relatifs aux épis de blé établis dans le champ varient parfois considérablement de ceux qui sont calculés au bureau régional. Enfin, il importe que les appareils de laboratoire soient vérifiés afin d'assurer l'exactitude des données relatives au poids, à la teneur en eau et au rapport entre les grains et autres matières.

3.5 Perte attribuable à la récolte

Bon nombre des problèmes mentionnés ci-dessus peuvent se produire durant l'évaluation de la perte attribuable à la récolte. D'autres erreurs peuvent résulter des pertes causés par les oiseaux entre la récolte et la visite de prélèvement de l'échantillon. Il arrive aussi que l'on trouve tous les grains par terre, particulièrement lorsque la terre a été écrasée par les pneus du matériel agricole.

4. EFFORTS PERMANENTS VISANT À AMÉLIORER LA QUALITÉ

Maintenant que nous avons vu les problèmes typiques reliés aux données, il importe de déterminer ce qu'il faut faire pour y remédier. Le NASS prévoit plusieurs activités pour aider à contrôler et à réduire le taux d'erreur des données , notamment des séances de formation, des techniques et méthodes de contrôle qualitatif, et des études et des améliorations méthodologiques.

4.1 Formation

Le NASS applique la méthode selon laquelle on forme d'abord l'instructeur qui transmet à son tour ses connaissances selon la structure organisationnelle. Une séance de formation nationale a lieu pour les statisticiens des bureaux régionaux qui sont responsables de l'enquête dans leur état respectif. Il arrive souvent que la formation soit dispensée plusieurs mois avant le début de l'enquête. Ces statisticiens organisent ensuite une séance de formation dans leur état respectif au cours de laquelle ils retransmettent la matière apprise aux agents de dénombrement et aux surveillants.

Le programme de formation soulève plusieurs préoccupations. Premièrement, l'uniformité de la formation des agents de dénombrement d'un état à un autre peut constituer un problème selon la méthode consistant à former l'instructeur. De récentes coupures budgétaires ont accentué ce problème. La séance de formation initiale sur l'évaluation du rendement a été combinée à celle qui portait sur une autre enquête. De plus, une partie du matériel didactique plus dispendieux, par exemple les encarts en couleur pour les manuels de formation, a été éliminée.

Dans un esprit plus positif, on s'intéresse de plus en plus à l'utilisation de méthodes d'évaluation plus structurées dans le cadre de notre programme de formation. On a plus souvent recours à des exercices de contrôle portant sur les concepts, et les évaluations des résultats donnent lieu à des discussions stimulantes au sujet des méthodes de formation optimale. Par exemple, il semble qu'il y ait peu à gagner à demander aux agents de dénombrement de faire plus de trois heures d'exercices préparatoires à la formation à domicile. En outre, l'amélioration des notes obtenues aux exercices de contrôle est surtout constatée chez ceux qui ont plus de deux ans d'expérience. Enfin, on a effectué une enquête générale auprès des agents de dénombrement pour connaître leurs exigences concernant leur travail de même que leurs préoccupations. L'analyse de ces observations, qui ne fait que commencer, devrait donner des résultats intéressants.

4.2 Mesures et méthodes de contrôle qualitatif

Les méthodes de contrôle de la qualité appliquées aux opérations d'enquête et caractéristiques de celles qui sont utilisées par de nombreux organismes d'enquête sont aussi employées pour les enquêtes sur le RO du NASS. Depuis les premières enquêtes opérationnelles sur le RO du maïs et du coton en 1961, le NASS a produit bon nombre des mesures de qualité habituelles. Ces mesures comprennent les taux de refus, les taux d'inaccessibilité, le nombre d'échantillons par agent de dénombrement, le coût par échantillon, l'erreur d'échantillonnage et les taux de rejet au contrôle.

Les principales méthodes de contrôle de la qualité en temps réel sont les secondes vérifications du travail par les agents superviseurs et les contrôles des formules remplies. Les agents superviseurs retournent à un échantillon du travail effectué durant la première journée de travail de chaque agent de dénombrement et à un échantillon choisi au hasard du travail effectué ultérieurement pour vérifier les chiffres de nouveau. Ce processus constitue avant tout une autre étape de la formation car l'agent de dénombrement accompagne souvent l'agent superviseur. Trop peu de données sont recueillies pour permettre l'ajustement statistique des données d'enquête

complètes. De plus, certains comptes changent naturellement à mesure que des plants parviennent à maturité entre les visites.

Les données sont introduites manuellement et vérifiées à mesure que l'enquête progresse. Les contrôles des données sont faits par lots de façon périodique tout au long de l'enquête. Actuellement, les contrôles sont de simples vérifications de l'étendue et de la cohérence. L'élaboration de contrôles plus perfectionnés pourrait contribuer à améliorer la qualité des données. Par exemple, il pourrait s'agir de chercher les valeurs extrêmes et les points d'amplification dans le cadre de régressions de variables telles que le poids et le nombre de corps fructifères. Les vérifications de l'étendue peuvent varier en fonction de l'information sur les catégories; par exemple, le rendement de rangées étroites de soja présente une moyenne plus élevée que les plants en rangées larges. La possibilité de saisir des données à l'aide d'enregistreurs portatifs de données sur bande magnétique a été étudiée il y a sept ans et fait de nouveau l'objet d'un projet de recherche. Les problèmes de transmission des données qui ont interrompu les travaux dans le passé ont été résolus en grande partie, mais la question du coût peut encore constituer un obstacle, compte tenu de la petite taille de l'échantillon par agent de dénombrement. L'informatisation de la mesure du poids a été mise en oeuvre avec succès au bureau régional, réduisant ainsi les erreurs de transcription.

4.3 Études méthodologiques et améliorations

Le NASS se fonde sur des études visant à tester des hypothèses précises concernant les erreurs ou de nouvelles méthodes ainsi que sur des études générales de validation pour orienter le processus de modification des enquêtes en vue de l'amélioration des résultats. Des exemples des résultats d'études spéciales ont été présentés. Le NASS a recours aux études de validation depuis longtemps. Avant 1987, le plan de sondage de ces études visait typiquement un ou deux états, et de 16 à 32 champs étaient choisis. Les unités d'un champ étaient reproduites jusqu'à 32 fois par champ pour mesurer les sources d'erreurs reliées aux méthodes. Bien que les résultats aient varié selon l'année, l'état et la culture, une méta-analyse des études entreprises par Warren (1985) indique un biais relié à la méthode de 6 à 9 pour cent pour le maïs. Quoique non documentés de façon officielle, des résultats semblables ont été obtenus pour le soja. Ces biais ont subsisté en dépit de trente années d'études.

Le niveau de ces biais répétitifs nous a incité à évaluer nos méthodes de validation et de recherche. Le rendement étant le produit de nombreuses mesures, lesquelles comportent toutes plusieurs sources possibles d'erreur, nous nous sommes demandés dans quelle mesure la méthode de contrôle permettait d'obtenir les résultats voulus à la première application. Heureusement, nous avions une solution qui consistait à faire un échantillonnage double pour tenir compte du biais (Fecso, 1986). En gros, selon le nouveau plan de sondage, le rendement réel d'un sous-échantillon de champs est déterminé à partir du poids du grain récolté. Il s'agit là d'un rare exemple d'enquête permettant d'obtenir une valeur réelle. Le biais est considéré comme la différence entre les résultats obtenus par les méthodes d'enquête et le poids de la récolte pour les observations appariées.

Cette nouvelle méthode de validation a été utilisée pour le soja et a donné les résultats suivants:

ANNÉE	Biais estimé (boisseaux)	Erreur-type de l'estimation
1987	2.2	.9
1989	1.9	.8
1989	3.2	.9

Le principal critère du plan qui consistait à relever les biais de plus de 5% (1.6 boisseau) a été respecté. Les biais estimés pour ces trois années ont varié entre 6 et 9 pour cent. Des variations importantes du biais peuvent aussi être décelées, ce qui représente un avantage important lorsque les conditions d'enquête changent avec le temps. Les données préliminaires de 1990 indiquent qu'il est possible qu'un changement se soit produit.

Bien que l'estimateur de cet échantillonnage double place l'estimation d'enquête sur une base pleinement statistique (plutôt que de supposer l'existence d'un certain niveau de biais ou de tirer des conclusions pour l'ensemble des états à partir des résultats de deux états), la variance de l'estimation est plus grande que nous l'aurions souhaité. Il pourrait être utile de se pencher de nouveau sur la répartition des ressources du fait que certains aspects de l'enquête semblent faire l'objet d'un échantillonnage excessif par rapport à d'autres.

Il est possible que nous poursuivions l'élaboration des méthodes de contrôle du traitement pour les opérations d'enquête alors que nous embauchons de nouveaux employés pour la recherche sur le rendement. Les techniques de contrôle du traitement utilisées dans le secteur privé peuvent être modifiées pour répondre aux besoins des enquêtes. Des séries chronologiques simples de mesures telles que la mesure de quatre rangées par rapport à une rangée ou la différence entre l'unité un et l'unité deux peuvent fournir des indices concernant les changements des conditions de base de l'enquête. Des techniques statistiques plus complexes telles que LISREL (Reiser, Fecso et Chua, 1989) pourraient aussi s'avérer utiles.

5. CONCLUSION

L'objet de la présente communication est de donner un aperçu du nombre et de la complexité des tâches relatives à la mesure visuelle et mécanique des valeurs qui sont requises pour calculer le rendement. Les lecteurs intéressés sont invités à communiquer avec l'auteur pour obtenir de plus amples renseignements ou pour discuter de leurs expériences avec la mesure du rendement des cultures.

Les principes de gestion de la qualité totale peuvent être appliqués à nos efforts renouvelés de recherche sur le rendement. Nous devons cerner les principaux problèmes qui affectent la qualité de nos données afin de pouvoir mieux équilibrer la nécessité d'obtenir des données plus nombreuses et précises, les restrictions budgétaires et les limites personnelles , et les exigences relatives à l'obtention de données plus actuelles. Le fait que l'on accorde maintenant plus d'importance à ces éléments entraîne des changements dans l'enquête, ce qui fausse souvent les estimations, parfois de façon inattendue. Par conséquent, nous nous tournons vers des techniques telles que les méthodes de validation et les mesures du traitement pour repérer les changements relatifs à la précision et pour nous aider à déceler les sources d'erreur importantes. Si nos efforts réussissent, nous disposerons d'une façon efficace de gérer nos ressources limitées en vue de mettre au point les améliorations méthodologiques nécessaires au niveau de la précision, de la facilité du déroulement des opérations ou de la réduction des coûts. À part ces idées de "contrôle de la qualité", on prévoit que l'utilisation de panels, la nature géographique et multivariée des données contribueront à accroître l'efficacité des méthodes.

BIBLIOGRAPHIE

Fecso, R. (1986). Sample Survey Quality: Issues and Examples from an Agricultural Survey, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Fecso, R. (1990). A Review of Errors of Direct Observation in Crop Yield Surveys, en révision pour publication.

Francisco, C.A., Fuller, W.A., et R. Fecso (1987). Propriétés statistiques des estimateurs de la production végétale, *Techniques d'enquête*, 13, 1, 53-70.

Mahalanobis, P.C. (1946). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. *Journal of Royal Statistical Society*, 109(4), 325-370.

Reiser, M., Fecso, R., et Chua, M. (1989). Some Panel Aspects of the Objective Yield Survey, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Reiser, M., Fecso, R., et Taylor, K. (1987). A Nested Error Model for the Objective Yield Survey, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Warren, F.B. (1985). Corn Yield Validation Studies, 1953-83, Washington, D.C., USDA, SRS, Staff Report No. YRB-85-07.

Zarkovich, S.S. (1966). Quality of Statistical Data, Rome, Food and Agricultural Organization of the United Nations.

SESSION 6

Mesure et amélioration de la qualité des dossiers administratifs utilisés pour remplacer les enquêtes traditionnelles

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

DONNÉES SUR LA FAMILLE CANADIENNE ET DOSSIERS FISCAUX: ÉVALUATION DE CRITÈRES QUALITATIFS ET TENDANCES DES DONNÉES

J.M. Leyes[1]

RÉSUMÉ

Statistique Canada a produit des données sur la famille à partir des dossiers fiscaux. Dans cette communication, ces données sont évaluées relativement aux données d'enquête et du recensement pour cinq critères: concepts et définitions, couverture de la population, biais de couverture de la population, biais de couverture du revenu, intervalle de variation de la variable. Pour obtenir une comparaison transversale, on a comparé aux estimations de la population de 1989 les données de 1988 sur la famille obtenues à partir des dossiers fiscaux. On a également comparé les séries chronologiques de données fiscales aux données de l'enquête sur les dépenses des consommateurs, qui est un supplément annuel de l'enquête sur la population active. Cette communication contient en outre un aperçu des orientations futures qui pourront favoriser l'amélioration de la couverture des données sur la familles produites à partir des dossiers fiscaux.

MOTS CLÉS: Biais; recensement de la population; comparaisons macro.

1. INTRODUCTION

En 1979, Statistique Canada entreprenait une étude pour évaluer les possibilités offertes par les dossiers administratifs pour la production de statistiques sociales sur les petites régions (Statistique Canada, 1979). L'évaluation a révélé que c'étaient les déclarations d'impôt des particuliers (soit les déclarations T1) qui donnaient la couverture la plus complète et qui offraient les meilleures possibilités pour la production de statistiques sociales. Dans ce document, nous allons évaluer les statistiques établies à partir des dossiers fiscaux en les comparant aux données produites à partir des deux sources suivantes: a) estimations démographiques officielles et b) résultats de l'enquête sur les finances des consommateurs (EFC).

Au moment d'entreprendre le programme d'élaboration de données, nous avons posé, comme hypothèse de départ, que les statistiques produites à l'aide des dossiers fiscaux comporteraient certaines lacunes, en particulier:

i. **Biais de couverture de la population.** Le système fiscal repose sur le revenu individuel. Étant donné qu'entre le milieu et la fin des années 70, 60% seulement des Canadiens avaient produit une déclaration d'impôt, on a jugé que la couverture des dossiers fiscaux était insuffisante pour permettre la production de statistiques sociales.

ii. **Biais dans la répartition par âge.** Le profil par âge des déclarants est différent de celui obtenu à partir des estimations démographiques officielles et le biais que cela entraîne a été jugé inacceptable.

iii. **Biais dans la répartition du revenu.** Une proportion importante de personnes âgées et de jeunes ont de faibles revenus et ne produisent pas de déclaration d'impôt. Par conséquent, les données tirées des dossiers fiscaux ne permettent pas de produire des statistiques fiables aux fins de l'exécution des programmes de l'État à l'intention de ces groupes cibles.

---

[1] J.M. Leyes, Division des données régionales et administratives, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6.

iv. **Biais de couverture du revenu.** Puisque ce ne sont pas tous les genres de revenu qui sont imposables, on a jugé que les déclarations T1 ne permettaient pas de produire des données complètes sur le revenu des Canadiens.

v. **Disponibilité des variables.** Comme les dossiers administratifs servent exclusivement à l'exécution des programmes, des variables disponibles ont été jugées inadéquates pour la production de statistiques sociales.

vi. **Concepts et définitions.** Les concepts et définitions utilisés dans les enquêtes-ménages et le recensement ne peuvent être appliqués que d'une façon approximative aux données fiscales annuelles.

Les hypothèses ci-dessus décrivent les lacunes que comporteraient les statistiques produites à partir des dossiers administratifs en général, et des déclarations T1 en particulier. Depuis le moment où ces lacunes ont été identifiées, des statistiques ont été produites sur les familles déclarantes à partir des dossiers fiscaux. Dans la section 3 et les suivantes, nous allons vérifier les trois premières hypothèses en examinant, pour les années 1982 à 1988, les statistiques sur les familles déclarantes.

2. ÉLABORATION DE DONNÉES SUR LES FAMILLES DÉCLARANTES

Statistique Canada a créé la notion de "famille déclarante" pour faire pendant à la notion de "famille de recensement", laquelle est définie de la façon suivante:

> "Époux et épouse (avec ou sans enfants jamais mariés, peu importe leur âge) ou parent seul (peu importe son état matrimonial) avec un ou plusieurs enfants jamais mariés (quel que soit leur âge) vivant dans le même logement. Aux fins du recensement, les personnes vivant en union libre sont considérées comme "actuellement mariées", peu importe leur état matrimonial légal; elles figurent donc comme une famille époux-épouse dans les tableaux sur la famille de recensement" (Statistique Canada, 1982, p. 65).

Le concept de famille de recensement peut être appliqué dans les enquêtes-ménages parce qu'on y demande au répondant d'indiquer le lien entre toutes les personnes qui habitent le logement. Pour reconstituer les familles lorsqu'on travaille avec des dossiers fiscaux, il faut se servir de renseignements secondaires (par exemple, l'état matrimonial du déclarant, le montant des exemptions et des crédits fiscaux, l'âge et l'adresse des déclarants, les dépenses pour enfants à charge, etc.).

Cela dit, on comprend aisément qu'il est impossible de produire, à partir des dossiers fiscaux, des données qui coïncident parfaitement avec le concept de famille de recensement. Les principales catégories de personnes qui posent des difficultés sont: les enfants d'âge adulte (lorsqu'ils habitent chez leurs parents, qu'ils aient ou non déjà été mariés) et les conjoints de fait. Les dossiers fiscaux permettent habituellement de reconstituer sans trop de difficulté les familles de recensement qui consistent en un couple marié avec enfants à charge. Par ailleurs, des progrès ont été faits en ce qui concerne la reconstitution des familles monoparentales et des familles dans lesquelles les partenaires sont conjoints de fait (ci-après familles de fait).

En 1984, Statistique Canada a commencé à produire des estimations sur les familles canadiennes à partir des 2données contenues dans les déclarations des particuliers (déclarations T1). La reconstitution des familles à partir des déclarations T1 a été faite en suivant les six étapes que voici:

i. Le déclarant qui donne le numéro d'assurance-sociale (NAS) de son conjoint est jumelé à ce dernier pour former une famille époux-épouse.

ii. D'autres familles époux-épouse sont formées en jumelant les déclarants qui indiquent être mariés (mais qui ne donnent pas le NAS du conjoint).

iii. Les enfants déclarants[2] qui habitent chez leurs parents sont jumelés à ces derniers.

iv. L'étape quatre est une étape intermédiaire qui sert à éliminer les doubles- comptes, à isoler les familles époux-épouse ne comptant qu'un seul déclarant et à attribuer un code postal unique aux membres de la famille de même qu'un code de composition de la famille à chaque unité familiale.

v. Les couples formés de conjoints de fait sont reconstitués par jumelage de personnes classées dans la catégorie "parent seul" et "personne hors famille".

vi. À l'étape six, on ajoute les membres non déclarants de la famille (cette opération se fait par imputation).

Maintenant que nous avons fait une brève introduction et avons expliqué en quoi consistaient les données sur les familles déclarantes, nous pouvons examiner des statistiques produites à partir des dossiers fiscaux.

3. RÉSULTATS EMPIRIQUES

3.1 Comparaison de la couverture de la population: fichier des familles déclarantes (FF-T1), 1982-1988, et estimations démographiques officielles, 1983-1989[3],[4].

Les deux tableaux dans cette section comparent les chiffres obtenus à partir des sources ci-dessus: le tableau 1 donne les chiffres de population totale et les répartitions des familles selon le genre et le tableau 2 donne les chiffres de population pour l'ensemble du Canada et pour chaque province et territoire.

3.1.1 Biais de couverture de la population: fichier des familles déclarantes (FF-T1), 1988, et des estimations démographiques officielles de Statistique Canada, 1989

Pour les besoins de la comparaison présentée dans le tableau 1, les données du fichier des familles déclarantes (FF-T1) ont été classées selon les six catégories suivantes:

- La section A (lignes 1 et 2) indique le nombre de déclarants et la variation annuelle en pourcentage de ces derniers.

- La section B (lignes 3 à 7) compare le nombre estimé de personnes dans les familles déclarantes calculé à partir du FF-T1 et le nombre de personnes dans les familles établi à partir des estimations démographiques officielles de Statistique Canada. (Lorsqu'on a construit le FF-T1, on a créé un enregistrement pour chaque membre de la famille et pour chaque personne hors famille, y compris les personnes ajoutées à une famille par imputation). Le chiffre figurant sur la ligne 3 du tableau 1 correspond donc au chiffre de population établi à partir du FF-T1).

- La section C (lignes 8 à 12) compare le nombre de familles époux-épouse établi à partir du FF-T1 aux estimations officielles des familles époux-épouse de Statistique Canada (dans l'un et l'autre cas, les familles époux-épouse comprennent les familles de fait).

---

[2] L'expression "enfants déclarants" désigne les personnes qui a) indiquent être célibataires; b) ont moins de 30 ans; c) habitent chez leurs parents et d) produisent une déclaration d'impôt.

[3] Pour réduire le coût du traitement des données du FF-T1, la plupart des statistiques sur les familles présentées dans ce document ont été établies à partir d'un échantillon d'enregistrements du fichier.

[4] La période de référence des données du FF-T1 ne coïncide pas parfaitement avec celle des estimations démographiques officielles. La meilleure comparaison est d'utiliser les données du FF-T1 pour une année d'imposition donnée et les estimations de la population de l'année suivante, par exemple, de comparer les données des FF-T1 pour 1988 aux estimations démographiques pour 1989.

- La section D (lignes 13 à 17) compare le nombre de familles monoparentales calculé à partir du FF-T1 au nombre établi à partir des estimations officielles de Statistique Canada.

- La section E (lignes 18 à 21) donne une estimation du nombre de familles de fait. La seule autre enquête des estimations des familles de fait est le recensement, qui a lieu tous les cinq ans. Par conséquent, la comparaison ne porte que sur une seule année, 1985.

- La section F donne le nombre de personnes hors famille établi à partir du FF-T1 pour chacune des années comprises durant la période 1982-1988 (la seule enquête permettant d'établir une estimation annuelle des personnes hors famille est l'EFC).

Voici les faits saillants qui se dégagent du tableau 1:

- L'accroissement de la population estimé à partir du FF-T1 (déclarants et personnes incluses par imputation) a été légèrement inférieur à l'accroissement calculé à partir des déclarations T1 seulement (tableau 1, lignes 2 et 4).

Tableau 1

Tableau sommaire: comparaison des estimations produites à partir du FF-T1 et des estimations démographiques officielles, 1982-1988

LIGNE no	ANNÉES D'IMPOSITION						
	1982	1983	1984	1985	1986	1987	1988
NOMBRE DE DÉCLARANTS (provinces et territoires)							
1. Déclarants (T1 - en milliers)	15,166	15,243	15,467	15,526	15,971	16,687	17,251
2. Variation en pourcentage		0.5	1.5	0.4	2.9	4.5	3.4
POPULATION TOTALE (provinces et territoires)							
3. Estimation d'après FF-T1 (en milliers)	23,628	23,725	23,736	23,839	24,016	24,838	25,155
4. Variation en pourcentage		0.4	0.0	0.4	0.7	3.4	1.3
5. Est. démographique (en milliers)	24,787	24,978	25,165	25,353	25,617	25,912	26,219
6. Variation en pourcentage		0.8	0.8	0.8	1.1	1.2	1.2
7. Taux de couverture (3)/(5)	95.3	95.0	94.3	94.0	93.7	95.9	95.9
FAMILLES ÉPOUX-ÉPOUSE (Canada, à l'exclusion du Yukon et des T.N.-O.)							
8. Estimation d'après FF-T1 (en milliers)	5,510	5,524	5,570	5,528	5,592	5,753	5,866
9. Variation en pourcentage		0.3	0.8	-0.7	1.1	2.9	2.0
10. Est. démographique (en milliers)	5,722	5,773	5,824	5,875	5,932	5,987	5,995
11. Variation en pourcentage		0.9	0.9	0.9	1.0	0.9	0.1
12. Taux de couverture (8)/(10)	96.3	95.7	95.6	94.1	94.3	96.1	97.9
FAMILLES MONOPARENTALES (Canada, à l'exclusion du Yukon et des T.N.-O.)							
13. Estimation d'après FF-T1 (en milliers)	830	873	894	934	940	965	937
14. Variation en pourcentage		5.1	2.4	4.5	0.6	2.7	-2.9
15. Est. démographique (en milliers)	768	796	824	852	882	912	936
16. Variation en pourcentage		3.6	3.5	3.5	3.5	3.4	2.6
17. Taux de couverture (13)/(15)	108.1	109.7	108.5	109.6	106.5	105.7	100.1
FAMILLES DE FAIT (incluses à la ligne 8, provinces et territoires)							
18. Estimation d'après FF-T1 (en milliers)	205	207	242	221	281	336	365
19. Variation en pourcentage		0.8	16.8	-8.5	27.3	19.4	8.6
20. Est. du recensement (en milliers)				487			
21. Taux de couverture (18)/(20)				45.4			
PERSONNES HORS FAMILLE (Canada, à l'exclusion du Yukon et des T.N.-O.)							
22. Estimation d'après FF-T1 (en milliers)	2,973	3,071	3,153	3,269	3,540	3,794	4,096
23. Variation en pourcentage		3.3	2.7	3.7	8.3	7.2	8.0
24. Estimation d'après l'EFC (en milliers)	3,282	3,348	3,433	3,512	3,625	3,770	3,829
25. Variation en pourcentage		2.0	2.5	2.3	3.2	4.0	1.6

Sources - FF-TI: données préliminaires, non publiées, 1982-1987, totalisations produites à partir d'un échantillon de 5%; 1988, totalisations produites à partir de l'ensemble des enregistrements du fichier. Estimations démographiques, nos 91-204, 91-210, 91-529 au catalogue.

EFC: 1982-1988, Revenus des familles. Familles de recensement, n° 13-208 au catalogue, tableau 14, janvier 1990.

- Pour la période 1982-1988, les estimations de la population du FF-T1 se situait entre 93.7 et 95.9% des estimations officielles de Statistique Canada (tableau 1, ligne 7)[5].

- Pour l'année 1988, le taux de couverture obtenu à partir du FF-T1 pour les familles époux-épouse était proche de 98% (tableau 1, ligne 12).

- On a réduit la surestimation du nombre des familles monoparentales calculé à partir du FF-T1 au cours de la période 1982-1988 (ligne 17) et, grâce à l'amélioration des méthodes de couplage des enregistrements, il a été possible d'accroître le taux de couverture des familles de fait[6].

- Pour les cinq premières années de la période considérée, l'estimation des personnes hors famille calculée à partir du FF-T1 était inférieure à l'estimation obtenue à partir des résultats de l'EFC (lignes 22 et 24). Pour les deux dernières années de la période, l'estimation du FF-T1 était supérieure à celle établie à partir de l'EFC. Les variations annuelles mises en évidence par l'une et l'autre source étaient comparables (lignes 23 et 25).

3.1.2 Comparaison de la couverture par province: fichier des familles déclarantes (FF-T1), 1988, et estimations démographiques officielles, 1989

La comparaison des estimations obtenues à partir du FF-T1 et les estimations démographiques officielles ne pourrait être complète sans examen des chiffres au niveau infra-national. Le tableau 2 compare les chiffres des dix provinces et des deux territoires pour l'année d'imposition 1988.

Tableau 2
Comparaison, par province et territoire, entre les chiffres de population produits à partir du FF-T1 pour 1988 et les estimations démographiques officielles pour 1989

PROVINCE	FF-T1 1988 (en milliers)	EST. DEM. 1989 (en milliers)	RAPPORT (en pourcentage) (FF-T1/EST. dém.)
Manitoba	1,082.4	1,084.2	99.8
Terre-Neuve	558.1	570.0	97.9
Nouveau-Brunswick	698.7	718.5	97.2
Ontario	9,230.5	9,569.5	96.5
CANADA	25,154.8	26,218.5	95.9
Québec	6,398.5	6,688.7	95.7
Nouvelle Écosse	847.2	886.8	95.5
Saskatchewan	958.7	1,007.0	95.2
Alberta	2,305.5	2,429.2	94.9
Yukon	24.0	25.4	94.5
Territoires du Nord-Ouest	50.4	53.4	94.4
Colombie-Britannique	2,878.7	3,055.6	94.2
Ile-du-Prince-Edouard	122.1	130.2	93.8

Sources - FF-T1, 1988: totalisations produites à partir de l'ensemble des enregistrements du fichier.
Estimations démographiques, 1989: estimations postcensitaires préliminaires (1er juin), n° 91-210 au catalogue, tableau 1, février 1990.

[5] Pour l'année 1986, on a utilisé une version préliminaire du fichier des déclarations d'impôt des particuliers. C'est pour cette raison que le taux de couverture du FF-T1 pour 1986 est inférieur au taux obtenu pour les autres années de la période. Si l'on n'avait pas utilisé un fichier préliminaire, le taux de couverture pour cette année aurait été supérieur à 93.7%.

[6] Grâce à l'expérience acquise, le personnel de Statistique Canada a pu améliorer sa méthode de couplage des enregistrements et plus de familles nonparentales ont été jumelées pour créer des familles de fait.

Comme le montre le tableau 2, le taux de couverture assuré par le FF-T1 se situait entre 99.8% (pour le Manitoba) et 93.8% (pour l'Ile-du-Prince-Edouard).

3.2 Biais dans la répartition par âge: comparaison des estimations produites à partir du FF-T1, 1988, et des estimations démographiques officielles, 1989

Le tableau 3 fait voir la répartition par âge de la population. Pour produire l'estimation de la population âgée de moins de 18 ans à partir du FF-T1, on s'est abondamment servi de données imputées (92%). C'est pour cette raison qu'il a été impossible de définir de plus petites tranches d'âge pour la population de moins de 18 ans ou de déterminer le sexe des enfants dans ce groupe[7].

Tableau 3
Comparaison selon la tranche d'âge: chiffres produits à partir du FF-T1, 1988, et estimations démographiques officielles, 1989

GROUPE D'ÂGE	FF-T1, 1988 (en milliers)	EST. DÉM., 1989 (en milliers)	RAPPORT (FF-T1/ EST. DÉM.)	CHIFFRES IMPUTÉS, FF-T1, 1988 (en milliers)	POURCENTAGE DES CHIFFRES DU FF-T1 IMPUTÉS
0 - 18	7,165.4	6,982.2	102.6	6,623.2	92.4
19 - 29	4,645.2	4,820.1	96.4	123.6	2.7
30 - 34	2,277.9	2,338.4	97.4	86.6	3.8
35 - 39	2,038.2	2,116.5	96.3	93.8	4.6
40 - 44	1,809.3	1,898.3	95.3	106.6	5.9
45 - 49	1,412.9	1,486.0	95.1	121.3	8.6
50 - 54	1,205.3	1,249.5	96.5	143.8	11.9
55 - 59	1,155.4	1,214.3	95.1	162.5	14.1
60 - 64	1,052.5	1,142.1	92.2	158.9	15.1
65 - 69	914.2	1,026.9	89.0	124.8	13.7
70 - 74	646.0	758.4	85.2	83.3	12.9
75+	832.3	1,185.6	70.2	75.7	9.1
TOTAL	25,154.7	26,218.5	95.9	7,904.1	31.4

Sources - FF-T1, 1988: totalisations (provinces et territoires) produites à partir de l'ensemble des enregistrements du fichier, non publiées.
Estimations démographiques, 1989: estimations postcensitaires préliminaires (1er juin), n° 21-210 au catalogue, tableau 2, février 1990 (provinces et territoires).

À noter par ailleurs que pour produire l'estimation du nombre d'enfants vivant chez leurs parents à partir du FF-T1, on a posé qu'il n'y avait pas, dans les familles, d'enfants de plus de 29 ans.

Dans la colonne 4 du tableau 3 (Rapport), on peut voir que dans les groupes d'âge de moins de 65 ans, le taux de couverture était de 90%[8] ou mieux.

Voici les faits saillants qui se dégagent du tableau 3:

- Il y a surestimation de l'ordre de 2.6% de la population de moins de 19 ans.

---

[7] Les déclarations T1 peuvent nous fournir certains éléments d'information sur les enfants à charge, notamment : lien avec le déclarant et date de naissance. A noter cependant que les déclarations T1 renferment la date de naissance de tous les enfants pour lesquels des Allocations familiales sont versées (soit 85% environ des enfants au Canada). Un projet est actuellement en cours pour attribuer un âge aux enfants inclus dans les familles par imputation.

[8] Le pourcentage de déclarants chez les 65 ans et plus est passé de 55.6% en 1985 à 70.9% en 1988.

- Pour les 6 groupes d'âge entre 19 et 59 ans, le taux de couverture le plus bas a été enregistré dans la tranche des 45 à 49 ans (95.1%) et le taux le plus élevé (97.4%) dans la tranche des 30 à 34 ans.

- Dans les tranches pour les personnes qui ont 60 ans ou plus, le taux de couverture diminue progressivement pour s'établir à 70.2% pour les 75 ans et plus. (Bien qu'on ne l'indique pas dans ce tableau, le taux global de couverture de la population de 65 ans et plus était d'environ 80% pour l'année 1988).

3.3 Biais dans la répartition du revenu: comparaison des estimations produites à partir du FF-T1 et des résultats de l'EFC

Considérons la troisième hypothèse dont nous avons fait état à la page 1. On prévoyait que les données sur le revenu produites à partir des dossiers fiscaux seraient biaisées. Les raisons de ce biais sont les suivantes:

a. Dans la population en général, on trouve un certain nombre de personnes à faible revenu qui ne produisent pas de déclaration d'impôt, et,

b. par conséquent, la mesure de la tendance centrale du revenu obtenue à partir du FF-T1 sera supérieure à celle que l'on obtiendrait pour la population totale.

Pour évaluer cette hypothèse, nous avions la possibilité de comparer les estimations produites à partir du FF-T1 soit aux chiffres du recensement de la population soit aux résultats de l'EFC, qui est une enquête annuelle. Et comme le recensement ne nous fournissait des données que pour une seule année, 1986, contre 7 années dans le cas de l'EFC, nous avons décidé de comparer les chiffres du FF-T1 aux résultats de l'EFC.

Dans la comparaison nous examinons deux groupes: les familles et les personnes hors familles (ou personnes seules).

3.3.1 Comparaison du revenu médian des familles: estimations produites à partir du FF-T1 et estimations de l'EFC, 1982-1988

Le tableau 4 donne des séries chronologiques du revenu médian des familles, pour la période 1982-1988, estimé à partir du FF-T1 et l'EFC[9].

Voici les faits saillants qui se dégagent du tableau 4:

- Pour toutes les années de la période, le revenu médian estimé à partir du FF-T1 est inférieur à l'estimation de l'EFC.

- Pour la période 1982-1985, le revenu médian estimé à partir du FF-T1 correspond à 95% de l'estimation de l'EFC.

- Pour l'année d'imposition 1986, année au cours de laquelle le crédit pour taxe fédérale sur les ventes a été introduit, et à nouveau pour l'année 1987, le revenu médian estimé à partir du FF-T1 ne représentait plus que 92% de l'estimation de l'EFC. (Ce crédit d'impôt a été conçu à l'intention des personnes à faible revenu et des personnes qui ont des revenus non imposables. Rappelons que nous avons au départ posé comme hypothèse que ce serait pour ces deux groupes que la couverture du FF-T1 serait la plus facile).

- En 1988, le taux de couverture du revenu est passé à 95.6%, soit le taux annuel le plus élevé dans cette série chronologique.

---

[9] L'EFC est un supplément annuel de l'enquête sur la population active. L'EFC est semblable au supplément de mars de l'enquête américaine Current Population Survey (CPS).

Tableau 4

Comparaison du revenu des familles, 1982-1988: FF-T1 et l'EFC

ANNÉE	REVENU MÉDIAN				RAPPORT EN POURCENTAGE (FF-T1/EFC)
	FF-T1	AUGMENTATION EN POURCENTAGE	EFC	AUGMENTATION EN POURCENTAGE	
1982	28,154		29,537		95.3
1983	28,806	2.3	30,419	3.0	94.7
1984	30,603	6.2	32,079	5.5	95.4
1985	32,140	5.0	33,950	5.8	94.7
1986	INTRODUCTION DU CRÉDIT POUR TAXE FÉDÉRALE SUR LES VENTES				
1986	33,135	3.1	36,019	6.1	92.0
1987	35,279	6.5	38,059	5.7	92.7
1988	38,668	9.6	40,430	6.2	95.6
Augmentation en pourcentage, 1982-88		37.3		36.9	

Sources - FF-T1: Pour les années 1982 à 1987, les totalisations ont été produites à partir d'un échantillon de 5%. Pour 1988, les totalisations ont été produites à partir de l'ensemble des enregistrements du fichier. (Les territoires ont été exclus de la comparaison.)

EFC: Les données des années 1982 à 1988 ont été tirées de la publication annuelle n° 13-208 au catalogue.

Bien que la série chronologique ci-dessus ne permette pas de tirer des conclusions définitives, elle donne à penser que les données fiscales assurent une couverture plus exhaustive de la population à faible revenu qu'on ne s'y attendait. Il est clair que l'accroissement du nombre de déclarants chez les personnes à faible revenu est attribuable à l'introduction de deux programmes de crédits d'impôt. Le programme du crédit d'impôt pour enfants, entré en vigueur en 1978, s'adresse aux familles à faible revenu qui ont des enfants et le programme du crédit pour taxe fédérale sur les ventes, entré en vigueur en 1986, à toutes les personnes à faible revenu. Par ailleurs, on prévoit que l'introduction du crédit pour la taxe sur les produits et services va se traduire par un autre bond dans le nombre de déclarants à faible revenu parce que le montant de ce crédit, qui est semblable au crédit pour taxe fédérale sur les ventes, est encore plus élevé[10].

3.3.2 Comparaison du revenu médian des personnes hors famille, FF-T1 et l'EFC, 1982-1988:

Le tableau 5 contient les données produites à partir des deux sources.

Voici les faits saillants qui se dégagent du tableau 5:

- Pour les années 1982 à 1985, l'estimation de l'EFC du revenu médian des personnes hors famille était, comme on s'y attendait, inférieur à l'estimation du FF-T1. Pour cette période, le revenu médian des personnes hors famille estimé à partir du FF-T1 représentait à 117.1 à 124.2% de l'estimation de l'EFC.

[10] La dernière révision du crédit pour taxe fédérale sur les ventes en a porté le montant à $70 par personne admissible. D'après les renseignements que nous avons obtenus, le montant du crédit pour de taxe sur les produits et services sera de $380 par trimestre (à souligner cependant que ce montant peu varier dans certaines circonstances).

- Avec l'introduction du crédit pour taxe fédérale sur les ventes en 1986, ce pourcentage a baissé à 104.8% en 1986 puis à 99.4% en 1987, et il a grimpé à nouveau en 1988, pour atteindre 105.8%.

Pour les personnes hors famille, notre hypothèse semble avoir été confirmée. Par ailleurs, l'effet observé en ce qui concerne l'introduction du crédit pour taxe fédérale sur les ventes renforce l'hypothèse: on avait prévu qu'avec l'introduction du crédit, un plus grand nombre de personnes à faible revenu produiraient une déclaration d'impôt, ce qui viendrait abaisser le revenu médian estimé à partir des données fiscales par rapport à l'estimation de l'EFC.

Tableau 5
Comparaison du revenu des personnes hors famille, 1982-1988: FF-T1 et l'EFC

ANNÉE	REVENU MÉDIAN				RAPPORT EN POURCENTAGE (FF-T1/EFC)
	FF-T1	AUGMENTATION EN POURCENTAGE	EFC	AUGMENTATION EN POURCENTAGE	
1982	12,314		10,246		120.2
1983	12,116	-1.6	9,757	-4.8	124.2
1984	12,759	5.3	10,842	11.1	117.7
1985	13,330	4.5	11,383	5.0	117.1
1986	INTRODUCTION DU CRÉDIT POUR TAXE FÉDÉRALE SUR LES VENTES				
1986	12,966	-2.7	12,371	8.7	104.8
1987	13,235	2.1	13,317	7.6	99.4
1988	14,689	11.0	13,880	4.2	105.8
Augmentation en pourcentage 1982-1988		19.3		35.5	

Sources - FF-T1: - Pour les années 1982 à 1987, les totalisations ont été établies à partir d'un échantillon de 5%. Pour 1988, les totalisations ont été faites à partir de l'ensemble des enregistrements du fichier (les territoires ont été exclus de la comparaison).

EFC: - Les données des années 1982 à 1988 ont été tirées de la publication annuelle n° 13-208 au catalogue.

4. PRINCIPAUX PROJETS POUR 1989 ET LES ANNÉES À VENIR

Deux projets touchant à l'exploitation des données fiscales sont actuellement en cours:

- En 1989-1990, Statistique Canada a commencé à construire une base de données administratives longitudinales. La création de cette base a pour but de faciliter les travaux de recherche sur la variation longitudinale de la pauvreté, le bien-être et le revenu au Canada durant la période 1982-1986. La fraction de sondage choisie pour la création de la base de données est de 10%, ce qui correspond à la fraction retenue pour l'enquête par panel intitulée Panel Survey of Income Dynamics (PSID) conçue par l'University of Michigan il y a environ 20 ans (Duncan, 1984).

- Des travaux sont actuellement en cours afin de construire un fichier d'intégration de données administratives au moyen d'un couplage d'enregistrements multiples prélevés par échantillonnage. Ce fichier permettra la modélisation des données sur les familles déclarantes pour a) améliorer la couverture de la population et b) améliorer la qualité de données sur certaines des variables (par exemple, l'âge des enfants, les raisons pour lesquelles les chômeurs ont reçu des prestations d'assurance-chômage).

5. RÉSUMÉ ET COMMENTAIRES

Les données que renferme le fichier des familles (FF-T1) présentent des caractéristiques intéressantes: il s'agit de données annuelles qui permettent la production de statistiques sur les petites régions.

Pour ce qui est de la qualité globale des statistiques produites à partir du FF-T, si l'on considère qu'un taux de couverture de 95% est élevé, les comparaisons présentées dans ce document montrent que le fichier assure un couverture exhaustive de la population. En outre, il semble que la représentativité des statistiques produites ait progressivement augmenté durant la période 1982-1988. A noter toutefois qu'il continue d'y avoir sous-estimation des personnes à faible revenu (c'est-à-dire les jeunes et plus particulièrement les personnes âgées), bien qu'une amélioration de la couverture de la population à faible revenu a découlé de l'introduction, en 1986, du crédit pour taxe fédérale sur les ventes. Enfin, d'après l'expérience, il semble raisonnable de penser que l'introduction du crédit pour taxe sur les produits et services va permettre, dans les prochaines années, une amélioration de la couverture de la population en général, et de la population à faible revenu en particulier.

BIBLIOGRAPHIE

Duncan, G.J. (1984). Years of Poverty, Years of Plenty, Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan.

Statistique Canada (1982). *Dictionnaire du recensement, 1981*, Statistique Canada, 1981, Recensement de la population, n° 99-901 au catalogue, Ottawa, Canada.

Statistique Canada (1988). *Estimations intercensitaires des familles, Canada et provinces, 1981-1986*, n° 91-529 au catalogue, Ottawa, Canada.

Statistique Canada (1990). *Estimations postcensitaires des familles, Canada et provinces, 1987, 1988, 1989*, n° 91-204 au catalogue, Ottawa, Canada.

Statistique Canada (1990). *Estimations postcensitaires de la population, suivant l'état matrimonial, l'âge, le sexe et la composante de croissance, Canada, provinces et territoires, 1989*, n° 91-210 au catalogue, Ottawa, Canada.

Statistique Canada (1990). *Revenu des familles - Familles de recensement, 1988, Enquête sur les finances des consommateurs*, n° 13-208 au catalogue (annuelle), Ottawa, Canada.

SESSION 7

Amélioration de la qualité des bases de sondage et de leur utilisation

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

L'ÉCHANTILLONNAGE DES FLUX DE POPULATIONS HUMAINES MOBILES

G. Kalton[1]

RÉSUMÉ

On fait souvent des enquêtes sur des flux de populations de personnes comme celles qui vont dans les musées, les bibliothèques et les parcs; les électeurs; les personnes qui font des courses; les malades en consultation externe; les touristes; les voyageurs venant de l'étranger et les passagers des automobiles. Les plans d'échantillonnage d'enquêtes de ce genre prévoient en général un échantillonnage dans le temps et dans l'espace. Cette communication passe en revue et illustre les méthodes utilisées pour sonder les flux de populations humaines.

MOTS CLÉS: Populations mobiles; sondages des votants; enquêtes de circulation; échantillonnage dans le temps et dans l'espace; échantillonnage systématique.

1. INTRODUCTION

La majorité des enquêtes portant sur les populations humaines sont basées sur les ménages, habituellement avec un échantillon de ménages choisi au moyen d'un plan d'échantillonnage à plusieurs degrés et des particuliers sont choisis dans les ménages sélectionnés. L'enquête-ménage est une méthode puissante employée pour recueillir des données sur une gamme étendue de caractéristiques relatives à la population, comme des caractéristiques sociales, démographiques et économiques ainsi que des particularités en matière de santé, tout comme les opinions et attitudes de la population. Toutefois, la méthode n'est pas aussi efficace pour étudier les caractéristiques de populations mobiles. On peut distinguer deux types de populations mobiles: les personnes qui n'ont pas de domicile fixe habituel, comme les nomades et les sans abris ainsi que les membres de l'ensemble de la population qui font partie de la population mobile à l'étude parce qu'ils sont de passage, comme les personnes qui vont dans des bibliothèques et des parcs, les électeurs aux bureaux de scrutin, les personnes qui font des courses, les malades en consultation externe, les voyageurs et les passagers des automobiles. Dans la présente communication on passe en revue certaines questions portant sur les plans d'échantillonnage utilisés pour cette dernière catégorie de population mobile.

Bien que de nombreuses enquêtes portent sur les flux de populations humaines mobiles, les ouvrages généraux portant sur l'échantillonnage traitent peu des questions relatives à l'échantillonnage dans ces cas particuliers. L'objet de la présente communication est de décrire les plans d'échantillonnage généralement adoptés pour les enquêtes sur les flux de population humaines, de discuter de certaines des questions particulières qui se posent en matière d'échantillonnage et d'illustrer la gamme d'applications pour de telles enquêtes. Dans la prochaine section de la communication nous passons en revue le plan général d'échantillonnage dans le temps et dans l'espace utilisé pour échantillonner des personnes de passage et certaines des questions portant sur l'emploi de ce plan d'échantillonnage dans des situations particulières. La section 3 illustre ensuite l'application du plan d'échantillonnage dans une gamme de situations différentes. La section 4 renferme des conclusions.

2. ÉCHANTILLONNAGE DANS LE TEMPS ET DANS L'ESPACE

Il sera utile de considérer un exemple particulier pour décrire le plan d'échantillonnage général dans le temps et dans l'espace visant à échantillonner des flux de populations humaines. Supposons qu'on doit réaliser une

[1] G. Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan 48106-1248, U.S.A.

enquête auprès des visiteurs d'une exposition de sculptures tenue pendant l'été dans un parc d'une ville afin de trouver les caractéristiques socio-économiques des visiteurs, pour déterminer comment ils ont eu connaissance de l'exposition, quels moyens de transport ils ont utilisés pour se rendre au parc et, peut-être, leurs opinions à propos de l'exposition. Supposons que l'exposition se déroule du 1er avril au 30 septembre de l'année en question, qu'elle est ouverte de 10 h à 18 h tous les jours et qu'il y a trois endroits où les visiteurs peuvent entrer et sortir du terrain de l'exposition.

On considère généralement que la base de sondage employée pour une enquête de ce genre est une liste d'unités primaires d'échantillonnage (UPÉ) fondée sur l'intervalle de temps/l'endroit. On construit cette base de sondage en divisant la période de temps de l'enquête en un ensemble d'intervalles de temps pour chaque endroit. Une façon simple de construire les UPÉ pour l'exemple que nous étudions consisterait à diviser chaque jour d'exposition à chaque endroit en deux intervalles de temps, un allant de 10 h à 14 h et l'autre de 14 h à 18 h. Une façon plus complexe de construire les UPÉ pourrait comprendre l'utilisation d'intervalles de temps de longueur différente pour des jours différents et (ou) à des endroits différents. Une fois les UPÉ définies, on emploie souvent un plan d'échantillonnage à deux degrés. Dans la première étape, un échantillon d'UPÉ est choisi et, dans la deuxième étape, un échantillon de visiteurs est tiré, habituellement par échantillonnage systématique, dans les UPÉ échantillonnées.

Les spécifications réelles du plan d'échantillonnage utilisé pour une enquête visant les personnes de passage et qui utilise la base de sondage à deux degrés dépendent des caractéristiques de la population mobile à l'étude ainsi que des procédures utilisées pour faire la collecte des données d'enquête. Une caractéristique clé est la nature du flux de la population mobile. En particulier, y a-t-il une variabilité prévisible dans le taux du flux parmi les UPÉ? Par exemple, le flux à un endroit est-il plus élevé qu'à un autre, ou les flux pour certains intervalles de temps (disons les samedis dans l'après-midi) sont-ils plus élevés que pour d'autres? De plus, le flux dans une UPÉ est-il continu pendant tout l'intervalle de temps ou est-il inégal, avec des visiteurs qui arrivent (ou qui partent) en groupes importants? Ces deux aspects du flux ont un effet sur le plan d'échantillonnage utilisé pour l'enquête.

Si le flux est assez uniforme parmi les UPÉ et si les intervalles de temps dans les UPÉ sont identiques, alors le nombre de visiteurs par UPÉ est approximativement constant. Dans ce cas, on peut échantillonner les UPÉ avec des probabilités égales et appliquer un taux de sous-échantillonnage constant dans les UPÉ choisies pour produire un échantillon avec probabilité égale, ou "msepe", de visites. Les UPÉ peuvent être classées en deux dimensions ou plus (p. ex., le jour de la semaine, l'heure du jour et l'endroit) et on peut obtenir un échantillon bien équilibré entre ces dimensions à l'aide de l'échantillonnage de configuration (Yates, 1981; Cochran, 1977 et Jessen, 1978).

Dans nombre de cas, le niveau du flux varie entre les UPÉ d'une façon qui est partiellement prévisible. Par exemple, on peut savoir que la fréquentation de l'exposition de sculptures est généralement plus élevée lors du dernier poste du travail de chaque jour et au cours des fins de semaine et qu'elle est particulièrement faible les lundis. Ainsi, les UPÉ comprennent un nombre différent de visiteurs, c'est-à-dire, qu'ils sont de taille différente. La procédure habituelle utilisée pour traiter des UPÉ de taille différente consiste à les échantillonner avec des probabilités proportionnelles à leur taille (PPT), ou avec des probabilités proportionnelles à leur taille estimée (PPTE). Dans le contexte actuel, la taille réelle des UPÉ n'est pas connue à l'avance, il faut donc utiliser des tailles estimées. L'échantillonnage des UPÉ avec PPTE donne de bons résultats pourvu qu'on puisse faire des estimations raisonnables de leur taille. Quand des UPÉ sont choisies par échantillonnage avec PPTE, l'application dans les UPÉ choisies de taux de sous-échantillonnage qui sont inversement proportionnelles aux tailles estimées des UPÉ produit un échantillon "msep" global des visites. En général, un des attraits de l'échantillonnage avec PPTE (avec des estimations raisonnables de la taille) est que la taille des sous-échantillons des UPÉ ne varie pas beaucoup d'une UPÉ à l'autre. Cette caractéristique est particulièrement intéressante pour le travail sur le terrain lors d'enquête portant sur les personnes de passage. Quand des UPÉ portant sur le temps/l'endroit sont obtenues par échantillonnage avec PPTE, on ne peut appliquer l'échantillonnage de configuration pour une stratification en profondeur. On peut plutôt employer une sélection contrôlée à cette fin (Goodman et Kish, 1950; Hess et coll., 1975).

Une considération importante dans tout plan d'échantillonnage à deux degrés est l'affectation de l'échantillon entre les unités primaires et les unités secondaires d'échantillonnage, c'est-à-dire, combien d'UPÉ choisir et

combien d'éléments choisir dans chaque UPÉ sélectionnée. Dans le cas des enquêtes sur les personnes de passage, les procédures à utiliser sur le terrain ainsi que la nature du flux à l'intérieur des UPÉ ont un effet considérable sur cette affectation. Le but du plan d'échantillonnage est d'utiliser pleinement les travailleurs sur le terrain affectés à une UPÉ sélectionnée tout en maintenant un échantillon probabiliste des personnes qui entrent dans l'endroit (ou qui en sortent) pendant l'intervalle de temps échantillonné.

Dans de nombreuses enquêtes portant sur les personnes de passage on utilise des questionnaires à remplir soi-même; dans ce cas, le travail sur le terrain pour le plan d'échantillonnage à deux degrés décrit plus haut est composé du comptage des personnes au moment où elles entrent dans l'endroit échantillonné (ou lorsqu'elles en sortent) pendant l'intervalle de temps, du choix de chaque k^e personne pour un échantillon systématique et du fait de demander aux personnes sélectionnées de remplir le questionnaire. Si le flux n'est pas dense et s'il est réparti également pendant l'intervalle de temps, il se peut qu'un seul travailleur sur le terrain puisse remplir toutes les tâches. Quand c'est le cas, l'intervalle d'échantillonnage k peut être choisi de façon à donner à ce travailleur suffisamment de temps pour effectuer toutes les tâches sans travailler à la limite de ses possibilités. Si, toutefois, le flux est dense, que ce soit de façon constante ou intermitente, il se peut que l'on doive employer deux travailleurs, un qui n'aurait qu'à compter les entrants (ou les sortants) et à déterminer les personnes échantillonnées et le second pour remettre les questionnaires et dire aux répondants comment ils doivent le remplir et le retourner. Quand le travail sur le terrain est organisé de cette façon, on peut choisir l'intervalle d'échantillonnage pour que le second travailleur soit pleinement occupé, tout en s'assurant qu'il est en mesure de distribuer le questionnaire à toutes (ou presque toutes) les personnes choisies. La non-réponse peut-être une préoccupation importante quand la collecte des données se fait à l'aide d'un questionnaire à remplir soi-même. Il est souvent possible de conserver la non-réponse à un niveau acceptable quand les personnes choisies remplissent et remettent le questionnaire sur place. Toutefois, quand on leur remet le questionnaire en leur demandant de le remplir plus tard et de le retourner par la poste, le niveau de non-réponse peut-être très élevé et, de plus, il n'existe généralement pas de façon d'effectuer le suivi des non-répondants.

Quand on a recours à des interviews comportant l'interrogation directe pour recueillir des données, l'équipe chargée du travail sur le terrain pour une UPÉ comprend généralement un compteur et un petit groupe d'intervieweurs. La taille de l'équipe d'intervieweurs dépend de la régularité du flux et de la durée de l'interview. Puisque les personnes de passage ne voudront vraisemblablement pas être retardées pendant une longue période, la majorité des interviews seront nécessairement courtes. Il est toutefois possible de réaliser des interviews plus longues si les personnes échantillonnées sont en attente, comme lorsqu'elles sont dans une file d'attente ou dans une salle d'attente d'un aéroport. Le choix de l'intervalle d'échantillonnage doit être tel qu'il y a toujours (ou presque toujours) un intervieweur libre pour interviewer la prochaine personne échantillonnée et que les intervieweurs ne passent pas trop de temps à attendre que la prochaine personne à interviewer soit choisie. Si le flux est irrégulier, on doit prendre les dispositions nécessaires pour traiter les périodes de pointe (par exemple, l'arrivée, à l'exposition de sculptures, d'un autocar rempli de visiteurs).

La sélection avec PPTE des UPÉ permet de rendre égale la taille des sous-échantillons pour chaque UPÉ échantillonnée. Dans les enquêtes réalisées par interrogation directe, la charge de travail de l'intervieweur est ainsi à peu près la même pour chaque UPÉ sélectionnée et, par conséquent, la taille de l'équipe d'intervieweurs affectée à chaque UPÉ peut être la même. Il se produit toutefois un problème quand la mesure de la PPTE utilisée pour sélectionner l'UPÉ lors de la première étape est entachée d'une erreur considérable. Par exemple, un orage peut réduire substantiellement le nombre de visiteurs de l'exposition de sculptures un samedi après-midi en particulier ou, un congé non prévu peut augmenter considérablement le nombre de visiteurs un autre jour. Dans le premier cas, le fait d'appliquer dans cette UPÉ un intervalle d'échantillonnage inversement proportionnel à sa taille estimée laissera les intervieweurs très souvent inoccupés; alors que, dans le second cas, cela entraînera une charge de travail que les intervieweurs ne pourront pas traiter. Une modification qui peut être adoptée dans de tels cas consiste à changer l'intervalle d'échantillonnage au début de la collecte des données pour employer un intervalle qui est plus approprié au flux réel des visiteurs. Puisque cette modification détruit la propriété "msep" de l'échantillon, il faut utiliser des poids dans l'analyse de l'enquête.

Une restriction générale qui s'applique à l'échantillonnage systématique des visiteurs pour des UPÉ sélectionnées est que, si la longueur de l'intervalle d'échantillonnage est rendue suffisante pour permettre aux intervieweurs de traiter les flux lors des périodes de pointe, ces derniers passent une bonne partie de leur temps sans travail. Par contre, si l'intervalle d'échantillonnage est réduit, les intervieweurs sont plus occupés, mais ils ne peuvent

traiter les flux lors des périodes de pointe. Diverses méthodes ont été proposées pour surmonter ces problèmes (Heady, 1985). Une méthode consiste à prendre un échantillon systématique d'époques (disons à toutes les 10 minutes) et de choisir le prochain visiteur qui entre après chaque époque échantillonnée. Cette méthode pourrait présenter un certain intérêt sur le terrain, mais elle ne produit pas un échantillon probabiliste des visiteurs. La probabilité d'être choisie des personnes qui arrivent lors des périodes de grande activité est moindre, comme c'est le cas pour les personnes qui voyagent en groupe et les habitudes de déplacement de ces dernières peuvent avoir un effet inconnu sur la probabilité qu'ont ces personnes d'être sélectionnées. L'échantillon produit par cette procédure n'est certainement pas un échantillon "msep". On peut tenter de compenser le biais dû à la sélection qui défavorise les visiteurs arrivant pendant les périodes de grande activité en divisant l'intervalle de temps pour des UPÉ sélectionnées en un ensemble d'intervalles beaucoup plus courts et en tenant un journal des arrivées dans chacun de ces intervalles. On peut alors employer la pondération pour compenser la variation dans le flux pendant les intervalles plus courts.

Une autre méthode qui peut être employée à la place de l'échantillonnage systématique des visiteurs consiste à choisir la prochaine personne à entrer (ou à sortir) après la fin de la dernière interview. Dans cette méthode, les premières personnes qui arrivent après des intervalles dans le flux de visiteurs, peut-être les chefs de groupes, ont évidemment des chances plus élevées d'être choisies. Il se peut aussi que les intervieweurs accélèrent ou ralentissent délibérément l'interview qu'ils sont à réaliser afin d'éviter ou de choisir une personne en particulier. Pour ces raisons, on a employé des variantes de cette méthode dans le cadre desquelles on choisit la $n^{\text{ième}}$ personne qui entre ou qui sort après la fin d'une interview, où n peut avoir comme valeur 2, 3, 4, ou 5. Toute version de cette procédure produit, toutefois, un échantillon non probabiliste, avec le risque de biais que cela entraîne. Ces méthodes qui peuvent être utilisées à la place d'un échantillonnage systématique simple des visiteurs peuvent permettre d'utiliser beaucoup plus efficacement le temps des travailleurs sur le terrain, mais il faut sacrifier quelque chose quand on passe d'un plan d'échantillonnage probabiliste à un plan d'échantillonnage non probabiliste.

Les visiteurs peuvent être échantillonnés soit au moment où ils entrent dans un endroit, soit au moment où ils le quittent. Si l'on recherche des données sur les activités des visiteurs dans cet endroit et sur leurs opinions à propos de cet endroit, l'échantillon doit être composé seulement des personnes qui quittent l'endroit. Dans d'autres cas, le choix entre le fait d'échantillonner les personnes qui entrent dans un endroit ou qui en sortent peut dépendre de la nature des flux de populations humaines. Il peut, par exemple, être difficile d'échantillonner et d'interviewer des personnes qui quittent un cinéma ou une salle de théâtre parce que tous les spectateurs quittent les lieux en masse et qu'ils ne voudront pas être retardés. Par contre, ces personnes peuvent être échantillonnées et interviewées facilement quand elles font la queue pour entrer dans le cinéma ou la salle de théâtre.

En concluant cette section, il faudrait attirer l'attention sur le fait que les échantillons décrits ici sont des échantillons de visites et non de visiteurs. Le plan d'échantillonnage standard à deux degrés peut produire un échantillon "msep" de visites, mais ce n'est pas la même chose qu'un échantillon "msep" de visiteurs à moins que chaque visiteur ne visite l'endroit étudié (l'exposition de sculptures) qu'une fois (ou que tous les visiteurs visitent l'endroit étudié le même nombre de fois). Pour la majorité des enquêtes sur les flux de populations humaines, la visite, plutôt que le visiteur, est l'unité d'analyse appropriée. Il y a, toutefois, des situations où l'unité d'analyse est problématique. S'il utilise la visite comme unité d'analyse, le chercheur pourrait facilement accepter des visites à l'exposition de sculptures lors de deux jours différents comme des visites distinctes, mais il pourrait ne pas accepter de traiter deux entrées la même journée (une, peut-être, après être sorti pendant une brève période pour des rafraîchissements) comme deux visites. L'utilisation du visiteur comme unité d'analyse présente des problèmes graves à cause de la difficulté que présentent les visites multiples et du fait que les visiteurs ne pourront pas déclarer leurs visites multiples. Il se peut qu'ils puissent se souvenir, avec assez de précision, de leurs visites antérieures, mais habituellement ils ne pourront prévoir, avec précision, leurs visites futures.

3. EXEMPLES

La présente section renferme certains exemples d'enquêtes sur les flux de populations humaines afin de montrer la gamme étendue d'applications et pour illustrer certaines des considérations spéciales inhérentes à des situations particulières.

3.1 Une enquête sur l'utilisation des bibliothèques

Une enquête sur l'utilisation des 18 bibliothèques de la University of Michigan a été réalisée en 1984 (Heeringa, 1985). On a demandé à chaque personne échantillonnée sortant d'une bibliothèque si elle avait utilisé les documents et les services de la bibliothèque pendant cette visite. Dans l'affirmative, on demandait à la personne de remplir un bref questionnaire de sept questions portant sur les documents consultés et sur les services utilisés. La majorité des 5 184 répondants ont rempli le questionnaire sur place et l'ont remis aux travailleurs réalisant l'enquête sur le terrain; d'autres les ont retournés par l'intermédiaire du service de messageries de l'université. Un taux de réponse de 96% a été obtenu.

Le plan d'échantillonnage suivait le plan d'échantillonnage à deux degrés selon le temps/l'endroit décrit dans la section 2. L'enquête visait toute l'année civile 1984. Chaque jour où les bibliothèques étaient ouvertes a été divisé en dix intervalles de temps de deux heures, de 7 h 30 jusqu'à 3 h 30 le matin du jour suivant, l'intervalle de deux heures étant choisi parce qu'il constituait un poste de travail approprié pour les personnes réalisant l'enquête sur le terrain. Les UPÉ ont alors été définies comme des combinaisons d'intervalles de temps/de bibliothèques. Les UPÉ ont été sélectionnées par échantillonnage avec PPTE, où la taille estimée pour une UPÉ était le nombre estimé de personnes sortant de cette bibliothèque dans la période de temps précisée. Des estimations grossières de ces nombres ont été obtenues à partir de la fréquentation quotidienne moyenne en novembre 1983, selon les chiffres enregistrés par les tourniquets quand ils étaient disponibles et des estimations fournies par les bibliothécaires quand ce n'était pas le cas et en fonction d'une hypothèse selon laquelle le volume des sorties des bibliothèques était deux fois plus élevé entre 9 h 30 et 17 h 30 qu'en d'autres temps. Les bibliothèques ont été stratifiées en quatre types et, dans chaque strate, on a utilisé une sélection contrôlée pour obtenir une distribution proportionnelle de l'échantillon parmi les bibliothèques, les jours de la semaine et les intervalles de temps.

Dans chaque UPÉ sélectionnée, on a choisi pour l'enquête un échantillon systématique de personnes sortant de la bibliothèque, avec l'intervalle d'échantillonnage déterminé afin de donner un échantillon global "msep" des visites. On a fourni aux travailleurs sur le terrain une feuille d'inscription sur laquelle figuraient les entiers de 1 à 430 et où les numéros sélectionnés étaient marqués. Tout ce que les travailleurs avaient à faire était de cocher un numéro pour chaque personne sortant de la bibliothèque et de choisir les personnes associées aux numéros échantillonnés. Cette méthode présente l'avantage que les pas d'échantillonnage fractionnaires sont faciles à traiter. Quand on prévoyait que le volume des sorties pour une UPÉ échantillonnée devait être faible, un seul travailleur devait faire le comptage et communiquer avec les personnes échantillonnées. Quand le volume des sorties était élevé, le travail était réparti entre deux travailleurs, un qui effectuait le comptage et l'autre qui communiquait avec les personnes échantillonnées. Il fallait aussi employer plus d'un travailleur dans les bibliothèques disposant de plus d'une sortie.

3.2 Une enquête sur les visites dans un musée

Une enquête par interrogation directe des visiteurs sortant du National Air and Space Museum à Washington, D.C. a été réalisée de la mi-juillet jusqu'en décembre 1988 (Doering et Black, 1989). L'interview, d'une durée d'environ quatre à six minutes, visait à recueillir des données sur les antécédents socio-démographiques de la personne échantillonnée, sur son lieu de résidence, sur ses activités lors de la visite, sur les objets exposés qui l'ont particulièrement intéressée, sur la raison de sa visite, sur la taille et le type de groupe, si elle faisait partie d'une visite en groupe et sur le mode de transport utilisé. Les enfants de moins de 12 ans et les personnes travaillant au musée étaient exclus de l'enquête. Les données ont été recueillies auprès de 5 574 répondants, avec un taux de réponse de 86%.

Chaque jour de la période d'enquête était divisé en deux demi-journées. Les interviews étaient réalisés pendant l'une de ces demi-journées à tous les deux jours, avec alternance entre les matinées et les après-midis. Pendant

l'été, le public pouvait utiliser trois sorties du musée, alors que plus tard au cours de l'année seulement deux d'entre elles étaient ouvertes. Pendant les demi-journées sélectionnées, la collecte des données d'enquête se faisait par rotation, sur une base horaire, entre les sorties qui étaient ouvertes.

L'équipe de travailleurs sur le terrain affectée à une sortie à une heure échantillonnée était composée d'un ou de deux compteurs et de deux intervieweurs. Le chef compteur utilisait un compteur mécanique ainsi qu'un chronomètre pour suivre le nombre de personnes sortant du musée et pour tenir un registre qui donnait le nombre de personnes sortant dans chaque intervalle de dix minutes pendant l'heure. Le chef compteur identifiait aussi les personnes à interviewer. Le choix des personnes échantillonnées était fait pour que les intervieweurs ne soient jamais inoccupés. Le chef compteur remarquait quand un intervieweur avait terminé une interview et était prêt à en commencer une autre et il choisissait alors la cinquième personne sortant après ce moment comme la prochaine personne échantillonnée. Les comptes des flux de personnes pendant une période de dix minutes étaient utilisés dans l'analyse pour élaborer des poids afin de compenser pour la variation dans la sélection aléatoire associée au flux variable de personnes dans le temps.

La distinction entre la "visite" et le "visiteur" est particulièrement saillante pour cette enquête. Les personnes pouvaient, bien entendu, visiter le musée pendant plusieurs jours au cours de la période de l'enquête et elles pouvaient aussi visiter le musée plusieurs fois pendant la même journée. Cette dernière possibilité est particulièrement probable dans le cas du National Air and Space Museum parce que l'entrée au musée est gratuite et, par conséquent, rien n'incite les visiteurs à entrer une seule fois. Compte tenu de cette situation, il pourrait être approprié de définir les entrées multiples au cours d'une même journée comme une seule visite pour certaines fins analytiques. Dans certains cas, cette définition pourrait être appliquée en limitant l'analyse aux personnes qui sortent du musée pour la première fois le jour échantillonné.

3.3 Sondages des votants

Un certain nombre des principales agences de presse réalisent des sondages auprès des électeurs les jours d'élection aux États-Unis (Levy, 1983; Mitofsky, 1991). Les électeurs sont échantillonnés au moment où ils quittent les bureaux de scrutin. On demande aux personnes sélectionnées de remplir un questionnaire bref et simple et de le remettre dans une boîte de scrutin. Un questionnaire typique comprend environ 25 questions dans lesquelles on demande comment le répondant a voté, quelle est sa position sur des questions clés, quelle est l'opinion du répondant sur divers sujets et quelles sont ses caractéristiques démographiques. Les taux de refus, pour les sondages des votants réalisés par la société CBS se sont établis, en moyenne, à 25% lors des dernières élections (Mitofsky et Waksberg, 1989).

L'échantillonnage des électeurs pour les enquêtes électorales emploie habituellement un plan d'échantillonnage simple à deux degrés. Lors de la première étape, on tire un échantillon stratifié avec PPTE des circonscriptions électorales, où la mesure utilisée pour la taille est le nombre d'électeurs dans la circonscription. Lors de la deuxième étape, on choisit un échantillon systématique des électeurs quittant le bureau de scrutin, avec un intervalle d'échantillonnage choisi pour produire un échantillon approximativement "msep" des électeurs dans les États. Habituellement, un seul intervieweur est affecté à chaque circonscription sélectionnée. Le travail sur le terrain est simple quand le bureau de scrutin a une seule sortie et qu'on permet à l'intervieweur de s'en approcher. Quand il y a deux sorties (ou plus), les intervieweurs vont d'une sortie à l'autre, travaillant à chacune d'entre elles pour des périodes de temps déterminées. Quand cela se produit, l'intervalle d'échantillonnage doit être modifié en conséquence. Dans certains États, on ne permet pas aux intervieweurs de s'approcher à moins d'une certaine distance des bureaux de scrutin et cela peut créer des problèmes si les électeurs ont alors la possibilité de partir dans différentes directions avant que l'intervieweur ne réussisse à communiquer avec eux.

3.4 Enquête sur les soins médicaux ambulatoires

La National Ambulatory Medical Care Survey (NAMCS) des États-Unis emploie un plan d'enquête basé sur les flux de populations humaines pour recueillir des données sur les consultations médicales dans le cas des médecins en pratique privée qui dirigent les soins aux patients (Bryant et Shimizu, 1988). La NAMCS a été réalisée un certain nombre de fois depuis qu'elle a été mise en oeuvre en 1973. Pour chaque enquête, la collecte des données a été répartie pendant toute l'année civile de l'enquête afin de fournir des estimations annuelles des caractéristiques des visites. On a toutefois demandé à chacun des médecins échantillonnés de fournir des

renseignements pour un échantillon des visites qu'il reçoit pendant une seule semaine. On obtient une couverture annuelle en demandant à différents médecins échantillonnés de produire une déclaration pour différentes semaines de l'année.

L'échantillon de la NAMCS est basé sur un plan d'échantillonnage complexe à trois degrés qui a varié dans le temps. Un aperçu sommaire du plan d'échantillonnage suffira pour les besoins actuels; pour plus de détails, le lecteur est prié de se reporter à Bryant et Shimizu (1988). La première étape du plan d'échantillonnage de la NAMCS est le choix d'un échantillon stratifié avec PPTE d'UPÉ aréolaires, sélectionné avec une probabilité proportionnelle à la taille de la population. Dans la deuxième étape, les médecins sont échantillonnés à partir de listes dans les UPÉ sélectionnées avec des intervalles d'échantillonnage différents d'une UPÉ à l'autre afin de tenir compte des probabilités inégales de sélection des UPÉ (dans les enquêtes plus récentes, différentes classes de spécialités sont échantillonnées à des taux différents). Les médecins échantillonnés sont ensuite répartis au hasard, d'une façon équilibrée, dans une des 52 semaines de déclaration de l'année. On demande à chaque médecin de relever des renseignements pour un échantillon systématique des visites de ses patients qui se produisent pendant la semaine échantillonnée; l'intervalle d'échantillonnage est choisi afin de donner approximativement 30 visites échantillonnées au cours de la semaine. Un intervalle d'échantillonnage de 1, 2, 3 ou 5 est choisi pour un médecin donné en fonction du nombre de visites au cabinet que le médecin s'attend à recevoir pendant la semaine et du nombre de jours où il ou elle prévoit recevoir des patients. Les procédures de travail sur le terrain consistent à tenir un journal de l'arrivée des patients à des fins d'échantillonnage, puis à remplir un bref relevé de seize éléments pour chaque visite échantillonnée.

La NAMCS est une enquête portant sur les visites des patients et non sur les patients. À ce titre, elle fourni des renseignements utiles sur la nature du travail des médecins en fonction des visites - la fréquence d'utilisation des tests de diagnostic, les thérapies fournies et les caractéristiques démographiques des patients reçus. Toutefois, l'enquête ne fournit pas d'estimations en fonction d'un patient, comme les traitements et les résultats pour les périodes où les patients sont malades.

3.5 Enquêtes sur les passagers internationaux

Un certain nombre de pays réalisent des enquêtes sur leurs voyageurs internationaux, tant les personnes qui entrent dans leur pays que celles qui le quittent, par voie terrestre, maritime ou aérienne. Dans la présente sous-section nous décrirons brièvement les plans d'échantillonnage utilisés pour une enquête sur les passagers internationaux qui se déplacent par voie aérienne réalisée par les États-Unis, pour des enquêtes sur les passagers internationaux qui voyagent par voie aérienne ou terrestre réalisées par le Canada et pour une enquête sur les passagers internationaux qui se déplacent par voie aérienne ou maritime réalisée par le R.-U.

La United States Travel and Tourism Administration réalise une In-flight Survey of International Air Travelers pour enquêter tant sur les voyageurs étrangers aux États-Unis que sur les résidents des États-Unis qui voyagent à l'étranger (voir, par exemple, United States Travel and Tourism Administration, 1989). L'enquête est réalisée avec la collaboration volontaire d'environ trente sociétés aériennes. Un échantillon stratifié des vols réguliers est sélectionné pour la troisième semaine de chaque mois et tous les passagers de ces vols sont inclus dans l'échantillon. On fournit aux sociétés aériennes qui participent à l'enquête une trousse d'enquête qui renferme des instructions et des questionnaires dans les langues appropriées pour chaque vol échantillonné. Le personnel de cabine de la société aérienne distribue, dans les aires d'embarquement ou au cours des envolés, des questionnaires à remplir soi-même à tous les passagers adultes et il les recueille avant le débarquement. La non-réponse constitue un problème grave pour ces enquêtes. Dans le cas de l'enquête de 1988 sur les visiteurs aux États-Unis, aucun questionnaire n'a été retourné de la moitié des trousses remises au cours de ces envolés. Dans le cas des envolés pour lesquelles quelques questionnaires ont été retournés, le taux de réponse estimé pour les personnes qui ne résident pas aux États-Unis était de 44% et pour les résidents des É.-U., ce taux n'était que de 20%.

La Section des voyages internationaux de Statistique Canada réalise des enquêtes sur les voyages internationaux aux aéroports ainsi qu'aux ports d'entrée terrestres du Canada. Les enquêtes sont entreprises en collaboration avec Revenu Canada - Douanes et Accise, et ce sont les agents des douanes qui sont responsables de distribuer les questionnaires à remplir soi-même et à retourner par la poste. L'exposé présenté ici est basé sur le rapport produit par la Section des voyages internationaux, Statistique Canada (1979). Il reflète les plans

d'échantillonnage qui s'appliquaient avant certains changements qui ont été apportés récemment. Des plans d'échantillonnage semblables ont été utilisés pour les ports d'entrée terrestres et pour les aéroports et, par conséquent, nous ne décrirons ici que le plan d'échantillonnage utilisé pour les ports d'entrée terrestres.

À un certain moment, le plan d'échantillonnage employé pour les ports d'entrée terrestres, dans le cas des résidents du Canada revenant au pays qui avaient passé au moins une nuit à l'étranger, consistait à distribuer des questionnaires d'enquête à tout groupe de voyageurs à chaque quatrième jour pendant toute l'année, les jours étant choisis par échantillonnage systématique. Ce plan s'est révélé impraticable parce qu'il arrivait trop souvent que les agents des douanes ne l'appliquaient pas correctement. Il a donc été remplacé par un plan basé sur une période de sondage dans le cadre duquel on a attribué à un port d'entrée terrestre deux périodes de sondage pour chaque trimestre de l'année, pendant lesquelles les questionnaires devaient être distribués. On prévoyait que ces périodes de sondage devraient durer de six à dix jours, avec des périodes de sondage successives commençant à des intervalles d'environ 6 1/2 semaines (Gough and Ghangurde, 1977). Le nombre de questionnaires envoyés à un port d'entrée terrestre pour une période de sondage particulière était déterminé à partir de la circulation prévue à ce port d'entrée. On a ensuite demandé aux agents des douanes de commencer la distribution des questionnaires une journée donnée et de continuer à les distribuer jusqu'à ce qu'il ne leur en reste plus. Ce plan d'échantillonnage est adapté à des restrictions opérationnelles découlant de l'emploi des agents des douanes, pour lesquels l'enquête n'est qu'une préoccupation secondaire, comme travailleurs sur le terrain pour l'enquête. Le plan d'échantillonnage a certains désavantages importants, mais le fait que le taux de réponse soit de 20% ou moins constitue peut-être une préoccupation plus sérieuse.

Dans les enquêtes sur les voyageurs internationaux réalisées aux É.-U. et au Canada on s'en remet à la collaboration d'autres organismes pour effectuer le travail sur le terrain. Cette collaboration présente des avantages remarquables au niveau des coûts, mais il faut sacrifier la possibilité d'appliquer des contrôles rigoureux sur les procédures de travail sur le terrain. Les enquêtes sur les voyageurs se déplaçant par voie aérienne ou maritime réalisées au R.-U. emploient des procédures d'interviews par interrogation directe plus coûteuses.

La International Passenger Survey réalisée au R.-U. en 1984 incluait, comme strates, les trois aérogares de Heathrow ainsi que les aéroports de Gatwick et de Manchester (Griffiths et Elliot, 1987). Dans chaque aéroport, les jours ont été divisés en matinée et après-midi et ces périodes constituaient les UPÉ. Un échantillon stratifié d'UPÉ a été sélectionné et un échantillon systématique de passagers a été choisi dans les UPÉ sélectionnées. Un échantillon d'UPÉ pour d'autres aéroports a aussi été inclus. Deux procédures différentes de collecte des données étaient utilisées dans les ports de mer. Dans certains ports de mer, les intervieweurs échantillonnaient et interviewaient les passagers sur le quai. Dans d'autres ports, les intervieweurs voyageaient sur le bateau, interviewant les passagers pendant la traversée. Dans le premier cas, les intervieweurs travaillaient par postes qui comprenaient plusieurs départs et le poste devenait l'UPÉ. Dans le dernier cas, les traversées étaient les UPÉ.

3.6 Enquêtes réalisées dans les centres commerciaux

Deux types d'enquêtes sont réalisées dans les centres commerciaux. L'un de ces types d'enquêtes vise à décrire les caractéristiques socio-économiques des personnes qui font des courses, la région où elles habitent ainsi que leurs activités de magasinage au centre commercial. Dans l'autre type d'enquêtes on utilise le centre commercial comme endroit commode pour obtenir des échantillons de personnes provenant de l'ensemble de la population de la région.

Un exemple d'une enquête du premier type est une étude qui a été réalisée afin d'étudier l'incidence de l'ouverture d'un hypermarché dans les faubourgs de la ville de Southampton en Angleterre (Wood, 1978). Des enquêtes auprès des personnes qui font des courses ont été réalisées dans quatre centres commerciaux avoisinants avant et après l'ouverture de l'hypermarché (et aussi dans l'hypermarché lui-même). Dans chacun de ces centres, la première étape du processus d'enquête était le dénombrement de tous les points de vente ainsi que de leurs heures d'ouverture. La deuxième étape consistait à compter les départs des groupes de personnes qui font des courses des magasins échantillonnés pendant les heures sélectionnées, le comptage étant effectué pendant quinze minutes de chaque période d'une heure. L'opération de comptage a été effectuée pendant un mois. En fonction des totaux obtenus, les interviews ont été réparties entre les genres de magasin et les jours

de la semaine et à des magasins et des heures particulières. On a ensuite demandé aux intervieweurs d'interviewer le nombre précisé de personnes quittant le magasin, en interviewant la personne qui quittait le magasin après qu'ils avaient terminé une interview. L'échantillon porte sur les visites dans un magasin et les personnes qui font des courses pouvaient visiter plusieurs magasins lors d'un déplacement en particulier au centre commercial. On a demandé aux répondants s'ils avaient déjà visité des magasins du centre commercial lors de ce déplacement particulier et aussi combien d'autres magasins ils comptaient visiter. Ces données ont été utilisées pour élaborer des poids afin d'analyser les déplacements.

Le second type d'enquêtes réalisées dans les centres commerciaux utilise les personnes sélectionnées aux centres commerciaux comme un échantillon à l'aveuglette de la population dans son ensemble. Les interviews sur le vif de ce genre sont couramment utilisées dans les études de marché (Bush et Hair, 1985; Gates et Solomon, 1982). Les procédures sont souvent suivies au hasard et les échantillons pourraient bien être biaisés. Les questions qui sont en jeu sont traitées par Sudman (1980), quand il se penche sur les procédures à suivre pour échantillonner les centres commerciaux, les emplacements dans les centres sélectionnés et les périodes de temps afin d'améliorer les plans d'échantillonnage, et par Blair (1983), Dupont (1987) et Murry et coll. (1989).

3.7 Enquêtes sur la circulation routière

On réalise souvent des enquêtes sur les passagers des voitures de tourisme pour étudier l'utilisation des ceintures de sécurité et la concentration d'alcool dans le sang des conducteurs. Une discussion complète des questions complexes relatives au plan d'échantillonnage utilisé dans ces enquêtes dépasse la portée de la présente communication; on se limitera plutôt à quelques observations générales.

La méthode de collecte des données à utiliser exerce une forte influence sur les procédures d'échantillonnage employées pour les enquêtes sur la circulation routière. Le port de la ceinture de sécurité est généralement étudié à l'aide de méthodes d'observation, alors que la mesure de la concentration d'alcool dans le sang fait habituellement appel à l'éthyloscopie. L'utilisation de la ceinture diagonale par les personnes qui occupent les sièges avants peut être observée dans la circulation en mouvement, mais l'emploi des ceintures ventrales et le port des ceintures de sécurité par les autres passagers d'un véhicule ne peuvent être observés que lorsque ce dernier s'est arrêté brièvement, par exemple, à des feux de circulation. L'absence d'éclairage dans les rues peut empêcher l'observation de l'emploi des ceintures de sécurité la nuit à certains endroits. L'éthyloscopie exige que le véhicule soit arrêté et cela ne peut se faire de façon sécuritaire qu'en des emplacements où le véhicule à l'arrêt n'entrave pas la circulation. Contrairement aux enquêtes par observation, les enquêtes entrevues dans le cadre desquelles on arrête les véhicules font face à un problème de non-réponse considérable.

Une méthode ingénieuse pour étudier l'emploi des ceintures de sécurité sur les routes entre les États est décrite par Wells et coll. (1990). Dans le cadre de cette étude, un observateur était assis derrière le conducteur d'une fourgonnette de tourisme qui se déplaçait à une vitesse inférieure à la vitesse moyenne de la circulation dans la voie de droite de la route. Depuis cette position, l'observateur pouvait déterminer si les occupants des sièges avants des automobiles, des camions légers et des fourgonnettes qui dépassaient la fourgonnette dans laquelle il prenait place dans la voie adjacente portaient leur ceinture diagonale.

Une méthode plus courante employée pour étudier le port des ceintures de sécurité consiste à faire des observations aux intersections des rues et aux sorties des autoroutes où l'on trouve des feux de circulation et parfois dans des centres d'achat et des terrains de stationnement (Ziegler, 1983; Bowman et Rounds, 1989). O'Day et Wolfe (1984) décrivent une enquête par observation de l'emploi des ceintures de sécurité au Michigan réalisée à l'aide de cette méthode. Ils ont échantillonné un certain nombre d'unités spatiales, sélectionné un certain nombre d'intersections avec feux de circulation dans ces unités, choisi des jours où les observations devaient être réalisées à ces intersections et échantillonné cinq heures d'observation lors de chaque jour sélectionné (période basée sur un plan qui prévoit une alternance d'une heure de travail et d'une heure libre pendant toute la journée). Les observations de l'utilisation des ceintures de sécurité ont été prises aux intersections sélectionnées aux moments précisés pour les véhicules qui s'étaient arrêtés aux feux de circulation. Quand plus d'un véhicule était arrêté, l'observation commençait avec le deuxième véhicule, à cause du biais associé au premier véhicule à s'arrêter à un feu de circulation. Afin d'obtenir des renseignements plus détaillés sur l'utilisation des dispositifs de protection pour enfants, on a aussi observé les véhicules qui entraient dans des centres commerciaux et des haltes routières.

La méthode habituelle utilisée pour analyser des données d'observation sur l'emploi des ceintures de sécurité consiste à calculer la proportion des personnes observées qui portaient leur ceinture de sécurité. Brick et Lago (1988) proposent une autre mesure, la proportion du temps estimé pendant lequel les occupants des sièges avants portent leur ceinture de sécurité dans les véhicules admissibles par rapport au temps total qu'ils passent dans ces véhicules. Pour leur enquête on a choisi un échantillon probabiliste de toutes les intersections des chaussées, qu'on y trouve ou non des feux de circulation. Afin d'éviter un biais dû à la sélection, on a dit aux observateurs l'endroit qu'ils devaient utiliser pour faire leurs observations et dans quelle direction ils devaient observer la circulation pendant la période précisée de 40 minutes au cours de laquelle ils devaient réaliser leurs observations. On a estimé le temps pendant lequel les occupants des véhicules étaient sur la route en divisant la longueur du segment de route menant à l'intersection par la vitesse moyenne estimée de la circulation sur ce segment de route. Ce temps estimé a été utilisé comme facteur de pondération dans l'analyse.

Les considérations en matière d'échantillonnage pour les enquêtes au bord de la route au cours desquelles on a recours à l'éthyloscopie sont généralement semblables à celles qui s'appliquent aux enquêtes sur l'utilisation de la ceinture de sécurité, sauf que les endroits où les données peuvent être recueillies doivent être des endroits où il est possible d'arrêter les véhicules sans danger. Au cours de la National Roadside Breathtesting Survey de 1986 aux É.-U., les agents de police locaux ont collaboré à l'enquête en demandant aux conducteurs sélectionnés de s'arrêter et en les orientant vers les intervieweurs chargés de réaliser l'enquête (Wolfe, 1986). Les interviews duraient environ de 5 à 6 minutes. Quand un intervieweur avait terminé une interview et que le répondant avait subi le test de l'éthyloscopie, l'intervieweur faisait signe au policier d'arrêter le prochain véhicule qui passait.

4. CONCLUSIONS

Comme les exemples des sections précédentes le montrent, des considérations relatives au travail sur le terrain ainsi que le côté économique de la collecte des données jouent des rôles importants dans le choix du plan d'échantillonnage employé pour les enquêtes portant sur les personnes de passage. La longueur de l'intervalle de temps utilisé pour définir les UPÉ peut, par exemple, être dictée par la longueur d'un poste de travail approprié pour les travailleurs sur le terrain et cela peut entraîner l'utilisation d'UPÉ comportant des variations internes importantes dans le taux du flux. Par exemple, dans une enquête sur les passagers qui arrivent à une gare, le poste de travail d'un intervieweur du matin peut comprendre un flux en période de pointe quand les banlieusards arrivent tôt le matin et un flux moins important par la suite. Si l'on n'avait pas à faire correspondre les intervalles de temps des UPÉ au poste de travail des travailleurs sur le terrain, il serait préférable d'éviter une telle variation de flux à l'intérieur des UPÉ puisque cela entraîne des problèmes pour ce qui est de la façon de tirer un sous-échantillon dans les UPÉ sélectionnées.

Quand le flux des personnes dans une UPÉ est inégal, l'emploi de l'échantillonnage systématique, ou de tout plan d'échantillonnage "msep", pour sélectionner des personnes crée une charge de travail qui varie dans le temps. Si cette variabilité dans la charge de travail est importante, on a de la difficulté à décider comment affecter le personnel à l'UPÉ pour effectuer le travail sur le terrain, particulièrement dans le cas d'une enquête comportant des interviews par interrogation directe. L'affectation d'un nombre suffisant de personnes pour traiter les flux correspondant à des périodes de pointe n'est pas économique puisque les intervieweurs seront souvent inoccupés en dehors des périodes de pointe. Il arrive parfois qu'il soit préférable d'affecter un personnel suffisant pour traiter un flux un peu inférieur à celui que l'on rencontre pendant les périodes de pointe. Cela introduira une certaine non-réponse au moment où le flux correspond à une période de pointe parce qu'aucun intervieweur ne sera disponible pour réaliser une interview avec certaines personnes échantillonnées, mais en procédant de la sorte, on fait une meilleure utilisation du temps des intervieweurs.

L'utilisation la plus efficace du temps des intervieweurs consiste à leur faire interviewer la première personne qui arrive (ou qui part) après qu'ils ont terminé l'interview qu'ils sont en train de réaliser. Les plans de ce genre ont cependant le désavantage de ne pas produire d'échantillons probabilistes et, par conséquent, il y a un risque de biais dans les estimations de l'enquête. Quand on peut concevoir des plans d'échantillonnage probabilistes rentables, ces derniers doivent être préférés. Toutefois, le choix d'un plan d'échantillonnage dans lequel on choisit la première (ou la deuxième, ou la troisième) personne qui passe après qu'un intervieweur a terminé une

interview est naturellement intéressant pour les enquêtes où les interviews sont réalisées par interrogation directe quand le flux de populations humaines est très variable et imprévisible.

REMERCIEMENTS

Je désirerais exprimer ma reconnaissance aux nombreux chercheurs qui m'ont généreusement fourni des renseignements relatifs aux enquêtes sur les flux de populations humaines auxquelles ils ont participé.

BIBLIOGRAPHIE

Blair, E. (1983). Sampling issues in trade area maps drawn from shopper surveys, *Journal of Marketing*, **47**, 98-106.

Bowman, B.L., et Rounds, D.A. (1989). *Restraint System Usage in the Traffic Population, 1988 Annual Report*, Washington, D.C. U.S. Department of Transportation.

Brick, M., et Lago, J. (1988). The design and implementation of an observational safety belt use survey, *Journal of Safety Research*, 19, 87-98.

Bryant, E., et Shimizu, I. (1988). Sample design, sampling variance, and estimation procedures for the National Ambulatory Medical Care Survey, *Vital and Health Statistics*, Series 2, 108, Washington D.C.: U.S. Government Printing Office.

Bush, A.J., et Hair, J.F. (1985). An assessment of the mall intercept as a data collection method, *Journal of Marketing Research*, 22, 158-67.

Cochran, W.G. (1977). *Sampling Techniques* (3rd ed.), New York: John Wiley & Sons.

Doering, Z.D., et Black, K.J. (1989). Visits to the National Air and Space Museum (NASM): Demographic Characteristics, Working Paper 89-1, Institutional Studies, Smithsonian Institution.

Dupont, T.D. (1987). Do frequent mall shoppers distort mall-intercept survey results?, *Journal of Advertising Research*, Août-Sept., 45-51.

Gates, R., et Solomon, P.J. (1982). Research using the mall intercept: state of the art, *Journal of Advertising Research*, 22, 4, 43-49.

Goodman, R., et Kish, L. (1950). Controlled selection - a technique in probability sampling, *Journal of the American Statistical Association*, 45, 350-372.

Gough, J.H., et Ghangurde, P.D. (1977). Survey of Canadian residents returning by land, *Survey Methodology*, 3, 215-231.

Griffiths, D., et Elliot, D. (1987). Sampling errors on the International Passenger Survey, unpublished paper, London: Social Survey Division, U.K. Office of Population Censuses and Surveys.

Heady, P. (1985). Note on some sampling methods for visitor surveys, *Survey Methodology Bulletin* (U.K. Office of Population Censuses and Surveys), 17, 10-17.

Heeringa, S.G. (1985). *The University of Michigan 1984 Library Cost Study: Final Report*, Ann Arbor, Michigan: Institute for Social Research, University of Michigan.

Hess, I., Riedel, D.C., et Fitzpatrick, T.B. (1975). *Probability Sampling of Hospitals and Patients*, Ann Arbor, Michigan: Health Administration Press.

International Travel Section, Statistics Canada (1979). *Data Collection and Dissemination Methods for International Travel Statistics in Canada*, Ottawa: Statistique Canada.

Jessen, R.J. (1978). *Statistical Survey Techniques*, New York: John Wiley & Sons.

Levy, M.R. (1983). The methodology and performance of election day polls, *Public Opinion Quarterly*, 47, 54-67.

Mitofsky, W.J. (1991). A short history of exit polls. In *Polling in Presidential Election Coverage*, P. Lavrakas and J. Holley (eds.), Newbury Park, California: Sage.

Mitofsky, W.J., et Waksberg, J. (1989). CBS models for election night estimation. Paper presented at American Statistical Association San Diego Winter Conference.

Murry, J.P., Lastovicka, J.L., et Bhalla, G. (1989). Demographic and lifestyle selection error in mall-intercept data, *Journal of Advertising Research*, Février-Mars, 46-52.

O'Day, J., et Wolfe, A.C. (1984). *Seat Belt Observations in Michigan - August/September 1983*, Ann Arbor, Michigan: University of Michigan Transportation Research Institute.

Sudman, S. (1980). Improving the quality of shopping center sampling, *Journal of Marketing Research*, 17, 423-31.

United States Travel and Tourism Administration (1989). *In-flight Survey: Overseas and Mexican Visitors to the United States. Survey Period: January-December 1988*, Washington D.C.: United States Travel and Tourism Administration.

Wells, J.K., Williams, A.F., et Lund, A.K. (1990). Seat belt use on interstate highways, *American Journal of Public Health*, 80, 741-742.

Wolfe, A.C. (1986). *1986 U.S. National Roadside Breathtesting Survey: Procedures and Results*, Ann Arbor, Michigan: Mid-America Research Institute.

Wood, D. (1978). *The Eastleigh Carrefour: A hypermarket and its effects*, London: U.K. Department of the Environment.

Yates, F. (1981). *Sampling Methods for Censuses and Surveys* (4th ed.), London: Charles Griffin.

Ziegler, P.N. (1983). *Guidelines for Conducting a Survey of the Use of Safety Belts and Child Safety Seats*, Washington, D.C. U.S. Department of Transportation.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

AMÉLIORATION DE LA QUALITÉ DE LA BASE-LISTE DU RECENSEMENT DE L'AGRICULTURE DES ÉTATS-UNIS

C.Z.F. Clark[1]

RÉSUMÉ

La liste d'envoi postal du recensement de l'agriculture des États-Unis est compilée par recoupement de listes d'établissements établies à partir de données agricoles, statistiques, administratives et de données sur les produits. Des renseignements de base sur le processus d'établissement de cette liste sont présentés et une attention particulière est portée à cinq aspects de la qualité: contenu de la liste, précision des renseignements liés aux adresses, caractère unique des exploitations de la liste, couverture de l'univers et rapport coût-efficacité des méthodes d'établissement. Les procédures de listage qui se rapportent à ces critères de qualité sont analysées. Les résultats d'évaluations officielles des sources des enregistrements de la liste, de la couverture du recensement, de l'effet du modèle d'analyse discriminante sur le contenu et la couverture, des questionnaires non livrables et des enregistrements en double sont présentés. Les plans et projets futurs d'amélioration de la qualité de la liste sont examinés en fonction des résultats prévus.

MOTS CLÉS: Base-liste; collecte postale des données; agriculture.

1. ÉLABORATION DE LA LISTE D'ENVOI POSTAL DU RECENSEMENT DE L'AGRICULTURE

Le recensement de l'agriculture a pour objet de dénombrer tous les établissements dont les ventes de produits agricoles se sont élevées ou auraient dû s'élever à $1 000 ou plus l'année du recensement. Depuis 1969, le recensement de l'agriculture aux États-Unis, a recours à la méthode de l'envoi et retour par la poste de questionnaires plutôt qu'à un dénombrement sur place des exploitations agricoles. La liste d'envoi postal du recensement constitue la base de sondage des exploitations agricoles potentielles. Tous les efforts sont faits pour obtenir une réponse par la poste des exploitations dont l'adresse figure sur la liste. Ce n'est qu'en cas d'échec et donc en dernier recours qu'une collecte par téléphone des données est entreprise, et ce uniquement dans le cas des exploitations qui semblent importantes ou à caractère unique, ou encore des exploitations situées dans des comtés où le taux de réponse est inférieur à 75%.

La base-liste du recensement comprend des particuliers, des entreprises et des organismes reconnus comme ayant un certain lien avec l'agriculture. Une nouvelle liste est établie avant chaque recensement. La liste de 1987 a été compilée à partir des enregistrements du recensement de 1982, des dossiers administratifs du Internal Revenue Service (IRS) [service national du revenu] et de la Social Security Administration [administration de la sécurité sociale], et des dossiers statistiques du National Agricultural Statistical Service [service national de la statistique agricole] qui relève du ministère de l'Agriculture des É.-U. De plus, des listes ont été obtenues d'organismes de l'État ou du gouvernement fédéral, d'associations corporatives et autres organismes du genre, concernant des exploitations importantes ou spécialisées (par ex. les pépinières et les serres, les fermes de cultures spécialisées, les fermes avicoles, les fermes piscicoles, les fermes à bétail, les exploitations de culture d'aliments pour le bétail, les exploitations détentrices d'un permis de pacage). Le recensement de 1982 a fourni une liste d'entreprises comptant un ou plusieurs établissements de production agricole, laquelle a été mise à jour

[1] C.Z.F. Clark, Agriculture Division, U.S. Bureau of the Census, Washington, D.C. 20233 U.S.A.

en fonction des renseignements contenus dans la liste que tient le Census Bureau, appelée Standard Statistical Establisment List [liste des établissements statistiques types].

Le processus périodique d'élaboration de la liste d'envoi postal comprend les tâches suivantes: acquisition des diverses listes servant de sources d'information, adoption d'un mode de présentation unique pour toutes les adresses de la liste, couplage des enregistrements, élimination des enregistrements en double, tri des enregistrements, détermination du genre de questionnaire de recensement devant être expédié à chacun et préparation d'une étiquette d'adresse pour chaque enregistrement. Pour les trois derniers recensements, un couplage en deux étapes des enregistrements a été effectué en plus de les soumettre à un tri. Dans le cadre des recensements de 1978 et de 1982, entre les deux étapes de couplage, les données de la Farm and Ranch Identification Survey (FRIS) [enquête visant à repérer les fermes et les ranchs] furent utilisées pour découvrir les exploitations possiblement non agricoles et pour éliminer de la liste préliminaire les adresses en double. Au recensement de 1987, un modèle statistique d'analyse discriminante a été utilisé pour éliminer de la liste les enregistrements les moins susceptibles de correspondre à des fermes et identifier les enregistrements qui ne semblent pas correspondre à des exploitations agricoles afin qu'on leur envoie un questionnaire de recensement beaucoup moins détaillé. Le modèle a été testé en fonction de sa capacité de servir à l'élaboration de la liste entre les deux étapes de couplage. Utilisant l'une ou l'autre des méthodes de tri, le processus de couplage en deux étapes a permis l'utilisation d'enregistrements de référence plus à jour, comparativement à un couplage unique.

En 1987, des paramètres généraux pour la taille et la composition de la liste d'envoi postal du recensement ont été établis en vue de réduire le fardeau des répondants et de mieux contrôler les coûts. La liste d'envoi postal du recensement a été limitée à un maximum de 4.2 millions d'adresses, dont pas plus de 3.2 millions devaient recevoir soit le questionnaire courant de quatre pages, soit le questionnaire-échantillon détaillé de six pages. On autorisait l'envoi à jusqu'à 1.2 million de répondants du questionnaire abrégé de deux pages. Ces limites sont nettement inférieures à celles du recensement de 1982 dans le cadre duquel le questionnaire abrégé de l'enquête FRIS a été envoyé à 3.0 millions d'adresses et le questionnaire du recensement à 3.65 millions (y compris 1.7 million de répondants à l'enquête), soit un total de 5.35 millions de contacts pour le recensement. La liste définitive de 1987, de 4.1 millions d'adresses, respectait les exigences susmentionnées concernant la collecte des données. Sur un total de 4 098 693 adresses - 3 192 287 adresses ont reçu le questionnaire courant de quatre pages ou le questionnaire-échantillon de six pages tandis que 906 406 adresses recevaient le questionnaire abrégé de deux pages.

2. CONTENU DE LA LISTE

2.1 Sources des enregistrements de la liste d'envoi postal

La liste d'envoi postal du recensement de l'agriculture constitue la base du dénombrement. Le contenu et le degré de couverture de la liste sont donc des facteurs déterminants de l'intégralité de ce dénombrement. Des efforts concertés sont déployés pour que la liste puisse être fondée sur l'ensemble des principales sources de renseignements sur les exploitations agricoles. Aux États-Unis, il s'agit des dossiers fiscaux des contribuables ayant tiré un revenu de l'agriculture, des dossiers d'emploi des employeurs du secteur agricole, des dossiers statistiques des organismes fédéraux qui recueillent des données agricoles, des dossiers administratifs des organismes fédéraux ou d'État administrant des programmes agricoles et autres programmes connexes et des dossiers privés des associations corporatives agricoles et d'autres organismes du genre. Les principales sources des enregistrements de la liste sont restées fondamentalement les mêmes lors des trois derniers recensements. Cependant, certains changements, assez importants pour être notés, ont été apportés au contenu de la liste de 1987.

En effet, en 1987, le Census Bureau a été en mesure d'utiliser la liste des exploitations agricoles du National Agricultural Statistical Service (NASS) [service national de la statistique agricole] couvrant l'ensemble des 50 États, alors que des données sur seulement 31 États étaient disponibles au moment de l'établissement de la liste du recensement de 1982. La liste du NASS est constamment tenue à jour. Toute nouvelle donnée concernant le nom ou l'adresse des exploitations est ajoutée à la liste dès réception et les renseignements sur les exploitations agricoles et les produits sont mis à jour en fonction de données d'enquête récentes et de données

administratives. Toutefois, le NASS n'a pas accès aux dossiers fiscaux des exploitants agricoles. Les données de la liste sont classées par catégorie une fois par année. La liste du NASS a été élaborée de façon à offrir une bonne couverture des grandes exploitations et d'exploitations spécialisées en vue de produire des estimations à base duale au niveau de l'État. Selon des études du NASS fondées sur un échantillon tiré de sa base de sondage aréolaire, la liste comprendrait environ 55% des exploitations agricoles des États-Unis qui correspondent à la définition du recensement.

La liste d'envoi postal du recensement de 1987 n'a pas utilisé, comme aux recensements de 1978 et 1982, la liste du Agricultural Stabilization and Conservation Service (ASCS) [service de stabilisation et de conservation des terres agricoles]. Cette décision a été prise pour plusieurs raisons, notamment parce que les renseignements importants contenus dans les enregistrements de l'ASCS étaient fort probablement déjà inclus dans la liste courante du NASS, compte tenu du fait que ces renseignements sont pris en compte automatiquement dans le processus même d'établissement de la liste du NASS. En outre, la liste de l'ASCS était très importante -- 5 millions d'enregistrements. Enfin, les données relatives aux noms et adresses s'étaient révélées moins fiables que celles d'enregistrements d'autres sources utilisées lors de recensements précédents. En effet, dans le cadre du processus de couplage, un grand nombre des enregistrements de l'ASCS n'ont pu être appariés avec des enregistrements d'autres sources et ont dû faire l'objet d'un tri (706 000 - 25% - des 3 millions d'enregistrements de l'enquête FRIS de 1982 provenaient uniquement de la liste de l'ASCS). Par la suite, nous avons constaté qu'un grand nombre des adresses concernées ne correspondaient pas à une exploitation agricole. Parmi les adresses d'exploitations agricoles du recensement de 1982, 1.4% seulement avaient comme seule source la liste de l'ASCS.

La liste d'envoi postal de 1987 a tenu compte des enregistrements du ministère de l'Agriculture de l'État du Massachusetts, à la demande du gouvernement de cet État. Le ministère était d'avis que les données du recensement de l'agriculture de 1982 ne fournissaient pas une couverture très complète des exploitations de cet État et estimait que l'utilisation des 10 600 enregistrements de sa liste améliorerait la couverture des données du recensement. Des codes spéciaux ont été introduits lors de l'établissement de la liste d'envoi postal et du traitement des données afin de pouvoir déterminer toute amélioration de couverture attribuable à l'utilisation de cette liste. Si l'évaluation est favorable, nous analyserons la possibilité d'utiliser d'autres listes du genre dans les États pour lesquels le taux de sous-dénombrement du recensement agricole est relativement important.

2.2 Évaluation des sources des enregistrements de la liste d'envoi postal

Après le recensement de l'agriculture de 1982, une évaluation poussée des sources des enregistrements utilisées pour l'établissement de la liste finale d'envoi postal a été entreprise en se servant pour cela d'un échantillon pondéré d'enregistrements du recensement. Les sources des enregistrements ont été évaluées par rapport aux facteurs suivants: taux de réponse des personnes dont le nom figure sur la liste; proportion des noms de la liste correspondant à des exploitants agricoles; proportion de l'univers des fermes de recensement couverte par la liste; nombre et proportion des fermes dénombrées uniquement grâce à une liste donnée; et caractéristiques des fermes dénombrées uniquement grâce à la liste en question. Les résultats de l'évaluation montrent que 11.5% de toutes les fermes dénombrées au recensement provenaient de la seule liste du Internal Revenue Service (IRS), comparativement à 2.9% pour ce qui est de la même liste au recensement précédent et à 2.0% dans le cas de la liste du NASS. L'évaluation révèle également que les proportions de fermes dénombrées uniquement grâce à la liste du IRS ou aux listes spéciales de production agricole sont beaucoup plus élevées que pour les autres listes de référence. Les résultats de l'évaluation des sources de la liste d'envoi postal de 1982 sont résumés dans le tableau 1 (Gaulden, 1985). Une évaluation similaire de la liste du recensement de 1987 est en cours et les résultats ne sont pas encore connus.

Tableau 1: Evaluation des sources des enregistrements de 1982

Liste de référence	Nombre d'enr. dans la liste finale	Taux de réponse	Taux de fermes	Taux de couverture	Taux de fermes uniques	Taux de couverture unique
Toute source	3 629 425	87.4	61.9	Sans objet	Sans objet	Sans objet
IRS	2 688 850	90.5	69.1	85.6	9.3	11.5
Fermes du rec. de 1978	1 913 800	95.2	75.9	70.4	3.2	2.9
Fermes NASS	1 638 219	88.8	69.7	51.6	2.6	2.0
ASCS	2 204 275	87.6	71.1	69.9	1.5	1.4
Listes spéciales prod. agric.	139 075	91.5	59.4	3.8	9.1	0.6

Taux de réponse - Proportion de l'ensemble des enregistrements de la liste finale pour lesquels on a obtenu une réponse dans le cadre du recensement.
Taux de fermes - Proportion de répondants représentant des exploitations agricoles.
Taux de couverture - Proportion de l'ensemble des fermes tirées de cette liste.
Taux de fermes uniques - Proportion de répondants représentant des fermes tirées uniquement de cette liste.
Taux de couverture unique - Proportion de l'ensemble des fermes provenant uniquement de cette liste.

De plus, en 1985, une étude de plus petite envergure a été entreprise conjointement par le Census Bureau et le NASS pour mesurer et comparer le degré de couverture de l'univers des fermes obtenu à partir des listes du recensement et du NASS, ensemble et séparément. Le dénombrement sur place effectué dans quatre États (Washington, New York, Iowa et la Georgie) dans le cadre de la June Enumerative Survey (JES) de 1984, à partir d'un échantillon tiré d'une base aréolaire, devait fournir une couverture complète des segments échantillonnés dans chaque État. Pour chaque unité de l'échantillon de la base aréolaire, le NASS a déterminé s'il y avait appariement, appariement probable ou non appariement avec les unités de la liste du NASS. Le Census Bureau a pour sa part précisé si chaque unité de l'échantillon de la base aréolaire correspondait à une ferme du recensement de 1982, à un répondant n'exploitant pas une ferme ou encore à un non-répondant.

Le Census Bureau a estimé que la couverture de l'univers des fermes dans les quatre États était de 75% à la suite de la seule utilisation de la liste des fermes du recensement, de 53% à la suite de la seule utilisation de la liste du NASS et de 80% à la suite de l'utilisation combinée de ces deux listes. La couverture s'améliorait lorsqu'on tenait compte uniquement des fermes dont les ventes ont été supérieures à $2 500, soit 80% avec la seule liste du recensement, 61% avec la liste du NASS et 86% avec les deux listes. Les estimations variaient considérablement d'un État à l'autre, la couverture attribuable à l'utilisation des deux listes allant de 71% pour l'État de Washington à 88% pour l'Iowa. Dans le cas de la liste du recensement, les estimations sont un peu moins élevées que ce à quoi on aurait pu s'attendre, en raison du fait que la liste datait déjà de deux ans. Les estimations du NASS différaient quelque peu de celles du Census Bureau à cause de l'utilisation de règles d'appariement différentes. Compte tenu des évaluations de la liste d'envoi postal de 1982 et des résultats de l'étude d'appariement avec les unités des listes du recensement et du NASS, on a jugé que les listes du IRS constituaient des sources de référence qui devaient absolument être utilisées en vue d'améliorer la couverture des fermes dans le cadre du recensement de 1987 (Gaulden, 1986 et Davie, 1986).

2.3 Tri des enregistrements de la liste d'envoi postal

Bien que tous les enregistrements utilisés pour l'établissement de la liste d'envoi postal du recensement de l'agriculture ont un lien quelconque avec l'agriculture, cela ne garantit pas que chacun d'eux représente une exploitation active unique, correspondant à la définition d'une ferme pour les besoins du recensement. La complexité des types d'exploitation agricole rend difficile la tâche de déterminer le caractère unique d'une exploitation donnée. Les données d'identification et les données agricoles contenues dans chaque enregistrement

de la liste ne suffisent pas généralement à assurer qu'il s'agit bien d'une ferme de recensement. En 1978 et 1982, un questionnaire de la FRIS a été envoyé par la poste aux adresses dont le statut d'exploitation agricole était incertain ou qui pouvaient correspondre à des exploitations déjà pris en compte, ce qui a permis de réduire le nombre des enregistrements de la liste d'envoi postal finale du recensement. Le questionnaire avait été conçu de façon à obtenir des précisions sur la production agricole des répondants et pour découvrir lesquels exerçaient des activités multiples. Même après utilisation de ce questionnaire d'enquête, le nombre des enregistrements de la liste finale représentait près du double du nombre des exploitations agricoles américaines dénombrées lors de quelques uns des derniers recensements.

Le modèle d'analyse discriminante appliqué à la liste d'envoi postal de 1987 a aidé à atteindre le premier objectif, soit l'élimination des enregistrements ne correspondant pas à des fermes, mais pas le deuxième, soit le repérage des enregistrements en double. Une évaluation du modèle d'arbre de classification (Schmehl, 1990) a démontré que ce dernier était efficace pour séparer les enregistrements en groupes selon la proportion de fermes de recensement prévues dans chaque groupe. Les caractéristiques des enregistrements comme la source de l'enregistrement de la liste d'envoi postal, le nombre de listes de référence sur lesquelles figure l'enregistrement, la valeur prévue des ventes de produits agricoles, l'emplacement géographique et la situation au recensement agricole de 1982 ont été utilisées pour séparer les enregistrements de la liste d'envoi postal de 1982 en fonction des groupes du modèle. La proportion des enregistrements correspondant à des fermes du recensement de 1982 dans chaque groupe du modèle a été calculée de façon à fournir une estimation de la probabilité qu'un répondant situé à une adresse de la liste d'envoi postal de 1987 ayant les caractéristiques du groupe puisse être un exploitant agricole (Clark, 1989 et Owens, 1989). Des améliorations au modèle, prévues pour 1992, visent à accroître sa précision et sa capacité de prévoir en tenant compte de nouvelles variables des enregistrements et en n'étant plus limité par les contraintes actuelles concernant la taille des groupes du modèle. Ces améliorations pourraient par conséquent contribuer à éliminer un plus grand nombre d'enregistrements de la liste finale d'envoi postal.

2.4 Taille de la liste d'envoi postal

Les changements liés aux sources de référence utilisées pour le recensement de 1987 étaient censés entraîner une réduction du nombre des enregistrements distincts après couplage et appariement et nous permettre de pouvoir dresser une liste des exploitations agricoles potentielles dont la taille se rapprocherait de la limite de 4.2 millions d'enregistrements fixée. Avant la mise en oeuvre, il n'était pas certain dans quelle mesure le modèle d'analyse statistique discriminante réussirait à séparer en groupes les enregistrements de la liste en fonction de la probabilité qu'ils correspondent à une exploitation agricole. Comme aucune enquête de préclassement n'avait été autorisée, nous ne pouvions nous permettre d'établir une longue liste et de procéder ensuite au tri des enregistrements pouvant ne pas correspondre à une ferme. La meilleure stratégie à adopter compte tenu de ces contraintes semblait être de limiter le nombre des enregistrements de référence jugés les moins fiables.

Le nombre total d'enregistrements utilisés au départ en vue de l'établissement de la liste du recensement de 1987 était de 13.5 millions, comparativement à 18.9 millions en 1982. Après le couplage des enregistrements en double en fonction du nom, de l'adresse, du Social Security Number (SSN) [numéro d'assurance sociale], ou du Employer Identification Number (EIN) [numéro d'employeur], il restait encore 6.0 millions d'enregistrements à la fin de la seconde étape du processus de couplage. Parmi les enregistrements couplés, environ 1.6 million provenaient exclusivement de sources non agricoles et ont été éliminés de la liste d'envoi postal. (Des enregistrements correspondant à des exploitations non agricoles tirés à la fois de la liste du NASS et de la liste du recensement de 1982 ont été inclus dans le processus initial de couplage afin qu'on puisse différencier les cas douteux d'exploitations agricoles des exploitations des entreprises non agricoles actuelles en déterminant si les enregistrements classés antérieurement comme ne correspondant pas à des fermes pouvaient être couplés avec des enregistrements récents tirés de listes d'exploitations agricoles.) Un total de 180 000 enregistrements représentant des non-répondants au recensement précédent et dont les ventes prévues de produits agricoles devaient être inférieures à $2 500 ont aussi été éliminés, ce qui laissait 4 275 000 enregistrements. Un nombre additionnel de 175 000 enregistrements ont été enlevés sur la base des probabilités du groupe auquel ils appartenaient, telles que déterminées par le modèle d'analyse discriminante, pour un solde de 4.1 millions d'enregistrements dans la liste finale d'envoi postal. L'échantillon du recensement a été prélevé dans cette liste et un questionnaire abrégé a été envoyé aux adresses des 900 000 enregistrements restants, les moins susceptibles de représenter des fermes de recensement.

2.5 Résultats de la collecte des données du recensement

Des données de recensement ont été obtenues pour 3 404 036 (86.2%) des adresses de la liste d'envoi postal du recensement. Aux recensements de 1982 et de 1978, les taux de réponse étaient sensiblement les mêmes, soit 85.1% et 87.3%, respectivement. Parmi les questionnaires reçus, 1 826 702 correspondaient à des exploitations agricoles, 1 533 776 à des exploitations non agricoles et 43 558 ne pouvaient être classés. A la fin du processus, on comptait au total 542 248 non-répondants et 148 252 questionnaires retournés au maître de poste. Une méthode de pondération a été utilisée pour tenir compte des exploitations agricoles non répondantes (en fonction d'enquêtes menées par les divers États auprès de non-répondants) à la fin du recensement. Grâce à cette méthode, nous avons estimé que 263 057 des non-répondants avaient une exploitation agricole, ce qui porte à 2 087 759 le nombre des fermes de recensement sur un total de 4 095 061 questionnaires envoyés par la poste.

3. PRÉCISION DES RENSEIGNEMENTS CONTENUS DANS LES ENREGISTREMENTS

3.1 Méthodes utilisées pour accroître la précision des renseignements figurant dans les enregistrements

Dans le cadre de l'établissement de la liste d'envoi postal, certaines méthodes ont été employées pour améliorer la précision des renseignements contenus dans les enregistrements de la liste finale (Sandusky, 1984 et Gaulden, 1990). Une présentation uniforme a été adoptée pour chaque enregistrement de référence relativement au nom (contrôle du nom), à l'adresse de voirie ou de route postale rurale, au nom de la localité et aux zones de codes. Une telle normalisation était nécessaire aux fins d'application de tous les procédés automatisés de couplage et de repérage des enregistrements en double. Les enregistrements ont d'abord été couplés en fonction des numéros d'identification, soit le numéro d'assurance sociale et le numéro d'employeur. Tous les enregistrements non appariés après ce premier couplage et susceptibles de représenter la même exploitation (enregistrements en double) ont été codés par région géographique et ont fait l'objet d'une recherche de parties de nom et d'un recodage de nom avant d'être soumis à un appariement automatisé selon le nom et l'adresse par code ZIP [code postal] ou région. Un autre couplage a été effectué en fonction des enregistrements en double repérés dans des listes antérieures. Les enregistrements en double ont été supprimés, conservant l'enregistrement dont la source présentait la meilleure qualité du point de vue des adresses et des renseignements. Tous les enregistrements susceptibles d'être des doubles (identifiés à l'aide de nombres arrangés par paires) ont été vérifiés à la main. L'adresse et les éléments d'identification devant figurer dans l'enregistrement final étaient fondés sur les renseignements provenant de l'ensemble des enregistrements couplés.

Le contrôle du nom était essentiel aux fins de déterminer si un enregistrement était ou pouvait être un enregistrement en double. Cette vérification s'est faite au moment du couplage initial selon le numéro d'identification. Pour assurer un contrôle adéquat des noms dans les enregistrements agricoles, un dictionnaire de "listage successif" a été conçu. Ce dictionnaire contenait plus de 1 000 termes et abréviations (comme Farm, Dairy, Bros) qui pouvaient figurer dans la zone du nom sans nécessairement correspondre au nom en tant que tel. Le contrôle du nom s'est fait par lecture de droite à gauche de la zone principale du nom jusqu'à la rencontre d'un terme non numérique de trois caractères ou plus ne figurant pas déjà dans le dictionnaire de listage successif. Les quatre premiers caractères de ce terme devenaient le code de contrôle du nom. Un releveur de nom (code d'indicateur) a été introduit dans l'enregistrement, indiquant dans quelle zone se situait le terme duquel a été dérivé le code de contrôle du nom. Un petit nombre d'enregistrements (la demie d'un pourcent) n'avaient aucun code de contrôle du nom (par ex. A & B FARMS INC).

Des zones de codes aux fins de traitement ont été attribuées à l'enregistrement initial pour les besoins du couplage des enregistrements et faciliter l'utilisation de l'information la plus fiable possible dans l'enregistrement final. En particulier, chaque enregistrement a reçu un code de priorité de nom et d'adresse, lequel a servi lors du couplage à déterminer lesquels parmi les enregistrements de référence en double devaient être conservés. Le code de priorité était fixé en fonction de l'actualité probable des renseignements concernant l'adresse de l'enregistrement de référence.

Chaque enregistrement a aussi reçu une mesure de taille fondée sur des indicateurs pertinents de l'enregistrement de référence. Il s'agit de la TVP (total value of agricultural products) qui est une estimation de la valeur totale des produits agricoles qui pourraient être vendus pendant l'année de recensement. Chacune

des sources d'enregistrements utilisées avait une zone spéciale pour ce code de taille de sorte qu'au moment du couplage, le code de taille a pu être conservé pour tous les enregistrements sur lesquels figuraient un nom en transférant à l'enregistrement retenu les données provenant de l'enregistrement supprimé. Cela a permis de définir à la fois un "code de combinaison de sources" indiquant toutes les sources utilisées pour produire l'enregistrement final et un "code de taille final" fondé sur la fiabilité des renseignements de ce type de chaque source. Les codes de sources comme les codes de taille ont été des variables importantes du modèle d'analyse discriminante. Le code de taille final a été utilisé dans le cadre des opérations du recensement pour déterminer le genre de questionnaire à envoyer par la poste, la fraction de sondage et le genre de mesure de suivi à tenter auprès des non-répondants. Les deux codes ont servi à l'évaluation de la liste d'envoi postal du recensement.

Le système de codage géographique a été conçu de façon à assurer que tous les enregistrements utilisés dans le processus de couplage contenaient des codes géographiques normalisés et vérifiés, c'est-à-dire pour ce qui est des codes d'États et de comtés de recensement agricole, des codes alphabétiques des comtés et des codes ZIP. Ces données de codage étaient essentielles aux opérations d'envoi postal et de dépouillement des questionnaires ainsi que pour le couplage des enregistrements selon le nom et l'adresse. Les renseignements géographiques étaient tirés du fichier de référence de la Geography Division. Des contrôles de validité ont été effectués sur les codes ZIP en fonction des noms des bureaux de poste, utilisant le nom complet et les noms classés phonétiquement (code soundex). Au cours de l'étape subséquente de couplage selon le nom et l'adresse, le code ZIP a été utilisé comme facteur de blocage afin que seuls les enregistrements à l'intérieur d'un même code ZIP soient comparés aux fins d'appariement possible.

3.2 Évaluation de la précision des données des enregistrements

Plusieurs mesures liées à la collecte des données du recensement permettent une évaluation de la précision des données de nom et d'adresse contenues sur chaque enregistrement. Les questionnaires que le bureau de poste n'a pu livrer ont été renvoyés au bureau de dépouillement de Jeffersonville en Indiana, ce que nous appelons les PMR (post master returns) [retours au maître de poste]. Les PMR ont été triés par cas de décès, changement de nom ou d'adresse et autres raisons. Tous les PMR ont été postés à nouveau une fois les changements d'adresse faits, les questionnaires de personnes décédées étant adressés à "La succession de...". On a ensuite dressé la liste de tous les PMR retournés pour la deuxième fois dont on avait déterminé à l'avance qu'il pouvait s'agir d'exploitations importantes afin que ces cas puissent ultérieurement faire l'objet d'un suivi par téléphone. En 1987, il restait à la fin du recensement un total de 148 252 PMR (3.6% des adresses de la liste d'envoi finale), comparativement à 82 792 PMR en 1982 (2.3%). Cet accroissement du nombre des PMR est sans doute attribuable à l'absence de données d'une enquête comme la Farm and Ranch Identification Survey. Grâce à cette enquête, 446 000 PMR avaient pu être éliminés de la liste provisoire de 1982; 1 373 000 adresses dites vérifiées de la liste de la FRIS figuraient aussi dans la liste du recensement, bien que 522 000 d'entre elles correspondaient à des personnes n'ayant pas répondu à l'enquête.

Une indication de la précision des renseignements de chaque enregistrement peut être obtenue en examinant les résultats de l'enquête menée auprès des non-répondants. En effet, chaque État mène une telle enquête avant de mettre fin aux opérations de collecte des données du recensement. Cette enquête permet de produire une estimation du pourcentage de propriétaires d'exploitations agricoles parmi les non-répondants. Un questionnaire est envoyé par la poste à tous les non-répondants, lesquels sont par la suite rejoints par téléphone. En 1987, nous avons été en mesure de contacter et de classer 63.5% des 27 096 cas de l'échantillon; en 1982, nous avions contacté et classé 70.5% des 13 489 cas de l'échantillon. L'écart entre les deux recensements pour ce qui est de la réponse et de la classification est attribuable au changement de taille de l'échantillon, au changement de méthode de collecte des données (menée par des analystes en 1982 et par du personnel de soutien en 1987) et à la qualité des adresses de la liste.

D'autres renseignements sur la précision des adresses de l'échantillon de l'enquête auprès des non-répondants ont pu être obtenus grâce à une étude de 1 263 cas de non-répondants n'ayant pu être classés et qui faisaient partie de strates données dans six États. Au début de 1989, nous avons mené un suivi postal et téléphonique intensif pour tenter d'obtenir plus de renseignements au sujet de ces cas. Nous avons d'abord envoyé un questionnaire par courrier recommandé à chaque unité de l'échantillon. Nous avons obtenu une réponse de 488 (38.6% des cas). Les bureaux de poste ont retourné 382 formules (30.2%) pour décès, changement d'adresse ou autre raison (PMR). Après avoir passé tous les cas en revue, nous avons effectué un suivi téléphonique,

lorsque cela était possible, pour recueillir les données. Nous avons mené des entrevues sur place pour les autres cas et au total, nous avons recueilli des données au sujet de 880 des unités d'enquête (70%). Le suivi sur place nous a permis de constater qu'un grand nombre parmi ces unités n'avaient pas la bonne adresse. Le pourcentage élevé de PMR à la suite de l'envoi par courrier recommandé pouvait être attribué soit au fait que les bureaux de poste n'avaient pas communiqué de bonnes adresses au moment de l'envoi initial des questionnaires du recensement, soit au fait que les changements d'adresse n'avaient pas bien été pris en compte au moment des envois ultérieurs.

Chaque enregistrement de la liste contient d'autres renseignements utiles aux opérations de recensement, en plus de ceux se rapportant au nom et à l'adresse. La section 3.1 décrit l'utilisation qui a été faite des codes de sources de référence et des codes de taille dans les opérations du recensement. Après le recensement, le fichier des enregistrements de la liste d'envoi postal a été apparié au fichier de traitement du recensement et au fichier des fermes de recensement. Les totalisations établies à partir des résultats de l'appariement aident vraiment à mieux évaluer la qualité de la liste d'envoi postal du recensement. Certaines précisions au sujet de la qualité des sources des enregistrements ont été données à la section 2.2. Des totalisations faisant un rapprochement entre le code de taille de la liste d'envoi postal et la valeur réelle des produits agricoles vendus (TVP) l'année du recensement permettent d'évaluer la précision de ce code. Environ 35% de l'ensemble des enregistrements de la liste avaient un code de taille TVP identique à leur code de taille de liste d'envoi postal; 73% se situaient dans la tranche de code de taille immédiatement inférieure ou supérieure à celle de la liste d'envoi postal.

Tableau 2: Lien observé en 1982 entre la TVP réelle et la TVP prévue

Limite inférieure du code de taille de la liste d'envoi postal	Pourcentage d'enregistrements ayant le même code	Pourcentage d'enregistrements ayant un code d'une tranche inférieure ou supérieure
Ensemble des enregistrements	34.9	72.7
$1 000 000	3.1	80.0
500 000	37.7	82.2
200 000	51.9	84.3
100 000	49.1	74.0
80 000	22.6	68.0
60 000	24.9	60.5
40 000	31.9	73.0
20 000	39.8	74.6
10 000	34.2	73.0
5 000	27.9	64.0
2 000	32.8	77.4
1 000	34.5	87.5
Moins de $1 000	42.4	70.0

3.3 Améliorations possibles de la précision des renseignements liés aux adresses

L'augmentation du nombre des PMR observée au recensement de 1987, l'accroissement du nombre des unités non classées dans l'enquête auprès des non-répondants ainsi que le nombre élevé de mauvaises adresses dans l'étude restreinte menée auprès des non-répondants du recensement donnent à penser qu'il y a place pour des améliorations en ce qui concerne la précision des données relatives aux adresses. Nous étudions présentement la possibilité d'utiliser les services de vérification d'adresses que le United States Post Office [société des postes des États-Unis] pourrait offrir avant le recensement en se chargeant de mettre à jour les listes d'adresses fournies par le Census Bureau. Nous aurions notamment recours à ce service pour le repérage des adresses en double potentielles. Il semble que le nombre de ces dernières pourrait augmenter avec le changement des adresses de routes rurales en adresses de voirie. Notre système permet de repérer les enregistrements en double potentiels ayant le même nom et des adresses différentes.

4. CARACTÈRE UNIQUE DES EXPLOITATIONS DE LA LISTE

4.1 Procédures utilisées pour assurer le caractère unique des adresses et des exploitations de la liste

Les procédures d'appariement utilisées dans le cadre des opérations du recensement de l'agriculture ont pour objet de coupler les enregistrements hautement susceptibles de représenter le même nom et la même adresse. Les méthodes automatisées et manuelles sont conçues de façon à repérer les concordances presque parfaites. Les règles visent à assurer un degré élevé de couverture en conservant dans la liste les enregistrements pouvant représenter une exploitation distincte dont le nom et l'adresse ne varie que très peu par rapport à un autre enregistrement. Dans l'univers agricole, les exploitants de fermes ou de ranchs possèdent souvent des exploitations multiples de sorte qu'un exploitant donné peut être propriétaire d'une entreprise individuelle en plus de faire partie d'une ou plusieurs sociétés de personnes. Pour tenir compte de ces diverses formes d'exploitation, le processus d'établissement de la liste d'envoi postal du recensement comprend le couplage automatique d'enregistrements représentant des exploitations figurant sous des noms différents dont les liens d'affaires ont été établis lors de recensements antérieurs. Le programme d'édition des enregistrements de référence comprend aussi un signal d'avertissement PPC (possible partnership or corporation) [possibilité de société de personnes ou de société de capitaux] pour faciliter le couplage. Le signal PPC sert à éviter la suppression automatique d'enregistrements que l'ordinateur pourrait considérer comme figurant en double, tous les enregistrements reliés à un enregistrement PPC devant être vérifiés manuellement.

4.2 Évaluation du caractère unique des adresses et des exploitations de la liste

Le programme d'évaluation de la couverture mesure les erreurs pour ce qui est de la classification des fermes selon le type de questionnaire, des exploitations en double et des fermes non comprises dans la liste d'envoi postal. Les non-fermes classées à tort comme des fermes et les exploitations comprises en double sont à l'origine du surdénombrement des fermes au recensement. Le nombre total estimé de fermes comptées deux fois en 1987 (135 600) n'est guère différent de celui de 1982 (113 623). Toutefois, la proportion de fermes surdénombrées représentant des exploitations comptées en double est passée de 17% en 1982 (19 062 fermes) à 47% en 1987 (63 290). La hausse pour 1987 est principalement attribuable à l'absence d'une enquête de sélection avant le recensement.

Le programme d'édition initial a permis d'uniformiser les données concernant les noms et les adresses pour tous les enregistrements. Le programme d'uniformisation des adresses utilisé pour la liste d'envoi postal du recensement de l'agriculture a été spécialement conçu pour les adresses rurales. Nous sommes en train d'évaluer ce programme en le comparant à celui utilisé au recensement décennal et à celui élaboré pour la Post Enumerative Survey de 1990 [enquête postcensitaire]. Les résultats indiqueront s'il est nécessaire d'utiliser un autre programme d'uniformisation des adresses en vue d'accroître la précision de notre processus de couplage.

Nous évaluons également la possibilité d'utiliser un système de couplage probabiliste pour l'appariement selon le nom et l'adresse des enregistrements. Ce genre de méthode permettrait de préciser le degré de certitude requis pour les enregistrements appariés dans le système. Nous comparerons le programme probabiliste d'appariement utilisé pour apparier les enregistrements de la Post Enumerative Survey et les enregistrements du recensement décennal de 1990 avec le processus d'appariement selon le nom et l'adresse utilisé pour le recensement de l'agriculture. Nous procédons actuellement à la définition des fichiers nécessaires à cette évaluation. Des améliorations à la précision des programmes d'uniformisation et d'appariement devraient faire augmenter la proportion d'exploitations uniques dans la liste, diminuer le nombre de fermes dénombrées deux fois et réduire le coût de la vérification manuelle.

4.3 Projets visant à réduire le nombre des dénombrements en double

Nous étudions présentement plusieurs méthodes axées sur la réduction du nombre de fermes dénombrées en double dans le recensement de l'agriculture. Nous envisageons d'apporter des changements liés à l'établissement de la liste d'envoi postal ainsi qu'à la collecte et au traitement des données du recensement. Dans le premier cas, nous tiendrons compte des résultats de l'évaluation des règles d'appariement manuel et automatisé des enregistrements de la liste et du projet d'utilisation d'un programme d'appariement probabiliste. Nous examinons aussi la possibilité d'avoir recours au service de vérification des adresses du US Post Office. Enfin,

nous envisageons d'effectuer une enquête postale ou téléphonique restreinte à partir de certaines des adresses de la liste, du genre de la Farm and Ranch Identification Survey. La vérification des adresses par le Post Office ainsi que l'enquête sont deux mesures qui contribueraient à accroître la précision des adresses de la liste.

L'objectif premier de l'évaluation des règles d'appariement automatisé et manuel est la plus grande précision du processus. Toutefois, nous voulons également éviter d'augmenter le nombre des faux appariements, lesquels influent sur le taux de couverture du recensement. Il est sans doute préférable de conserver un plus grand nombre d'enregistrements en double potentiels dans la liste du recensement et de se tourner plutôt vers les méthodes de collecte et de traitement des données pour essayer de réduire les dénombrements en double. Une enquête postale d'essai suivie d'une réinterview par téléphone évaluera trois modèles différents d'instructions disant au répondant de retourner ensemble des questionnaires se rapportant à une même exploitation. Une telle façon de procéder vise à aider le répondant à faire la différence entre une exploitation individuelle et une exploitation multiple.

Nous étudions un ensemble de méthodes qui pourraient être utilisées en combinaison pendant la vérification des données en vue du repérage des dénombrements en double. Il s'agit d'une méthode d'appariement selon le numéro de téléphone, d'une liste des variables importantes se rapportant à tout l'univers et d'une méthode de tri alphabétique par comté des enregistrements. Nous avons utilisé la liste des variables de l'univers au recensement de 1982, mais non au recensement de 1987. Les numéros de téléphone étaient disponibles au moment de l'introduction des données de 1987, mais seuls quelques États ont utilisé une méthode d'appariement selon le numéro de téléphone lors du traitement des données.

5. TAUX DE COUVERTURE DE L'UNIVERS

5.1 Évaluation du champ couvert par la liste du recensement

L'objectif de tout recensement est de fournir un dénombrement complet de l'univers que l'on veut observer. Il s'agit d'un objectif extrêmement difficile dans le cas du recensement de l'agriculture en raison de l'impossibilité d'identifier les exploitations faisant partie de l'univers afin qu'elles puissent faire l'objet de l'enquête. Les résultats d'évaluations en règle de la couverture effectuées à chaque recensement de l'agriculture indiquent que la couverture de l'univers des fermes n'a jamais été complète, même lorsqu'un dénombrement sur place avait lieu, comme c'était le cas avant 1969. Selon ces résultats, le taux net de couverture variait de 85.0% à 96.6%. (Le taux net de couverture est défini comme le nombre total de fermes recensées divisé par le nombre de fermes estimé dans l'univers. Avant 1982, le numérateur était estimé par l'enquête aréolaire; pour 1982 et 1987, le numérateur correspondait au nombre de fermes de recensement. Toujours avant 1982, l'estimation du nombre de fermes dans l'univers (dénominateur) était fondée sur les échantillons de la base aréolaire et de classification; pour 1982 et 1987, l'estimation de la couverture des fermes était fondée sur l'estimateur de saisie-resaisie.) Bien que la couverture des exploitations agricoles s'établissait à 90% en moyenne pour les recensements précisés dans le tableau ci-dessous, le recensement a permis de dénombrer environ 98% de la production agricole.

Tableau 3: Couverture estimée des fermes de recensement: 1954-1987

ANNÉE DE RECENSEMENT	TAUX NET DE COUVERTURE DES FERMES
1987	92.8
1982	90.9
1978	96.6
1974	89.3
1969	85.0
1964	88.7
1959	91.6
1954	91.9

La meilleure couverture de l'univers des fermes a été réalisée au recensement de 1978. Pour ce recensement, la collecte des données s'est faite à partir d'une base de sondage double, soit un échantillon aréolaire de

segments en plus du dénombrement postal habituel. Une telle méthode a permis une amélioration substantielle de la couverture du recensement par État et pour l'ensemble du pays, particulièrement dans le cas des fermes dont les ventes étaient inférieures à $2 500, la partie de l'univers pour laquelle le dénombrement est le moins complet. En 1978, le pourcentage de ces fermes non recensées était de 6.5%, comparativement à 28.6% en 1982 et 32.3% en 1987. En raison de restrictions budgétaires, il n'a pas été possible de recourir à nouveau à une base de sondage double pour les recensements qui ont suivi.

Le programme d'évaluation de la couverture utilisé pour le recensement de l'agriculture de 1987 a été amélioré de façon à fournir des estimations des fermes ne figurant pas sur la liste d'envoi postal du recensement établie pour chaque État ainsi que des estimations plus fiables des exploitations mal classées ou comprises en double dans la liste d'envoi postal du recensent. Le pourcentage de fermes estimées dans la liste était sensiblement le même en 1987 et 1982, 89.2% contre 89.4%. La réduction de la taille de la liste d'envoi finale de même que l'absence d'une enquête de préclassement n'ont pas vraiment influé sur le degré de couverture de la liste. Les changements liés aux listes de référence utilisées en 1987, l'amélioration de la qualité des enregistrements contenus dans ces listes ainsi que l'efficacité du modèle d'analyse discriminante sont autant de facteurs qui ont contribué à maintenir une couverture constante malgré la réduction considérable du nombre total des enregistrements de la liste d'envoi postal finale. En revanche, le nombre estimé des exploitations dénombrées en double serait deux fois plus élevé qu'au dernier recensement. Aucune des méthodes utilisées en 1987 ne s'est révélée aussi efficace sur le plan de l'élimination des enregistrements en double que l'enquête FRIS menée avant le recensement.

5.2 Effet des méthodes d'établissement de la liste sur la couverture

L'idée derrière l'utilisation d'enregistrements de listes de référence pour établir la liste du recensement était l'élaboration de méthodes pouvant améliorer la couverture des exploitations agricoles dans le cadre du recensement. En particulier, l'enquête de préclassement et les règles précises de couplage automatisé et manuel ont été conçues pour assurer l'exhaustivité de la liste finale dans les limites des contraintes de taille et de coût. Toutefois, en raison des restrictions beaucoup plus sévères qui ont été imposées en 1987 au sujet de la taille de la liste et la tenue d'une enquête de préclassement, il fallait absolument faire en sorte que la liste finale contienne une très forte proportion d'exploitations agricoles pour assurer une couverture adéquate. En remplacement de l'enquête de préclassement, un modèle d'arbre de classification a été élaboré en vue d'atteindre un tel objectif.

Les processus de couplage automatisé et manuel visant à repérer les enregistrements en double influent sur la couverture de la liste. Des règles moins rigides dans ce domaine comportent le risque de l'élimination d'exploitations distinctes potentielles. Les règles actuelles en matière de repérage des enregistrements en double ont tendance à retenir les enregistrements en double potentiels dans la liste. Par exemple, trois enregistrements ayant la même adresse et des noms différents peuvent représenter entre 0 et 3 exploitations différentes; cependant, si les règles de couplage précisent qu'ils correspondent à une seule exploitation et qu'il s'agit en réalité de trois exploitations distinctes, alors deux exploitations pourront être manquées. Ces procédures peuvent augmenter la couverture, mais aussi les risques de dénombrements en double. Une évaluation de ces règles est prévue avant le recensement de 1992.

Une évaluation du modèle d'analyse discriminante d'arbre de classification a été entreprise afin de déterminer si le modèle est efficace pour l'établissement d'une liste d'envoi postal comportant une proportion élevée d'exploitations agricoles. Dans cette évaluation, la proportion d'enregistrements susceptibles de correspondre à des fermes a été comparée à la proportion réelle d'enregistrements représentant des fermes pour chacun des groupes créés par le modèle à partir des enregistrements conservés dans la liste d'envoi postal. L'évaluation a fait ressortir que pour la plupart des groupes du modèle, la proportion de cas réels était légèrement inférieure à ce qui avait été prévue (Schmehl, 1990). Le modèle qui sera utilisé pour le recensement de 1992 pourra tenir compte de plusieurs nouvelles variables comme la raison pour laquelle l'enregistrement avait été classé comme une entreprise non agricole au dernier recensement, les enregistrements du NASS classés comme des entreprises non agricoles ainsi que deux autres codes relatifs aux dossiers du IRS, un code sur la participation du répondant et un code d'activité agricole.

Dans le cadre de l'évaluation du modèle, une enquête-échantillon a été menée en se servant des enregistrements éliminés de la liste d'envoi postal du recensement (à l'exclusion des enregistrements de référence classés comme des entreprises non agricoles). Les enregistrements sélectionnés pour être éliminés de la liste à l'aide du modèle d'analyse discriminante se situaient dans des groupes pour lesquels la proportion des fermes devait être de 11.7% ou moins. Selon les résultats de l'enquête-échantillon, 14.6% des enregistrements qui avaient été éliminés représentaient des exploitations agricoles. Cette estimation devrait comporter un biais à la hausse étant donné que seulement 46.4% des unités de l'échantillon étaient classables et que les répondants qui sont des exploitants agricoles sont plus susceptibles que ceux qui ne le sont pas de répondre à l'enquête. Nous estimons donc que 25 500 des 175 000 enregistrements éliminés de la liste correspondaient à des exploitations agricoles.

D'après les résultats de l'évaluation de la couverture, 242 850 fermes au total ne figuraient pas sur la liste finale du recensement. Ainsi, environ 10% des fermes qui ne figuraient pas sur la liste finale étaient comprises dans la liste provisoire et les 90% qui restent figuraient ou bien sur la liste des adresses non agricoles et ont été éliminées ou bien ne figuraient sur aucune des listes. Par conséquent, le modèle d'analyse discriminante nous a fourni une assez bonne estimation de l'effet de la réduction de la taille de la liste d'envoi finale sur la couverture du recensement. Pour les recensements à venir, il faudrait mettre en balance cette estimation du sous-dénombrement du recensement et les coûts et le fardeau supplémentaire d'une liste d'envoi plus importante.

5.3 Améliorations possibles de la couverture de la liste

Au recensement de l'agriculture de 1978, les estimations des fermes non comprises dans la liste d'envoi postal du recensement pour chacun des États avaient été publiées. À cette fin un échantillon aréolaire et un dénombrement sur place avaient été utilisés. Selon ces estimations, la couverture de la liste du recensement variait considérablement d'un État à l'autre. Cependant, des estimations par État n'ont pas été produites pour le recensement de l'agriculture de 1982 ou pour le programme d'évaluation de la couverture. En l'absence d'estimations pour 1982, il était difficile de déterminer les faiblesses géographiques de la liste en vue de cerner les domaines d'amélioration possibles. Une des principales raisons pour lesquelles le programme d'évaluation de la couverture du recensement a été élargi en 1987 était justement la production d'estimations par État des fermes non comprises dans la liste d'envoi postal du recensement afin de pouvoir évaluer la couverture du recensement au niveau de chaque État. Une telle information était particulièrement importante pour la liste d'envoi de 1987 à cause des changements substantiels apportés aux listes servant de sources aux enregistrements de la liste de recensement (la liste du NASS est passée de 31 à 50 États et la liste de l'ASCS n'a pas été utilisée).

Les estimations par État des fermes ne figurant pas sur la liste d'envoi postal du recensement de 1987 variaient d'un minimum de 1.4% dans le Dakota du Nord à un maximum de 26.6% dans la Virginie de l'Ouest . Des estimations du pourcentage total de fermes américaines qui ne figuraient pas sur la liste d'envoi postal sont présentées pour chacune des divisions de recensement dans le tableau 4. Une analyse de ces estimations révèle que la liste d'envoi postal du recensement fournit la moins bonne couverture dans le cas des États de la Nouvelle-Angleterre et du sud de l'Atlantique et la meilleure couverture pour ce qui est des États du Centre Nord-Ouest et du Centre Nord-Est. En se fondant sur ces données et sur les données par État (tableau 3 du rapport d'évaluation de la couverture du recensement de l'agriculture de 1987), le Census Bureau étudiera diverses méthodes susceptibles d'améliorer la couverture de la liste dans certains États précis. Si l'évaluation des données provenant du ministère de l'Agriculture de l'État du Massachusetts indique que l'utilisation de la liste en question améliore sensiblement la couverture du recensement dans cet État, le Census Bureau envisagera d'acquérir de telles listes dans les autres États.

Tableau 4: Estimations du sous-dénombrement en 1987 selon la division de recensement

	Fermes de recensement - données publiées	Fermes estimées ne figurant pas sur la liste d'envoi postal	Pourcentage de fermes non sur la liste d'envoi postal
États-Unis	2 087 759	249 306	10.7
Nouvelle-Angleterre	25 158	7 767	23.6
Milieu de l'Atlantique	98 324	17 330	15.0
Centre Nord-Est	364 872	33 593	8.4
Centre Nord-Ouest	497 110	23 423	4.5
Sud de l'Atlantique	239 687	56 565	19.1
Centre Sud-Est	249 556	27 767	10.0
Centre Sud-Ouest	334 608	42 987	11.4
Rocheuses	124 210	17 142	12.1
Pacifique	154 234	22 732	13.3

6. COÛTS LIÉS À L'ÉTABLISSEMENT DE LA LISTE

6.1 Facteurs de coûts liés à l'établissement de la liste

Il est difficile de déterminer avec précision le coût réel de l'établissement de la liste d'envoi postal dans le coût global du programme de recensement de l'agriculture qui s'établissait à $65 millions en 1987. Les coûts sont principalement associés aux salaires du personnel professionnel. Une grande partie du travail se fait dans le cadre de projets d'élaboration, de mise en oeuvre et d'évaluation de la liste d'envoi postal entrepris pendant trois des cinq années du cycle du recensement. Les deux autres années sont consacrées aux activités de planification et de recherche. L'équipe de professionnels qui travaillent à l'établissement de la liste doit être formée d'au moins trois statisticiens pendant toute la durée du cycle de l'enquête et de trois programmeurs à l'étape de l'élaboration, de la mise en oeuvre et de l'évaluation.

Au nombre des autres coûts associés à l'établissement de la liste d'envoi postal du recensement de l'agriculture de 1987, on compte l'approvisionnement en enregistrements de référence, le traitement et le géocodage automatisés de ces enregistrements, les salaires du personnel de soutien chargés de vérifier les enregistrements en double potentiels et la préparation des étiquettes des questionnaires. Les enregistrements de référence sont obtenus du Census Bureau à un coût minime étant donné qu'ils sont les produits de la préparation d'autres fichiers de données statistiques et administratives et ils ne constituent donc pas des coûts de collecte distincts. Par exemple, en 1987, nous avons payé le NASS $30 000 pour ses 2.4 millions d'enregistrements et le IRS nous a chargé moins de $125 000 pour ses 6.0 millions d'enregistrements. Le nombre des enregistrements utilisés dans les opérations de couplage aux fins d'établissement de la liste influe sur le coût global du couplage automatisé et sur les coûts de la vérification manuelle effectuée par le personnel de soutien. En 1987, ce nombre était de 5.4 millions (30%) inférieur à ce qu'il était en 1982. Toutes les opérations de traitement automatisées sont exécutées à l'aide de l'ordinateur central du Census Bureau avec les coûts et les frais généraux qui s'y rapportent. Nos estimations de ces coûts sont incomplètes en raison d'erreurs relatives au montants imputés et de changements apportés aux algorithmes d'imputation des coûts. Une indication du montant des frais de personnel de soutien peut être obtenue en considérant le nombre de paires d'enregistrements en double potentiels soumis à la vérification, soit 767 448 paires en 1987 comparativement à 1 332 000 en 1982 pour ce qui est de la première étape du couplage.

Les salaires des professionnels ainsi que les coûts de traitement automatisé associés au modèle d'analyse discriminante se sont révélés peu dispendieux en comparaison des frais considérables des Farm and Ranch

Identification Survey menées en 1978 et 1982. Cependant, la seule application du modèle a été l'élimination de la liste des adresses ne correspondant pas à des exploitations agricoles. Le modèle ne satisfaisait pas aux deux autres objectifs de la FRIS qui sont d'obtenir des données plus à jour concernant les adresses et de repérer les exploitations en double. Les coûts d'une activité indépendante de collecte de données comme cette enquête sont relativement élevés, mais ils sont dans une certaine mesure liés à la taille de l'échantillon (3.0 millions d'unités en 1982). Le coût de chaque unité d'enquête additionnelle est peu élevé compte tenu de l'envergure de cette activité de collecte.

6.2 Coûts des améliorations éventuelles

Les améliorations qui pourraient être apportées à la liste d'envoi postal du recensement de l'agriculture ont déjà été mentionnées dans des sections antérieures de ce document. Ces changements ont été analysés du point de vue des caractéristiques de qualité d'une telle liste: contenu, précision des adresses, caractère unique des enregistrements et couverture de l'univers. Parmi les changements envisagés, il y a l'utilisation de listes d'exploitations agricoles tenues par les États ou autres, la vérification des adresses par le Post Office, la tenue d'une enquête de préclassement à partir de certaines adresses de la liste, l'adoption de nouvelles règles de couplage automatique ou manuel, l'amélioration du programme d'uniformisation des adresses et d'appariement selon le nom et l'adresse et enfin l'amélioration du programme d'analyse statistique discriminante. Dans l'attente de plans précis, nous pouvons uniquement formuler des généralités au sujet de ces coûts.

De façon générale, les sommes déboursées pour l'acquisition des listes d'exploitations agricoles des États et autres organismes pour les besoins de l'établissement de la liste d'envoi postal étaient minimes. Les coûts associés à l'utilisation de nouvelles listes seraient ceux associés au traitement des enregistrements additionnels ainsi qu'aux ressources de programmation nécessaires à l'uniformisation de la présentation des diverses listes. Comme ce type de ressources est rare à l'heure actuelle, nous serions portés à favoriser l'ajout sélectif de nouvelles listes si les fournisseurs ne peuvent pas adopter nos critères de présentation.

La vérification des adresses par le Post Office jouerait un des rôles d'une enquête de préclassement - soit vérifier et corriger les renseignements liés aux adresses. Le coût d'un tel service serait sans doute nettement inférieur à celui d'une enquête de préclassement. Nous pourrions par la suite adresser un questionnaire ou téléphoner aux adresses susceptibles d'être des exploitations en double (par ex. couplage selon le nom de la personne et le nom de l'exploitation). Cette méthode en deux étapes pourrait être une solution plus avantageuse financièrement qu'une enquête à grande échelle comme la FRIS, menée par le passé.

Les règles de couplage automatisé des enregistrements sont présentement conçues en fonction de la non-suppression par l'ordinateur d'enregistrements en double potentiels à moins d'être relativement certains que les enregistrements sont en double. Des règles d'appariement moins sévères pourraient entraîner une diminution du nombre des paires d'enregistrements en double potentiels soumis à une vérification manuelle et donc une réduction des frais de personnel de soutien. Il faudrait cependant mettre en balance l'utilisation de règles moins sévères et la perte éventuelle de couverture du recensement. Toutefois, les coûts de vérification par du personnel de soutien pourraient de toutes manières augmenter en 1992, que des changements soient ou non apportés aux règles de couplage. En effet, le Postal Office prévoit apporter des changements au niveau des adresses rurales en passant d'un système de routes rurales et de casiers postaux à un système de numéros de voirie. Une des conséquences possibles d'un tel changement est l'accroissement du nombre des enregistrements ayant le même nom, mais des adresses différentes et donc du nombre des enregistrements en double potentiels.

Il est difficile pour le moment d'évaluer les répercussions sur les coûts de changements au programme d'uniformisation des adresses, au programme d'appariement selon le nom et l'adresse ou au modèle d'analyse discriminante. Les deux premiers programmes font l'objet de recherches. On étudie notamment la possibilité d'avoir recours à des mini-ordinateurs à la place de l'ordinateur central, ce qui ne veut pas dire que nous pourrions exécuter sur mini-ordinateur l'ensemble des opérations d'établissement de la liste d'envoi postal de 1992. La question la plus importante est sans doute le coût des ressources nécessaires au développement d'une méthodologie appropriée, compte tenu du type de programmation choisi. Quant au modèle, les changements qui devraient être apportés aux contrôles en vue d'améliorer la précision des estimations de la proportion des fermes dans chaque groupe sont très peu coûteux. L'ajout de nouvelles variables exigerait pour sa part des ressources professionnelles ainsi que des crédits de traitement informatique supplémentaires.

7. RÉSUMÉ

L'établissement d'une liste d'envoi postal de grande qualité pour une activité de collecte de données importante n'est pas une tâche facile. Cela exige un programme permanent de recherche, d'évaluation et d'élaboration, que la liste soit tenue et mise à jour périodiquement ou recréée cycliquement comme c'est le cas de la liste du recensement de l'agriculture. Tout changement au niveau des enregistrements de référence ou des méthodes de livraison postale ne peut qu'influer sur le processus d'établissement de la liste, exigeant l'adoption de nouvelles techniques. L'objet même pour lequel la liste a été conçue est un facteur dont on doit tenir compte avant de fixer des exigences de qualité.

L'établissement d'une liste d'exploitations agricoles en vue de mener un recensement comporte une série de défis de taille. Il est fort peu probable qu'on puisse un jour mener un recensement de presque toutes les exploitations agricoles correspondant à la définition du recensement à partir d'une seule liste. Depuis plusieurs recensements, le taux de couverture des exploitations agricoles se situe dans le voisinage de 90%. Toutefois, dans le cas des exploitations plus importantes, ce taux grimpe à 95% et la couverture des activités agricoles et économique atteint les 98%. Le seul maintien d'un tel degré de couverture exige l'étude constante d'améliorations possibles.

Comme le recensement constitue la seule source de données détaillées par comté et de données sur des produits agricoles rares, il est nécessaire de produire une liste d'envoi aussi complète que possible en utilisant à cette fin des enregistrements de nombreuses sources. Aux États-Unis, les dossiers fiscaux se sont révélés une source de renseignements essentielle puisqu'elle représente à elle seule 11.5% des enregistrements de la liste finale de 1982. Ce résultat peut être attribué en partie à la rotation importante des exploitations agricoles d'un recensement à l'autre (seulement 71% des fermes dénombrées en 1978 étaient encore des fermes en 1982). Nous nous retrouvons donc avec le problème constant de l'identification des exploitations qui ont abandonné les affaires et de la recherche des nouvelles exploitations.

Les nombreuses formes juridiques différentes sous lesquelles les fermes et les ranchs sont exploités ont nécessairement un effet sur les enregistrements en double qu'on retrouve dans la liste. Comme nous avons pu le constater en 1987, notre système général de traitement du recensement doit être doté de contrôles pour aider à résoudre le problème inévitable de l'envoi en double de questionnaires aux adresses figurant sur la liste. Nous n'avions pas prévu que l'élimination de l'enquête de préclassement avant le recensement aurait un tel effet sur le nombre des dénombrements en double. Nous devons toujours évaluer tout changement de méthodologie visant à accroître la qualité d'un aspect de la production de la liste en fonction de ses répercussions possibles sur d'autres critères de qualité.

REMERCIEMENTS

Le système de couplage de base des enregistrements de la liste d'envoi postal du recensement de l'agriculture a été mis au point en 1978 sous la direction de D. Dean Prochaska avec le concours précieux de Ralph Graham, Jane Sandusky, Tommy W. Gaulden, Billy Stark, Steve Schobel et Steve Hess. Ces mêmes personnes sont à l'origine des modifications et des améliorations proposées et mises en oeuvre en 1982 et 1987. Dedrick Owens, sous la direction de Ruth Ann Killion, a élaboré le modèle d'arbre de classification d'analyse statistique discriminante qui a été utilisé en 1987 en remplacement d'une enquête de préclassement. Le lecteur trouvera dans le corps de ce document des références aux évaluations officielles des divers aspects du processus de collecte postale des données qui ont été réalisées par le personnel de la Division de l'agriculture. Charles P. Pautler Jr., Tommy Gaulden et Jane D. Sandusky ont accepté de fournir renseignements, critiques et suggestions utiles pour ce document. Les travaux graphiques sont l'oeuvre de Melody Atkinson, Diana Marz et Antoinette Wooten. Le manuscrit des actes du symposium a été préparé par Sharon Lewis et Rosemary Vespoint.

BIBLIOGRAPHIE

1987 Census of Agriculture, U.S. Summary and State Data, 1, 71, Appendices A et C.

1987 Census of Agriculture Coverage Evaluation Report, 2, 2.

Clark, C.Z.F. (1989). Mail Data Collection Methodology and Research for the U.S. Census of Agriculture, *Proceedings of the 47th Session of the International Statistical Institute.*

Clark, C.Z.F. (1985). Mail Enumeration in the United States Census of Agriculture, *Proceedings of the 1991 Census Planning Conference*, Statistique Canada.

Clark, C.Z.F. (1984). Comparability of Data from the Censuses of Agriculture, *1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Davie, W.C. (1986). Special Study on Farms Undercounted in the 1982 Census of Agriculture Due to Misclassification, U.S. Census Bureau internal report.

Davie, W.C., Marz, D., et Wooten, A. (1986). 1986 Census/NASS List Study Matching Procedure, U.S. Census Bureau internal report.

Davie, W.C., Lorenzen, E., et Prochaska, D.D. (1984). Coverage Evaluation for the 1982 Census of Agriculture, 1984 *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Dea, J.Y., Gaulden, T.W. et Prochaska, D.D. (1984). Record Linkage for the 1982 Census of Agriculture Mail List Development Using Multiple Sources, *1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Gaulden, T.W. (1990). Development of the 1987 Census of Agriculture Mail List, U.S. Census Bureau internal report.

Gaulden, T.W. (1985). Evaluation of Mail List Sources Using the Mail List Research Sample, U.S. Census Bureau internal report.

Gaulden, T.W. (1986). Census/NASS List Match Study, U.S. Census Bureau internal report.

Lorenzen, E.B. (1986). Special Study of Overcounted Farms in the 1982 Census of Agriculture Coverage Evaluation, U.S. Census Bureau internal report.

Owens, D., Killion, R.A., Ramos, M., et Schmehl, R. (1989). Classification Tree Methodology for Census Mail List Development, *1989 Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Ruggles, D.R., Dea, J.Y., Kwok, F.K., et Carman, C.A. (1984). Evaluation of the Effectiveness of Data Collection Procedures for the 1982 Census of Agriculture, *1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Schmehl, R., et Ramos, M. (1990). Evaluation of Classification Tree Methodology for Census Mail List Development, *1990 Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Wright, K.E., Davie, W.C., et Sandusky, J.D. (1989). 1987 Coverage Evaluation Estimation, 1989 *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

AMÉLIORATION DE LA QUALITÉ DES DONNÉES DE LA LISTE DES ÉTABLISSEMENTS STATISTIQUES TYPES

P.S. Hanczaryk et M.L. Trager[1]

RÉSUMÉ

La liste des établissements statistiques types (LEST) est un fichier informatique normalisé, ou registre, de toutes les entreprises américaines et de leurs établissements. Le Bureau of the Census dresse, met à jour et utilise de façon permanente cette liste établie à partir des dossiers administratifs d'autres organismes gouvernementaux et des données recueillies par le Census Bureau. Les données d'entrée sont soumises à une série de programmes d'assurance de la qualité et de contrôle ayant pour objet d'en vérifier l'exactitude. Au cours des dernières années, nous avons élaboré plusieurs nouvelles techniques qui nous ont permis d'améliorer la qualité de la LEST sans entraîner d'augmentation du nombre de préposés à l'analyse. Figurent au nombre de ces techniques: un programme machine automatisé qui explore le nom commercial de l'entreprise à laquelle correspondent les enregistrements non classés et attribue à ces derniers le code de classification approprié; un programme de vérification des établissements qui met sur ordinateur les principes et les raisons invoqués par les analystes à l'étape de la correction; enfin, un réseau local de micro-ordinateurs qui permet l'accès en direct aux 7 millions d'enregistrements sur les établissements de la LEST.

MOTS CLÉS: Exactitude; automatisation; vérification.

1. INTRODUCTION

Un des principaux projets mis en oeuvre par le Census Bureau dans le secteur économique vise à améliorer la qualité des données de la LEST tout en respectant les limites financières imposées par un budget restrictif. Au cours des dernières années, l'amélioration de la qualité de la LEST s'est souvent accompagnée d'un accroissement des dépenses. Bien sûr, il est tout naturel que la qualité des données s'améliore lorsqu'on dispose de crédits plus élevés.

Or, compte tenu des importants déficits budgétaires fédéraux prévus dans un avenir immédiat, nous ne prévoyons, au cours des prochaines années, aucun accroissement du financement de la LEST en termes réels. Malgré tout, le Census Bureau s'est engagé à assurer l'amélioration continue de la qualité de la LEST. Pour respecter la politique générale de l'administration en matière de dépenses, ces améliorations devront être le fruit d'une utilisation plus efficace des ressources disponibles plutôt que d'une augmentation des crédits.

Récemment, le Census Bureau a pu réaliser des économies considérables en automatisant certaines opérations de routine auparavant exécutées par des commis ou des analystes. Ces processus automatisés, initialement mis en oeuvre à titre de mesures de réduction des coûts, ont également permis d'améliorer la qualité des données de la LEST. Le présent article donne une description détaillée de trois des opérations analytiques ayant été automatisées. Nous croyons que l'automatisation de ces opérations et d'autres opérations semblables nous permettra d'améliorer la qualité de notre liste sur les entreprises tout en respectant les contraintes budgétaires actuelles.

---

[1] P.S. Hanczaryk et M.L. Trager, U.S. Department of Commerce, Bureau of the Census, Washington, D.C. 20233 U.S.A.

La LEST est une liste informatisée de toutes les entreprises américaines et de leurs établissements. Le Census Bureau met à jour et utilise de façon permanente cette liste qui contient des données économiques de base sur environ 7 millions d'entreprises avec salariés[2]. Tous les cinq ans, des données sur un autre 7.5 millions d'entreprises sans salarié sont recueillies dans le cadre des recensements de l'économie. La liste comporte également un volet distinct sur les fermes, rassemblé une fois tous les cinq ans à partir des résultats du recensement de l'agriculture. Le nombre d'entreprises agricoles répertoriées dans la LEST est d'environ 3 millions.

Toutes les données nécessaires à la mise en oeuvre des programmes économiques et agricoles du Census Bureau sont systématiquement vérifiées et intégrées dans la LEST. La liste renferme pour chaque établissement des données sur le nom commercial principal et secondaire, les adresses postales et de voirie, les codes de la Classification type des industries (CTI), des données sur l'emploi et la paie, des données sur les ventes et les recettes, ainsi que nombre d'autres éléments d'information. La LEST fait l'objet, de façon continue, de mises à jour visant à refléter les renseignements les plus récents recueillis dans le cadre des recensements de l'économie et de l'agriculture ainsi que des nombreuses enquêtes réalisées par le Census Bureau. La liste est aussi régulièrement mise à jour à l'aide des données administratives reçues du Internal Revenue Service (IRS) et de la Social Security Administration (SSA).

Les analystes et le personnel de bureau soumettent les données de la liste à diverses mesures de contrôle visant à en assurer l'intégralité et la fiabilité. Figurent au nombre de ces mesures: le contrôle du contenu des fichiers de données reçus de sources administratives; l'évaluation de l'intégralité des données recueillies chaque année dans le cadre de la Company Organization Survey ainsi que de la couverture des entreprises à établissements multiples; le contrôle des relations annuelles entre les données agrégées à l'échelle de l'industrie et à l'échelle du comté; l'attribution de codes CTI aux établissements d'après le nom commercial; enfin, nombre d'autres opérations. Conjuguées à nos mesures informatisées de contrôle et d'imputation normalisées, ces opérations manuelles permanentes nous permettent de dresser une liste complète et exacte des établissements commerciaux faisant partie de l'univers étudié.

Le Census Bureau a entrepris un examen approfondi des diverses opérations effectuées par les analystes et le personnel de bureau en relation avec la LEST afin de déterminer quels sont les processus susceptibles d'être automatisés. Nous avons jusqu'à maintenant réalisé des progrès sensibles dans trois domaines: l'attribution de codes de classification des industries à partir de la description écrite de l'entreprise ou de son nom commercial, la correction des valeurs des données au cours des opérations de contrôle exécutées par les analystes, ainsi que l'élaboration d'un réseau de micro-ordinateurs permettant d'avoir un accès direct aux enregistrements de données de la LEST. Bien que nous travaillions encore à perfectionner ces processus, l'automatisation de ces tâches nous a déjà permis d'améliorer considérablement la qualité des données et de réaliser d'importantes économies.

Les deux premières opérations, soit l'attribution des codes de classification et la correction des valeurs des données, ont traditionnellement été exécutées par des employés qualifiés. Nous avons essayé de remplacer ces opérations manuelles par des programmes informatisés incorporant les méthodes et les techniques des analystes. La troisième opération, à savoir l'élaboration d'un réseau de micro-ordinateurs, permet d'avoir accès en direct aux 7 millions d'enregistrements sur les établissements de la LEST. Ce réseau, qui utilise nombre de nouvelles techniques de micro-informatique, nous a permis d'éliminer plusieurs opérations manuelles fastidieuses de l'une de nos principales procédures de contrôle. Le présent article explore certaines des caractéristiques de ces trois opérations.

---

[2] Entreprises avec salariés pendant l'année en cours et/ou l'année précédente.

2. ATTRIBUTION AUTOMATIQUE DES CODES DE LA CLASSIFICATION DES INDUSTRIES

2.1 Aperçu

Le système CTI et les systèmes connexes de classification des industries permettent de grouper les statistiques sur les entreprises selon un large éventail de catégories en vue de leur analyse et de leur diffusion. Le Census Bureau effectue certaines opérations visant à convertir la description des activités d'une entreprise ou le nom commercial de cette entreprise en code de la classification des industries. Auparavant, ces opérations étaient confiées à une équipe de préposés au codage. Or, bien qu'un commis soit souvent capable de reconnaître les termes clés de la réponse en toutes lettres et de déterminer quel est le code approprié, il n'est pas facile, en raison des nombreuses variations possibles pour chaque activité industrielle, d'attribuer ces codes à l'aide d'un programme machine.

Afin de surmonter cette difficulté, nous avons tenté de programmer l'ordinateur afin qu'il applique les techniques utilisées par les préposés au codage. À cette fin, nous avons d'abord élaboré un fichier de référence sur les réponses rencontrées fréquemment au cours de l'opération de classification et susceptibles d'être codées avec un haut degré de précision, puis nous avons déterminé quels étaient les mots clés à l'intérieur de chacune de ces réponses. Le fait d'éliminer les mots superflus et d'apparier les réponses uniquement avec des mots clés nous a permis de normaliser les réponses et, dans de nombreux cas, d'attribuer des codes de classification appropriés.

2.2 Codage automatisé des descriptions d'entreprises

Pour les fins du recensement de l'économie de 1987, le Census Bureau devait attribuer un code de la classification des industries à plus de 2.5 millions d'entreprises sans salarié, le coût estimatif de cette opération manuelle s'établissant à $500 000. Afin de réduire ces importants frais de codage et d'accélérer le processus de codage, nous avons élaboré un programme machine automatisé qui explorait le champ réservé à la description écrite et attribuait les codes de classification appropriés. Au total, nous avons traité à l'aide de ce programme 2 563 537 enregistrements non classés. De ce nombre, 1 675 505 enregistrements (soit environ 66.4% du total) se sont vu attribuer un code d'industrie par le programme et les frais de traitement afférents au codage automatisé ont été inférieurs à $500.

Outre ce programme de codage des descriptions, le Census Bureau devait élaborer un programme permettant d'attribuer un code d'industrie aux enregistrements sur les établissements de la LEST à partir de la raison sociale ou du nom commercial de l'entreprise. Bien qu'il soit possible que le nom commercial ne reflète pas l'activité industrielle exercée par l'entreprise, nous verrons qu'il peut être très avantageux d'en automatiser la procédure de codage.

2.3 Codage automatisé des noms commerciaux

Au cours des années de recensement de l'économie, nous déployons des efforts considérables pour attribuer un code à toutes les entreprises avec salariés répertoriées dans la LEST. Lorsqu'on ne peut obtenir les codes à partir des formules du recensement ou d'autres sources fiables, nous déterminons la nature de l'activité industrielle exercée en postant des formules de classification aux établissements visés. Naturellement, le pourcentage d'établissements non classés au cours des années de recensement de l'économie est plutôt faible.

Au cours des années intercensitaires, nous faisons plus largement appel aux dossiers administratifs pour classer les établissements. En particulier, nous utilisons les codes de classification des industries obtenus de la SSA et du IRS pour mettre à jour les codes CTI de nombreux établissements nouvellement répertoriés dans la LEST. Cependant, comme nombre de codes administratifs ne nous sont pas transmis à temps pour nous permettre de mettre à jour la LEST, il est possible que nos fichiers de la LEST comprennent jusqu'à 500 000 enregistrements sur des établissements non classés pour une année intercensitaire donnée. Or il importe de réduire le nombre de ces enregistrements puisque la LEST sert de source de données pour nos publications sur les Country Business Patterns (CBP) et divers autres programmes économiques annuels.

En vue de réduire le pourcentage d'enregistrements non classés, le personnel de bureau procède chaque année à un codage des noms commerciaux, qui consiste à attribuer un code CTI aux enregistrements non classés en se fondant sur le nom commercial de l'entreprise. Comme il est fréquent que ce nom ne contienne aucune indication de la nature de l'activité industrielle exercée, nombre d'enregistrements ne peuvent être codés. Généralement, le personnel de bureau peut attribuer un code à environ 40% des enregistrements non classés de la LEST. Si on suppose que le nombre de ces enregistrements est égal à 500 000, le personnel devrait donc être en mesure de coder 200 000 enregistrements tandis que les 300 000 autres demeureraient non classés.

Le Census Bureau a récemment procédé à une évaluation de l'opération de codage manuel des noms commerciaux. À cette fin, nous avons posté des formules de classification à 13 285 établissements s'étant vu attribuer un code d'industrie par codage manuel du nom commercial, puis nous avons comparé le code attribué au code obtenu à l'aide de la formule de classification (lequel code était considéré comme exact). Les résultats ont indiqué que 58.9% des enregistrements s'étaient vu attribuer un code à 4 chiffres exact, 64.4% un code à 3 chiffres exact, 70.2% un code à 2 chiffres exact, et 81.2% un code de division d'industries exact.

Bien que ces résultats nous aient déçus, l'évaluation nous a permis de cerner les groupes d'établissements pour lesquels on obtenait des codes incohérents. Ainsi, nous avons découvert que les codes attribués étaient beaucoup moins fiables pour les divisions de l'exploitation minière et du commerce de gros que pour les autres divisions. Le Census Bureau a donc mis en oeuvre une stratégie visant à éliminer les codes correspondants attribués au cours de l'opération de codage manuel des noms commerciaux.

Par ailleurs, l'évaluation nous a permis d'élaborer un fichier d'essai idéal pour effectuer des travaux de recherche sur le codage automatisé des noms commerciaux. Le personnel du Bureau a établi, à partir des résultats de l'évaluation, une liste de référence, constituée de mots clés et de combinaisons de mots clés, pouvant servir à déterminer avec un degré élevé de précision la nature de l'activité industrielle exercée par un établissement. Nous avons ensuite rédigé un programme machine permettant de solliciter cette liste de référence - lorsque le nom commercial figurant dans l'enregistrement de la LEST contient la combinaison appropriée de mots clés, le code CTI correspondant lui est attribué.

Le fichier d'essai codé nous a pour sa part permis de déterminer l'exactitude des diverses combinaisons de mots clés figurant sur la liste de référence. Les combinaisons qui ne satisfaisaient pas à des normes de qualité minimales ont alors été éliminées de la liste. À cette fin, nous avons trié les enregistrements de l'évaluation selon le code CTI attribué au cours de l'opération de codage des noms commerciaux et selon le code CTI définitif (c.-à-d. selon le code CTI attribué à l'aide de la formule de classification envoyée par la poste). Ensuite, nous avons effectué une évaluation systématique de l'exactitude des codes à 4 chiffres relatifs à chaque groupe d'industries. En règle générale, seuls les codes CTI ayant été attribués correctement dans un pourcentage de 67% à l'échelle du code à 4 chiffres et dans un pourcentage de 80% à l'échelle du code de division d'industries ont été retenus dans le système de codage automatisé.

2.4 Résultats

Les résultats de l'essai de codage automatisé des noms commerciaux ont été fort satisfaisants: 83.9% des enregistrements se sont vu attribuer un code à 4 chiffres exact, 86.2% se sont vu attribuer un code à 3 chiffres exact, 89.1% se sont vu attribuer un code à 2 chiffres exact, et 94.3% se sont vu attribuer un code de division d'industries exact. Bien que les codes attribués par le programme automatisé ne soient pas aussi fiables que les codes CTI obtenus par le Census Bureau à l'aide des formules de classification envoyées par la poste, leur fiabilité est supérieure à celle des codes attribués manuellement à partir des noms commerciaux.

En plus d'automatiser l'attribution des codes CTI, le programme permet de séparer les enregistrements non codés en deux groupes: ceux qui sont susceptibles d'être codés à la main et ceux qui ne sont pas susceptibles de l'être. Comme nous ne soumettons aux préposés au codage que les enregistrements susceptibles d'être codés, leur charge de travail s'en trouvera d'autant réduite.

Nous prévoyons que l'adoption de cette approche systématique à l'égard du codage des noms commerciaux (attribution automatisée de certains enregistrements non codés couplée à l'élimination des enregistrements incodables de l'opération de codage manuel) permettra au Census Bureau de réaliser des économies importantes

tout en améliorant de beaucoup la qualité des données. De fait, elle se traduira par une réduction de 50% de la charge de travail du personnel affecté au codage manuel des noms commerciaux.

2.5 Méthodologie

Comme nous l'avons mentionné ci-devant, nous avons élaboré une liste de référence constituée de mots clés et de combinaisons de mots clés utilisés pour attribuer un code CTI à l'aide du programme de codage automatisé des noms commerciaux. Dans certains cas, il suffit que le nom commercial comporte un seul mot clé pour qu'il y ait appariement. Ainsi, un nom commercial comprenant le mot "Accounting" se verrait attribuer le code correspondant à "Accounting and bookkeeping", sans qu'il soit besoin d'un autre qualificatif. Par ailleurs, si le nom commercial comporte deux mots clés incompatibles (par ex. Lawyer/Accounting ou Accounting Books), il ne pourra être codé.

La plupart des noms commerciaux doivent toutefois comporter plus d'un mot clé pour pouvoir être appariés. Ainsi, bien que le terme "Appliance" soit un mot clé, il ne sera codé que s'il est accompagné d'un autre mot clé. Les combinaisons de mots clés "Appliance Store" et "Appliance Center" se verront attribuer le code correspondant à "Household appliance stores", tandis que les combinaisons "Appliance Repair" et "Appliance Repairs" se verront attribuer le code correspondant à "Electrical and electronic repair shops".

La première opération exécutée par le programme de codage des noms commerciaux consiste à décomposer le nom commercial d'entrée en mots distincts. À cette fin, les mots sont définis comme des caractères alphabétiques consécutifs délimités par des espaces ou par d'autres caractères non alphabétiques. Ensuite, chaque mot est comparé à une liste de mots clés constituée de mots distincts tirés de la liste de référence. Tous les mots du nom commercial ayant pu être appariés à un des mots de la liste de mots clés sont alors concaténés, puis comparés à une liste prédéterminée de combinaisons de mots clés. Enfin, les mots concaténés pouvant être appariés à l'une des combinaisons de la liste se voient attribuer le code CTI approprié.

L'exemple ci-après illustre les diverses opérations exécutées par le programme pour coder le nom commercial "ABC TERMITE & PEST CONTROL".

Description d'entrée "ABC TERMITE & PEST CONTROL"

1re opération:	Conversion des caractères non alphabétiques en espaces (la perluète devient un espace).	"ABC TERMITE PEST CONTROL"
2e opération:	Décomposition du nom commercial d'entrée en mots distincts.	Wd 1 ABC Wd 2 TERMITE Wd 3 PEST Wd 4 CONTROL
3e opération:	Comparaison de chaque mot à une liste de mots clés	Wd 1 ABC - Pas d'appariement Wd 2 TERMITE - Appariement Wd 3 PEST - Appariement Wd 4 CONTROL - Appariement

4e opération: Concaténation des mots appariés à un des mots de la liste: "TERMITEPESTCONTROL"

5e opération: Appariement de la chaîne concaténée à la liste de référence de descriptions (également concaténées). Attribution du code d'industrie à 4 chiffres approprié aux enregistrements appariés:
"ABC TERMITE & PEST CONTROL" Le code CTI 7342 est attribué.

Cette technique nous permet d'apparier des mots clés sans qu'il soit nécessaire de vérifier si ces mots apparaissent dans chaque chaîne de caractères. Ainsi, les noms commerciaux "ABC TERMITE & PEST

CONTROL", "JOHNSONS TERMITE AND PEST CONTROL" et "TERMITE AND PEST CONTROL OF DADE COUNTY" se verront attribuer le même code puisqu'ils comprennent les mêmes mots clés.

L'illustration ci-devant a pour objet de donner un aperçu général du fonctionnement du programme, pour plus de clarté, nous avons omis d'y mentionner plusieurs caractéristiques de ce dernier. Or j'aimerais maintenant exposer le mécanisme utilisé pour repérer les enregistrements qui ne sont pas susceptibles d'être codés à la main. Comme nous l'avons vu ci-devant, afin de réduire l'envergure de l'opération de codage manuel, nous ne remettrons aux préposés au codage que les enregistrements susceptibles d'être codés.

De nombreux établissements non classés ne peuvent pas être codés manuellement parce que le nom commercial de l'entreprise n'indique pas la nature de l'activité industrielle qu'elle exerce. Ainsi, il est clair qu'on ne saurait attribuer un code à des entreprises possédant un nom commercial du genre "ABC Corporation" ou "JJ Enterprises". Il était donc essentiel, afin d'accroître l'efficacité de l'opération de codage manuel, d'éliminer ce type d'enregistrement du fichier de codage.

Il est toutefois possible de coder à la main certains autres enregistrements n'ayant pas été codés à l'aide du système automatisé, surtout du fait que la liste de référence n'est pas exhaustive. En effet, compte tenu du nombre incalculable de mots clés pouvant figurer dans les noms commerciaux et du temps limité mis à la disposition du personnel, nous n'avons pas été en mesure de prévoir toutes les combinaisons possibles. Les noms commerciaux énumérés ci-après illustrent bien le genre d'enregistrements n'ayant pas été codés à l'aide du système automatisé, mais pouvant être codés à la main:

- SUITLAND HOSPITAL PROFIT SHARING PLAN
- CAROL'S GIFT AND NOVELTY SHOPPE
- JEFFERSONVILLE MOPED SALES
- CANDY STORE - ABC CANDY
- etc.

Bien que chacune de ces descriptions renferme des mots clés, les combinaisons de ces mots clés n'ont pu être appariées à aucune des combinaisons de la liste de référence. Ainsi, le nom commercial "SUITLAND HOSPITAL PROFIT SHARING PLAN" comporte quatre mots clés pouvant être appariés, à savoir "HOSPITAL", "PROFIT", "SHARING" ET "PLAN". Bien que notre liste de référence renferme l'expression concaténée "PROFITSHARINGPLAN", elle ne comprend pas l'expression plus spécifique "HOSPITALPROFITSHARINGPLAN" et le nom commercial n'a pu être codé. Il ne fait aucun doute qu'un préposé au codage qualifié réaliserait que ce nom commercial indique la mise en oeuvre d'un régime de participation aux bénéfices et serait en mesure d'attribuer le code approprié[3].

En attendant l'établissement d'une liste de référence complète, nous avons tenté d'élaborer une méthode pratique de repérage des noms commerciaux susceptibles d'être codés à la main. Nous en sommes venus à la conclusion que les noms commerciaux ne comportant aucun mot clé n'avaient que très peu de chances d'être codés à la main, tandis que les mots comportant au moins un mot clé étaient fort susceptibles de l'être. En envoyant les enregistrements non codés dans deux fichiers de sortie distincts, selon qu'ils comprennent ou non un mot clé, nous avons été en mesure d'établir un mécanisme nous permettant d'atteindre tous nos objectifs. En effet, seul un petit nombre d'enregistrements ne renfermant pas de mot clé peuvent être codés à la main.

Le Census Bureau se propose d'utiliser le système de codage automatisé des noms commerciaux dès le mois de novembre afin d'attribuer des codes d'industrie aux établissements non classés de la LEST. Nous prévoyons être en mesure d'attribuer un code à près de 25% des enregistrements et d'en éliminer environ 30% du processus de codage. En conséquence, nous pourrons réduire dans une mesure de près de 55% le nombre des enregistrements soumis à l'opération de codage manuel des noms commerciaux. En supposant que l'univers de

---

[3] Nous comptons incorporer aux futurs programmes de codage automatisé un mécanisme permettant de tenir compte de tels cas. Certaines combinaisons de mots clés devraient être codées, malgré la présence d'un mot clé additionnel dans le nom commercial.

départ comporte 500,000 enregistrements non classés, nous réduirons d'environ 275 000 le nombre d'enregistrements devant être codés à la main cette année.

Au total, le Census Bureau attribuera donc des codes valides à environ 200 000 enregistrements (125 000 à l'aide du système de codage automatisé et 75 000 au cours de l'opération de codage à la main). Le pourcentage d'enregistrements de la LEST non classés descendra en conséquence d'environ 7.1% à 4.3%.

2.6 Améliorations futures

Bien que nous soyons satisfaits des progrès réalisés jusqu'à maintenant, nous comptons apporter dans le futur un certain nombre d'améliorations au système afin d'en accroître l'utilité. Premièrement, nous prévoyons compléter l'actuelle liste de référence en y ajoutant de nombreuses descriptions supplémentaires. Ces ajouts, qui seront effectués à la lumière des résultats de l'opération de codage manuel de cette année, seront constitués des noms commerciaux ayant été codés par les préposés. Cette liste de référence plus complète permettra d'accroître le pourcentage d'enregistrements codés au cours des opérations subséquentes. Deuxièmement, nous étudierons de quelle façon on peut utiliser les autres éléments d'information figurant dans les enregistrements de la LEST, comme les codes de la forme juridique de l'entreprise ou les chiffres sur l'emploi et la paie, pour faciliter le codage. Troisièmement, nous avons l'intention de normaliser les mots clés en éliminant les pluriels et les suffixes. Ainsi, les termes "Radiology", "Radiologist" et "Radiologists" pourraient tous être représentés par le dérivé de base "Radiolog". Non seulement cette technique assurerait-elle une réduction du nombre de mots clés nécessaires, mais elle pourrait permettre d'accroître le taux d'appariements réussis en prenant en considération toutes les formes possibles d'un mot.

Nous prévoyons apporter ces améliorations et d'autres améliorations semblables au système de codage automatisé des noms commerciaux au cours des deux prochaines années. Nous nous sommes fixé pour objectif de disposer d'un système de codage des noms commerciaux complètement automatisé, ne nécessitant aucune intervention manuelle, d'ici 1993. En outre, nous aimerions adapter cette technique automatisée à d'autres opérations de codage mises en oeuvre au Census Bureau. Ainsi, il est peut-être possible d'élaborer un programme semblable pour attribuer les codes de niveau de produit dans le cadre du recensement des manufactures de 1992.

3. RÉSOLUTION AUTOMATISÉE DES CAS DE REJET AU CONTRÔLE

Le Census Bureau a également réalisé des progrès au chapitre de l'automatisation de certaines opérations de contrôle exécutées par des analystes. Les secteurs économiques produisent de nombreuses listes de cas rejetés au contrôle devant être résolus par les analystes. Souvent, ces derniers résolvent les cas d'établissement à valeurs aberrantes en examinant la relation qui existe entre les divers éléments d'information et indicateurs sur l'enregistrement de la LEST. En donnant une forme exacte aux règles de résolution et aux raisons invoquées par les analystes et en incorporant ces règles et raisons dans un programme machine, nous avons réussi à informatiser certaines opérations auparavant exécutées par des analystes. Ces processus informatisés ne sont ni innovateurs ni complexes, mais visent simplement à quantifier les connaissances des analystes en vue de reproduire leurs jugements.

Dans le cadre de la tâche d'imputation relative à la LEST, les enregistrements rejetés au contrôle sont transmis aux analystes des CBP afin d'être examinés et éventuellement corrigés. Ces enregistrements sont groupés en plusieurs catégories, selon les comparaisons de données effectuées. Les analystes des CBP corrigent les cas de rejet en se fondant sur les relations entre les données et sur d'autres indicateurs figurant sur l'enregistrement de la LEST. Toutefois, ces analystes ne disposent que d'un temps limité pour traiter les listes de cas de rejet, qui comptent plus de 125 000 enregistrements sur les établissements. Les résultats de ces corrections doivent permettre de mettre la LEST à jour à l'intérieur d'un délai de quelques semaines, avant la clôture des activités de traitement de fin d'année. En raison de ces contraintes, les analystes des CBP examinent uniquement les enregistrements à valeurs aberrantes et nombre d'enregistrements rejetés au contrôle ne sont pas corrigés. En conséquence, les vérifications subséquentes des analystes des CBP sont beaucoup plus difficiles à traiter.

Nous étions déterminés à élaborer une procédure automatisée permettant de corriger adéquatement les enregistrements rejetés au contrôle; il nous apparaissait essentiel d'examiner tous les enregistrements rejetés au contrôle et non pas seulement les enregistrements à valeurs aberrantes. Dans un premier temps, nous avons demandé à trois de nos meilleurs analystes de corriger un échantillon de cas représentatifs tiré des listes de rejets de l'année précédente. Comme il arrive généralement en pareil cas, les analystes ont dû, pour rendre nombre de décisions, exercer leur jugement en sus d'appliquer certains principes reconnus.

En se fondant sur les corrections apportées, les analystes ont ensuite tenté de quantifier l'ensemble de règles utilisées au cours du processus de correction. C'est cette étape du processus d'élaboration qui s'est avérée la plus fastidieuse. En l'occurrence, les analystes ont élaboré ces règles en utilisant une méthode itérative. Des règles de base ont d'abord été formulées et incorporées au programme machine, puis leur fiabilité a été évaluée en exécutant le programme sur un fichier échantillon d'enregistrements rejetés au contrôle. Plusieurs exceptions et ajouts éventuels aux règles ont ainsi été décelés et graduellement intégrés au programme. Ce processus a été répété à plusieurs reprises en vue d'établir un ensemble complet de règles prévoyant toutes les variantes nécessaires.

Le programme machine résultant nous a permis d'obtenir des résultats exceptionnels. Selon une comparaison entre les corrections apportées par le programme et les corrections apportées par les analystes des CBP, les résultats obtenus à l'égard des enregistrements corrigés des deux façons ont été remarquablement similaires, mais le programme a permis de coder un nombre beaucoup plus élevé d'enregistrements. À titre d'exemple, qu'il suffise d'indiquer que le processus manuel a permis d'apporter moins de 750 corrections pour l'année de référence 1987, tandis que le programme machine a respectivement généré 6 635 et 7 402 corrections pour les années de référence 1988 et 1989.

Nous réalisons que les conclusions atteintes à l'aide d'un système automatisé fondé sur le jugement et les techniques des analystes ne sauraient être parfaitement justes. Toutefois, le système est capable de rendre des décisions presque optimales à partir des données dont on dispose. En outre, l'utilisation d'un tel processus systématique assure que les corrections apportées sont cohérentes, tandis que les corrections apportées par les analystes peuvent varier en fonction des méthodes utilisées par chacun.

La mise en forme définitive et l'automatisation des méthodes utilisées par les analystes aux fins de cette opération ont permis au Census Bureau d'améliorer de beaucoup la qualité des données tout en réalisant des économies importantes. Cette approche ne prévoit aucune nouvelle méthode d'imputation ni stratégie de contrôle, mais remplace simplement les corrections effectuées par les analystes par des corrections automatisées.

4. RÉSEAU INTERACTIF DE MICRO-ORDINATEURS

4.1 Introduction

En outre, le Census Bureau a élaboré un réseau local de micro-ordinateurs qui permet d'avoir accès en direct aux 7 millions d'enregistrements sur les établissements répertoriés dans la LEST. Nous avons élaboré ce réseau afin d'éliminer plusieurs opérations manuelles fastidieuses que nécessitait le contrôle des données par les analystes des CBP. Tout d'abord, en assurant un accès direct aux enregistrements sur les établissements, le réseau interactif permet d'éliminer les encombrantes recherches sur microfilm auparavant utilisées par les analystes. En outre, il permet aux analystes d'introduire les corrections directement dans le système, sans qu'il soit besoin de faire transcrire les données par des employés de bureau.

La mise en oeuvre du réseau s'est traduite par une amélioration sensible de l'efficacité des tâches de contrôle effectuées par nos analystes des CBP. Entre autres, le réseau interactif nous a permis d'améliorer la qualité des données et de réduire les délais de publication. De plus, sa mise en oeuvre a entraîné une diminution des frais de traitement en permettant d'exécuter sur micro-ordinateur les opérations auparavant exécutées sur l'ordinateur central et d'éliminer une large part des activités d'introduction de données au clavier. En sus de nombreux autres avantages, le réseau offre également la possibilité d'établir rapidement et à un coût minime des totalisations spéciales sur l'univers des entreprises.

4.2 Le processus de la vérification du CBP

Les statistiques relatives au CBP fournissent des renseignements détaillés sur la structure industrielle des États-Unis. Plus précisément, les publications sur les CBP présentent des données agrégées sur le nombre d'établissements et l'emploi ainsi que sur la paie au premier trimestre et à la fin de l'année. Ces données sont totalisées par industrie pour tous les comtés et États du pays. Avant leur diffusion, les analystes soumettent les données à diverses mesures de contrôle. La principale de ces mesures, celle du contrôle des cellules, consiste à comparer les totaux récapitulatif de l'année en cours pour les établissements, l'emploi et la paie aux totaux récapitulatifs de l'année précédente à l'échelle de l'État et du comté. Lorsque l'écart entre ces totaux est supérieur aux limites de tolérance établies, on prélève les enregistrements appropriés au sein de la cellule afin de les examiner.

Auparavant, les analystes tentaient de corriger les divergences entre les données en consultant des microfilms qui, contrairement aux listes de contrôle ne comprenant que des renseignements de base, renfermaient des données complètes sur chaque établissement industriel. Ces microfilms existaient en deux versions: sur la première, les entreprises de l'univers étaient classées selon leur numéro d'identification; sur la seconde, elles étaient classées par État, par comté et par code CTI. Toutefois, en raison de contraintes d'espace, la seconde version contenait uniquement des données sur les établissements d'une certaine taille, c'est-à-dire ceux ayant des frais annuels de personnel de $125 000 ou plus. Dans certains cas, cette limite entravait les analystes dans leurs recherches.

Le contrôle des cellules permettait chaque année de repérer environ 10 000 corrections, que les analystes inscrivaient directement sur les listes de contrôle des cellules. Une fois toutes les corrections relatives à un État repérées, le personnel de bureau les transcrivait sur un document de report, puis les introduisait dans l'ordinateur central. Comme la plupart des opérations d'introduction de données au clavier, cette opération donnait lieu à de nombreuses erreurs de transcription ainsi qu'à certaines omissions.

Afin de mieux utiliser le temps de notre personnel et d'améliorer la qualité des données, nous avons élaboré un réseau interactif de micro-ordinateurs. Ce réseau a permis à la section de CBP de réduire de 3 mois le temps consacré à la réalisation des vérifications de contrôle par les analystes (ces vérifications peuvent être effectuées en 5 mois à l'aide du réseau interactif, alors qu'elles nécessitaient 8 mois auparavant). Encore plus important, il nous a permis d'améliorer la qualité des données publiées en rendant possible l'examen de nombre d'autres établissements à valeurs aberrantes, en interdisant certaines corrections illogiques et en éliminant de nombreuses erreurs de frappe.

4.3 Avantages offerts par un réseau interactif

En 1989, nous avons remplacé les fichiers sur microfilm auparavant utilisés pour les recherches analytiques par un seul fichier de référence renfermant des données essentielles pour chaque établissement de l'univers des entreprises. Le fichier a été stocké sur un disque optique auquel chaque analyste des CBP peut avoir accès par l'intermédiaire d'un réseau de micro-ordinateurs. Un mécanisme d'accès multiple a également été élaboré pour permettre à plusieurs analystes d'examiner les mêmes enregistrements simultanément.

Le fichier de référence est indexé de façon à permettre un accès sélectif selon deux séquences distinctes: par ordre de numéro d'identification, afin d'examiner un enregistrement précis, ou par État, par comté et par code CTI. Comme les listes de contrôle des cellules sont dressées par ordre d'État, de comté et de code CTI, la seconde séquence est très utilisée au cours des recherches relatives au contrôle des cellules.

Chaque enregistrement sur les établissements comporte plus de 50 champs, qui correspondent aux besoins d'analyse des CBP. En outre, nous avons établi une série de commandes relatives au contrôle des cellules afin d'accélérer le processus d'examen. Ces commandes permettent la consultation immédiate de diverses catégories d'enregistrements: établissements correspondant au code CTI suivant à l'intérieur du comté et de l'État; établissements figurant dans la première cellule de l'État suivant; établissements ayant cessé leurs activités; etc.

Voici quelques-uns des avantages offerts par le fichier de référence interactif.

- L'accès en direct est virtuellement instantané. Par contraste, la consultation des microfilms était une opération longue et fastidieuse: les bobines de microfilm étaient chargées sur des lecteurs et il fallait parfois plusieurs minutes pour repérer un enregistrement.

- Le réseau interactif permet à de nombreux utilisateurs de consulter les enregistrements simultanément. Auparavant, l'accès multiple était peu commode car il était fréquent que divers analystes aient besoin de la même bande de microfilms.

- Tous les enregistrements d'une cellule d'État/comté/code CTI peuvent être sollicités par l'intermédiaire du réseau interactif. Il s'agit d'un avantage important par rapport au microfilm, qui comprenait uniquement les enregistrements sur les établissements d'une certaine taille.

- Les frais afférents à la préparation du microfilm ont été éliminés. Au cours des dernières années, le Census Bureau devait verser une somme d'environ $15 000 à un entrepreneur pour exécuter ce travail. De plus, avant de remettre les fichiers à l'entrepreneur, il était nécessaire de trier les enregistrements sur l'ordinateur central, moyennant des frais importants. Bien qu'il soit toujours nécessaire de trier les enregistrements pour le réseau interactif, cette opération s'effectue maintenant à l'aide d'un micro-ordinateur, moyennant des frais minimes.

Par ailleurs, grâce au mode interactif, les analystes peuvent introduire les corrections de contrôle directement dans le réseau de micro-ordinateurs, ce qui permet d'éliminer le processus de transcription des données et de réduire considérablement le nombre d'erreurs de frappe. De plus, nous avons introduit des mesures de contrôle afin d'empêcher l'entrée de données erronées. Ainsi, le système ne permet pas de mettre la base de données à jour à l'aide d'un numéro d'identification invalide. En outre, les valeurs des données doivent être soumises à un contrôle de cohérence interne avant que les mises à jour soient acceptées. Ces mesures de contrôle, quoique modestes, nous ont permis de déceler de nombreuses erreurs. Les futures versions du programme permettront d'appliquer des mesures de contrôle plus complètes à l'échelle de l'industrie et de la région géographique.

Notre réseau de micro-ordinateurs vient également répondre à un autre besoin. Dans le passé, le Census Bureau a éprouvé certaines difficultés à répondre aux demandes de totalisations spéciales. Auparavant, ces tableaux étaient produits exclusivement à l'aide de l'ordinateur central. Or, non seulement était-il coûteux d'avoir recours à cet ordinateur, mais le personnel chargé de la programmation des totalisations était affecté à une division distincte à l'intérieur du Census Bureau. Il fallait donc élaborer des spécifications interdivisionnelles et les délais de préparation des tableaux s'en trouvaient prolongés.

Comme la mise à jour de l'univers des entreprises s'effectue maintenant à l'aide du réseau de micro-ordinateurs, il nous est possible de produire des totalisations spéciales de façon rapide et efficace, sans nécessiter l'intervention de diverses divisions. De plus, nous avons élaboré un programme général qui permet tant aux programmeurs qu'aux non-programmeurs de préparer des tableaux facilement.

Bien que nous soyons satisfaits des gains de productivité réalisés grâce au réseau de micro-ordinateurs, nous prévoyons y apporter certaines améliorations au cours des prochaines années. Tout d'abord, nous comptons ajouter de nouveaux éléments d'information aux enregistrements afin d'aider les analystes à prendre leurs décisions. Encore plus important, nous avons l'intention d'offrir aux analystes des possibilités de contrôle interactif leur permettant de faire les corrections directement dans les fichiers de données, de recalculer les totaux récapitulatifs révisés et les valeurs de tolérance, puis de les afficher à l'écran. Ainsi, ils pourront voir immédiatement l'incidence de la correction sur les totaux récapitulatifs. Si cette dernière n'entraîne pas une modification des totaux suffisante pour atténuer le problème, ils pourront alors apporter d'autres modifications.

4.4 Élaboration du réseau interactif de micro-ordinateurs

La première étape de la préparation du fichier de référence interactif consiste à transférer tous les enregistrements de la LEST de notre ordinateur central UNIVAC aux satellites. À cette fin, il faut soumettre

une base de données comprenant 7 millions d'enregistrements, occupant un espace mémoire de près de 2.5 gigaoctets, à plusieurs phases de traitement. Premièrement, le fichier est transféré sur une bande magnétique à 9 pistes (à compter de l'an prochain, grâce à l'installation de la base de données LEST sur un mini-ordinateur de Digital Equipment Corporation (DEC), le réseau à grande distance du Bureau nous permettra de transférer le fichier directement de notre mini-ordinateur DEC à notre réseau de micro-ordinateurs, et donc d'éliminer la fastidieuse étape de la conversion sur bande). Deuxièmement, les fichiers transférés sur bande magnétique doivent être stockés sur disques optiques, lesquels disques servent à la fois de support de sauvegarde et de support d'entrée pour les autres phases du traitement.

Le choix du disque optique comme support d'information s'imposait pour deux raisons: il constitue la façon la plus pratique de stocker une immense quantité de données sur un seul élément matériel (bien qu'à la suite de développements récents, le disque dur offre maintenant une capacité de stockage supérieure à 1 gigaoctet), ensuite, comme il est un support amovible, il permet un accès facile aux ensembles de données des années précédentes (il suffit pour ce faire de remplacer un disque optique par un autre).

Malgré les avantages offerts par les disques optiques et les réseaux de micro-ordinateurs en général, la réussite du présent projet nécessitait la mise en oeuvre d'une autre étape, celle de la compression des données. Le stockage et la sollicitation de données occupant un espace mémoire de 2.5 gigaoctets est une opération relativement longue et coûteuse. Nous étions donc déterminés à réduire au minimum le matériel (et donc les investissements) nécessaire pour construire le système, tout en assurant la plus grande rapidité possible d'accès à la base et d'extraction des enregistrements. La compression de l'ensemble de données a également eu pour avantage de permettre un accroissement spectaculaire de la vitesse de tri, de fusion, d'indexation et d'exécution d'autres étapes de prétraitement.

Avant d'évaluer les techniques de compression des données, nous nous étions fixé plusieurs objectifs que nous jugions essentiels à l'obtention d'un système efficace. Premièrement, nous voulions obtenir un taux raisonnable de compression (d'au moins 50%); deuxièmement, nous voulions utiliser un algorithme assez simple afin de réduire au minimum le temps de traitement consacré à la décompression des données et de permettre de maintenir facilement le code de programme ou de le modifier au besoin; troisièmement, nous voulions qu'il soit possible de comprimer les enregistrements un à un et non pas uniquement le fichier au complet; enfin, nous voulions que l'algorithme permette de comprimer les données sans perte d'information.

Les progiciels de compression de données offerts dans le commerce ne convenaient pas à nos besoins, puisqu'ils assuraient uniquement le traitement d'un fichier complet. Dans une base de données à accès sélectif, il est possible d'extraire un seul enregistrement de n'importe quelle adresse du fichier. Or, à l'évidence, les techniques de compression devant utiliser le fichier au complet pour constituer des groupes de caractères répétitifs seraient incapables d'assurer la décompression dynamique d'un seul enregistrement à la fois. De plus, les progiciels commerciaux ne pourraient être facilement intégrés à notre logiciel personnalisé de gestion de base de données et ne pourraient certainement pas être maintenus ou modifiés au besoin.

Nous nous sommes donc résignés à rédiger notre propre logiciel de compression - décompression. À cette fin, il nous fallait d'abord choisir un algorithme approprié. Comme nous l'avons vu, l'algorithme devait nous permettre de comprimer et de décomprimer un ensemble de données sans perte d'information. Or nombre de techniques de compression, surtout celles utilisées aux fins de la compression des images, ne peuvent assurer une restauration littérale des données initiales. Pour le traitement de l'information graphique, cette perte d'information n'est pas fatale et il arrive même qu'elle soit imperceptible. Cependant, dans notre univers des entreprises, même la perte d'un seul bit pourrait avoir de graves conséquences. Nous avons donc réduit le champ de nos recherches à un petit nombre d'algorithmes prometteurs, dont le mieux connu est sans doute la technique Lempel-Ziv Welch (LZW). Compte tenu des objectifs que nous nous étions fixés, la technique LZW présentait toutefois deux inconvénients majeurs: il s'agit d'un algorithme complexe qui, au mieux, aurait rendu difficile le maintien et la modification du logiciel, surtout pour les utilisateurs ignorant tout des complexités de la technique; en outre, comme la technique LZW consiste à repérer et à encoder les groupes de caractères répétitifs, elle est beaucoup plus efficace lorsqu'on l'applique aux longues chaîne de données, où la probabilité d'occurrence de sous-chaînes répétitives est plus élevée. L'application de la technique LZW à un seul enregistrement de la LEST à la fois ne nous permettait d'obtenir qu'un taux de compression de 45%. Bien qu'il ait été possible d'améliorer ce taux à l'aide d'autres opérations, comme la conversion de toutes les données

numériques en caractères codés binaires (ce qui constitue en soi une forme de compression), la technique LZW présentait des inconvénients trop graves.

Nous avons plutôt fait porter notre choix sur un algorithme de compression simpliste mais très efficace et facile à mettre en oeuvre, l'encodage binaire de longueur variable de Huffman. Les codes de Huffman permettent d'obtenir une configuration binaire distincte pour chacun des caractères à comprimer, les caractères dont l'occurrence est la plus fréquente étant ceux dont l'encodage nécessite le plus petit nombre de bits. De façon typique, les caractères qui apparaissent souvent dans un fichier (ou un enregistrement) peuvent être codés à l'aide de seulement 1 ou 2 bits (comparativement au nombre nominal de 8 bits que comporte chaque octet, ou caractère, selon le jeu de caractère ASCII). Il est possible que l'encodage des caractères dont l'occurrence est moins fréquente nécessite plus de 8 bits, mais cet inconvénient est plus que compensé par l'encodage des caractères à occurrence fréquente. Contrairement à la technique LZW et à d'autres techniques, l'encodage de Huffman permet de comprimer et de décomprimer les données instantanément au fil de la lecture du train binaire: il n'existe aucun lien d'interdépendance d'un caractère au suivant. De plus, cette technique de compression ne nécessite pas de longues chaînes de données pour repérer les configurations répétitives: elle permet de traiter chaque octet séparément. Par ailleurs, la rédaction d'un programme de compression-décompression de Huffman nécessite qu'on ait accès aux données au niveau du bit. La façon la plus simple de procéder consiste à utiliser un langage d'assemblage, bien qu'il soit possible d'employer d'autres langages permettant de modifier la configuration binaire, comme le langage C.

Grâce à sa simplicité, la technique de Huffman nous a permis d'utiliser une technique de compression additionnelle, l'encodage de la durée d'exécution, que nous avons appliquée à un seul caractère, le caractère blanc. De par sa nature, l'univers de la LEST comporte de nombreux espaces, surtout dans les champs réservés au nom commercial et à l'adresse (nous avons en outre tiré parti de la compressibilité des espaces en nous assurant que tous les champs numériques étaient remplis de caractères blancs à gauche, plutôt que de zéros). Au fur et à mesure que les données initiales de la LEST sont lues par le programme de compression, nous repérons les espaces consécutifs au sein du train de données afin de les encoder en ajoutant 4 bits additionnels à la suite du code de Huffman pour un espace. Ces 4 bits représentent un décompte des espaces additionnels présents dans la chaîne. Ainsi, il est possible de représenter jusqu'à 16 espaces consécutifs (128 bits) à l'aide de 5 bits, soit un code de Huffman de 1 bit représentant un espace et un quartet de bits additionnels pouvant représenter jusqu'à 15 espaces consécutifs.

Au total, nous avons obtenu un taux de compression de la LEST de 65 %, ce qui nous a permis de faire descendre la capacité de stockage nécessaire de 2 gigaoctets à 800 mégaoctets.

On construit les codes de Huffman en utilisant un arbre de Huffman comme celui illustré en annexe (vous trouverez une explication détaillée du processus dans le livre *Data Compression* de Gilbert Held, publié par John Wiley & Sons, Ltd.).

Pour construire un arbre de Huffman, il faut d'abord prétraiter le fichier initial afin de produire une distribution de fréquence de tous les caractères qu'il contient. Cette distribution est ensuite ordonnée de façon à ce que les caractères dont la fréquence d'apparition est la plus élevée se trouvent au sommet de l'arbre, avec leur fréquence d'occurrence indiquée. Alors, en commençant au bas de l'arbre, on combine les branches des deux caractères apparaissant le moins fréquemment et on additionne leurs fréquences pour produire une nouvelle branche. On part ensuite de cette nouvelle branche pour poursuivre le processus de combinaison des fréquences les moins élevées jusqu'à ce que toutes les branches aient été fusionnées en une. On attribue alors à chaque noeud ou intersection de branches un bit 0 d'un côté et un bit 1 de l'autre. On détermine enfin le code de Huffman (ou chaîne binaire) de chaque caractère, en commençant par la branche de poids faible, en accumulant les 0 ou 1 figurant à chaque noeud. Les configurations binaires obtenues pour chaque caractère peuvent ensuite être introduites dans une table à consulter que comporte le logiciel de compression.

Après avoir comprimé les données, il nous restait à déterminer la meilleure stratégie d'accès en vue d'extraire les enregistrements de la base des données. Nous avions le choix entre plusieurs techniques, dont la plus répandue consiste à stocker les enregistrements de données (comprimées) dans un fichier, et les clés d'accès connexes dans un autre fichier. Ainsi, lorsque le fichier initial est comprimé et chargé sur disque optique, un fichier distinct renfermant les enregistrements d'index est établi. Ce fichier d'index comprend des

enregistrements distincts pour chaque numéro d'identification de la base de données et indique la position de l'enregistrement de données correspondant dans la base de données comprimées (décalage en octets). Pour extraire un enregistrement, il faut consulter le numéro d'identification dans le fichier d'index, déplacer le pointeur de lecture-écriture dans le fichier principal jusqu'à obtention du décalage approprié, puis lire et décomprimer l'enregistrement. Pour solliciter plusieurs enregistrements à l'aide de la même clé, comme État/comté/code CTI, il faut disposer d'un autre fichier d'index renfermant un enregistrement pour chaque État/comté/code CTI et indiquant le décalage en octets dans le fichier principal du premier enregistrement sollicité avec cette clé. On peut alors lire et décomprimer des blocs d'enregistrements possédant la même clé d'accès. L'inconvénient est qu'il peut s'avérer assez inefficace de solliciter des blocs d'enregistrements avec la même clé, à moins que l'ordre de classement du fichier ne corresponde à celui de cette clé. Autrement, les enregistrements compris dans le groupe à extraire n'occupent pas des positions adjacentes dans le fichier et il faut solliciter le disque à de nombreuses reprises pour grouper tous les enregistrements. À l'évidence, le fichier ne peut être classé de plusieurs façons et les méthodes d'accès les plus rapides et les plus efficaces sont l'extraction d'un seul enregistrement à l'aide de son numéro d'identification et l'extraction d'enregistrements multiples à l'aide de la clé État/Comté/code CTI.

Bien que cette technique ait pour avantage de permettre l'utilisation de virtuellement n'importe quel langage de programmation, il est possible, selon le langage utilisé, d'avoir recours à d'autres structures de fichiers susceptibles de simplifier l'élaboration des méthodes d'accès et des fichiers d'accès connexes. Ainsi, certains compilateurs COBOL permettent d'avoir accès au fichier selon la méthode VSAM, qui est un rejeton de la méthode ISAM (méthode d'accès séquentiel indexé) ayant l'avantage de permettre l'utilisation d'enregistrements de longueur variable (on se rapellera qu'il est essentiel d'utiliser des enregistrements de longueur variable aux fins de la compression des données). La méthode VSAM nous permet de créer une base de données comprimées comportant de nombreuses clés d'accès ou combinaisons de clés distinctes sans qu'il soit nécessaire de créer, de maintenir et de solliciter des fichiers d'index distincts.

Bien que la première stratégie d'accès que nous ayons adoptée (ancienne méthode non VSAM) se soit révélée très efficace, nous sommes en train de réécrire le logiciel en utilisant une version du compilateur COBOL qui permet de recourir à la méthode VSAM. L'objectif premier de cet exercice n'est pas d'améliorer le rendement, mais d'augmenter le nombre de clés d'accès utilisées pour extraire les enregistrements (c.-à-d. d'élargir les "vues" potentielles de la base de données) sans utiliser un système trop complexe de fichiers d'index et de codage spécial.

Malgré sa taille importante, même selon les critères s'appliquant aux unités centrales, le système de gestion interactive de bases de données offre un aussi bon rendement que les autres systèmes interactifs que nous connaissons. L'accès instantané, la souplesse, les faibles coûts de revient et la facilité d'utilisation sont autant d'arguments qui militent en faveur de l'utilisation des réseaux de micro-ordinateurs comme solution de rechange valable aux options plus traditionnelles comme les ordinateurs centraux de grande puissance ou les mini-systèmes d'exploitation.

5. RÉSUMÉ

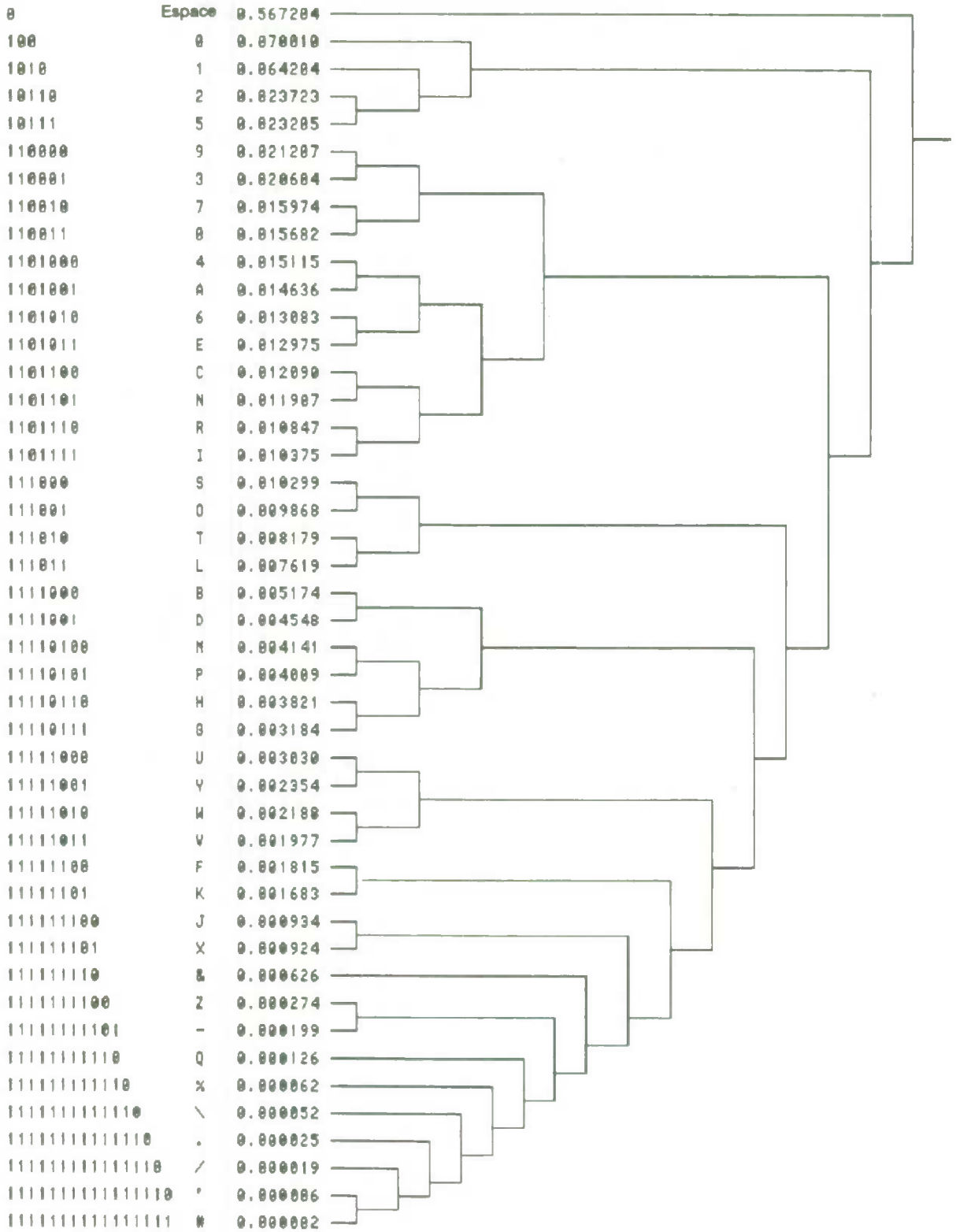
Le Census Bureau a entrepris un examen des opérations exécutées par les analystes et le personnel de bureau afin de déterminer les processus susceptibles d'être automatisés. L'élaboration de processus informatisés visant à remplacer les opérations manuelles peut nous permettre de réaliser des gains considérables de productivité.

Jusqu'à maintenant, nous avons entrepris l'automatisation de trois opérations: l'attribution des codes de classification des industries à partir des descriptions écrites des entreprises ou des noms commerciaux, la correction des valeurs des données dans les enregistrements rejetés au contrôle, et l'élaboration d'un réseau de micro-ordinateurs d'avant-garde. Nous comptons continuer à évaluer et à perfectionner ces trois processus automatisés tout en cherchant à automatiser d'autres opérations manuelles. L'automatisation de ces processus nous permettra d'améliorer la qualité de la LEST en respectant les contraintes budgétaires auxquelles nous sommes assujettis.

BIBLIOGRAPHIE

Held, G. (1988). *Data Compression* (2nd ed.), New York: John Wiley & Sons, Ltd.

Arbre de Huffman



SESSION 8

Amélioration du traitement des données et de l'estimation

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

UNE REVUE DE CERTAINES MÉTHODES DE MACRO-VÉRIFICATION VISANT À RATIONALISER LE PROCESSUS DE VÉRIFICATION

L. Granquist[1]

RÉSUMÉ

Cette communication présente des descriptions, des études, des résultats et des conclusions (y compris des recommandations) se rapportant à: la méthode des agrégats, la méthode de Hidiroglou-Berthelot (contrôle statistique), la méthode descendante, la méthode du tracé en boîte et la méthode graphique de la boîte. En faisant des simulations à l'aide de données d'enquête dans des conditions de production, on a trouvé que ces méthodes peuvent réduire le travail de vérification manuelle des données suspectes dans une proportion de 35 à 80% par rapport aux méthodes de micro-vérification classiques sans que la qualité en souffre. Ces méthodes sont faciles à comprendre et peuvent être intégrées sans difficulté à des systèmes de détection d'erreurs assistés par ordinateur par une simple addition de programmes. Les caractéristiques de chaque méthode sont comparées et la principale constatation est que les méthodes sont très comparables et qu'elles visent toutes à déterminer des bornes efficaces pour les micro-vérifications d'utilisation courante. Ces bornes devraient être fondées uniquement sur des statistiques établies d'après les données d'entrée pondérées.

MOTS CLÉS: Simulation; qualité des données; système interactif; micro-vérification.

1. OBJECTIF

Cette communication a pour objectif de présenter des méthodes de macro-vérification comme moyen de résoudre le problème de la survérification. Des études expérimentales portant sur le traitement des données et la production statistique ont révélé que les méthodes de ce genre sont supérieures aux méthodes de micro-vérification pour le contrôle des données quantitatives. Selon ces mêmes études, le travail de vérification manuelle serait réduit de 35 à 80%.

Pour faciliter la compréhension et mettre en évidence les caractéristiques essentielles de ces méthodes de macro-vérification, nous décrivons chacune des méthodes de façon schématique. Nous insistons sur les aspects rationnels de la macro-vérification en comparaison de la micro-vérification.

On trouvera la description détaillée des méthodes et des études qui sont à la base de cette communication dans les sources citées ainsi que dans la bibliographie.

Enfin, nous concluons cette communication par une brève analyse comparative des méthodes de macro-vérification et de micro-vérification.

---

[1] L. Granquist, Statistical Methods, Research & Development Department, Statistics Sweden, S-115 81 Stockholm, Sweden.

2. INTRODUCTION

Granquist (1984) définit la macro-vérification comme un moyen de déceler les erreurs de compréhension. (Ce type d'erreur comprend les erreurs dues à l'ignorance ou à une mauvaise interprétation des questions, des concepts ou des définitions, ainsi que les erreurs de tactique).

À Statistics Sweden, nous avons tout d'abord expérimenté les méthodes de macro-vérification sur des agrégats. Lorsqu'on découvrait un agrégat suspect, on cherchait premièrement à savoir si cette condition était attribuable à une ou à deux observations en particulier. Lorsque les agrégats étaient des moyennes, on s'est rendu compte que les méthodes de macro-vérification étaient plus efficaces que les méthodes de micro-vérification classiques pour détecter les données suspectes. Nous avons alors décidé d'approfondir cet aspect de la macro-vérification jusqu'à ce que nous parvenions à élargir l'usage de ces méthodes à Statistics Sweden. Nous sommes sur le point de conclure la première partie du projet et prévoyons amorcer la seconde partie l'année prochaine.

D'après les expériences que nous avons réalisées, nous avons jugé pertinent de définir la macro-vérification (ce type de méthodes de macro-vérification) comme une série de "contrôles statistiques" appliqués à des données pondérées. Les méthodes décrites dans cette communication peuvent être vues comme des méthodes visant à rendre plus efficaces les limites d'acceptation des méthodes de micro-vérification. Elles redéfinissent le travail de vérification en fonction de priorités.

3. LE PROBLÈME DE LA SURVÉRIFICATION

Nous pouvons formuler dans les termes suivants le problème fondamental des méthodes de micro-vérification classiques: à cause d'un trop grand nombre de contrôles effectués avec des limites trop étroites, de nombreux systèmes de détection d'erreurs produisent un trop grand nombre de messages d'erreur qu'il faut vérifier manuellement. Les commis préposés à cette opération ne sont pas en mesure d'évaluer l'importance de l'erreur signalée. Chacun de ces signaux a le même poids et exige la même quantité de ressources, mais beaucoup d'erreurs ont un effet négligeable sur les estimations parce qu'elles sont faibles ou qu'elles s'annulent. En règle générale, les limites d'acceptation sont fixées subjectivement suivant le principe de la "prudence", c'est-à-dire que l'on accepte uniquement les données dont la fiabilité ne fait aucun doute. Par exemple, une forme de contrôle courante dans les enquêtes-entreprises à Statistics Sweden est de signaler tout élément d'information qui a subi une variation de plus de 10% depuis la dernière enquête. De telles méthodes de micro-vérification se traduisent généralement par une forte *survérification*.

4. "DÉFINITIONS"

Cette communication porte sur les contrôles appliqués à des données quantitatives et qui ont pour but de repérer les données suspectes et de les soumettre à une vérification manuelle. Ce genre de contrôle peut être considéré comme l'inverse du contrôle de validation, qui sert à signaler les données erronées. Ferguson (1989) appelle le premier type "contrôle statistique". Ce type de contrôle utilise les distributions de données courantes établies à partir de tous les questionnaires ou d'une partie d'entre eux ou les données chronologiques de l'unité statistique afin de produire des limites acceptables pour les données d'enquête courantes.

Dans cette communication, le terme "macro-vérification" désigne une méthode par laquelle on décèle les données suspectes en appliquant des contrôles statistiques fondés sur les données d'entrée pondérées. Les bornes inférieure et supérieure d'une macro-vérification (macro-contrôle) doivent être établies *uniquement en fonction des données à contrôler* et *de l'importance relative de ces données dans l'ensemble.*

5. LA TECHNIQUE D'ÉVALUATION

Les méthodes ont été évaluées au moyen d'études de simulation appliquées à des données d'enquête. Les simulations ont été exécutées à l'aide de prototypes d'un système de vérification complet pour enquêtes. On a ensuite comparé les résultats de la méthode étudiée avec ceux des méthodes de micro-vérification appliquées

au cours du traitement des données de l'enquête. Les modifications rendues nécessaires par suite de la vérification des données de l'enquête ont été inscrites dans un fichier et notre étude consistait à déterminer (par quelques calculs) quelles données de ce fichier étaient signalées par la méthode de macro-vérification et lesquelles ne l'étaient pas. On mesurait l'effet de rationalisation par la réduction du nombre d'indicateurs et la "perte de qualité", par l'incidence des erreurs restantes (c'est-à-dire, les erreurs détectées au cours du traitement mais non signalées par la méthode de macro-vérification).

6. MÉTHODES DE MACRO-VÉRIFICATION

Nous décrivons ci-dessous *"la méthode des agrégats"*, *"la méthode de Hidiroglou-Berthelot"* (*"contrôle statistique"*) et *"la méthode descendante"* et donnons quelques résultats de l'application de ces méthodes. *"La méthode du tracé en boîte"* et *"la méthode de la boîte"* sont aussi abordées, mais comme variantes de la méthode des agrégats et de la méthode descendante respectivement. La méthode de la boîte est encore au stade de l'élaboration.

6.1 Méthode des agrégats

6.1.1 Documentation

La version originale de la méthode des agrégats est décrite en détail dans Granquist (1988b). Elle avait été élaborée en SAS comme prototype d'un système de vérification complet aux fins d'une enquête sur l'emploi et la rémunération (Survey on Employment and Wages -- SEW), lequel prototype devait être utilisé sur un gros ordinateur. Une version modifiée de cette méthode est décrite en détail dans Lindström (1990). Elle avait été élaborée en PC-SAS comme prototype pour la SEW, conçu pour les ordinateurs personnels.

6.1.2 Description de la méthode

Essentiellement, la méthode des agrégats consiste à appliquer des contrôles aux agrégats d'abord, puis à chaque enregistrement des agrégats suspects (signalés). Tous les enregistrements d'un agrégat suspect d'une variable quelconque constituent le fichier d'erreurs. C'est à ce fichier que sont appliqués les contrôles prévus pour les enregistrements.

Ce qui distingue plus particulièrement la méthode des agrégats est le mode d'établissement des limites d'acceptation. On détermine manuellement ces limites en examinant des listes d'observations ordonnées des fonctions de contrôle. Seulement les "s" plus grandes observations et les "m" plus petites observations des fonctions de contrôle figurent sur les listes.

La fonction de contrôle A (pour les agrégats) aussi bien que la fonction de contrôle F (pour les enregistrements) doivent être des fonctions de la valeur pondérée (selon le plan de sondage) des données d'entrée de la variable qui faisant l'objet du contrôle. En utilisant des valeurs pondérées dans la fonction A, on peut calculer les valeurs de contrôle pour les agrégats de la même manière qu'on le fait dans la micro-vérification classique. La macro-vérification peut être exécutée aussi facilement que la micro-vérification.

On peut se servir des listes d'observations de façon directe ou indirecte. Dans le premier cas, elles servent à un examen manuel des observations (lorsque celles-ci sont accompagnées d'identificateurs) tandis que dans le second cas, elles servent à fixer des limites d'acceptation pour un programme de détection d'erreurs, qui signale les données suspectes par des messages d'erreur en vue d'un examen manuel. L'utilisation d'un programme de détection d'erreurs a l'avantage de préserver la forme du processus initial. L'application de la méthode des agrégats n'exige que des programmes permettant l'impression des listes d'observations ordonnées. On peut ajouter facilement de tels programmes à l'ancien système.

Ce serait une amélioration notable si la méthode pouvait produire la liste des observations qui se trouvent aux extrémités des distributions, accompagnée de graphiques ou de paramètres statistiques comme la médiane, les quartiles, l'étendue l'écart interquartile et les graphiques. Des tracés en boîte (voir Tukey, 1975) seraient appropriés dans ce cas.

6.1.3 Applications

Les contrôles utilisés dans nos études sur la méthode des agrégats comprenaient un contrôle de rapports et un contrôle de différences. Il fallait que l'un et l'autre échouent pour qu'une valeur soit considérée suspecte. Plusieurs études de simulation ont été exécutées avec des données de la SEW pour différents mois. La principale constatation est la suivante:

La méthode des agrégats peut constituer l'élément fondamental du *processus de vérification* à l'*étape du traitement des données* d'une enquête. Elle *réduit* le travail de vérification dans des proportions de 35 à 80% sans que la *qualité* ou les *délais de production* en souffrent. La macro-vérification est une solution de rechange ou un complément réalistes pour les méthodes de micro-vérification et peut être appliquée durant le traitement des données dans les mêmes conditions que la micro-vérification assistée par ordinateur; elle réduit considérablement le travail de vérification manuelle.

Les études de simulation nous amènent à faire d'autres constatations:

Des erreurs graves peuvent échapper à la micro-vérification.

Les intervalles d'acceptation des contrôles devraient être beaucoup plus larges qu'ils ne le sont actuellement et ne devraient pas être centrés sur 1, pour les rapports, ni sur 0, pour les différences.

Une bonne façon de procéder est de fixer les limites d'acceptation aussi près que possible de la première valeur aberrante de part et d'autre.

Dans une étude réalisée avec le prototype PC-SAS (voir Lindström, 1990), une constatation très importante nous a obligés à élargir la "définition" de ce genre de méthodes de macro-vérification:

L'étude en question avait permis de constater une diminution de près de 80% du nombre de données signalées. Cependant, la perte de qualité était légèrement plus forte que dans les autres études. On a découvert que cette perte de qualité était attribuable à *un petit nombre* d'erreurs *importantes*, qui faisait que les agrégats n'étaient pas signalés. Ces erreurs pouvaient être détectées très facilement. Une solution intéressante dans les circonstances était de soumettre toutes les données à un contrôle de rapports avec des intervalles d'acceptation très larges. Cette opération pouvait se faire au stade de la micro-vérification ou tenir lieu de contrôle final.

On s'est ensuite interrogé sur l'utilité des contrôles d'agrégats. Ces contrôles ont pour seul avantage de permettre une économie de mémoire ou de temps machine. Or, lorsque la capacité de mémoire ou les coûts de traitement ne posent pas problème, les contrôles d'agrégats sont superflus.

Cette constatation nous a obligés à élargir la définition de la macro-vérification en adoptant la définition donnée plus haut. Il s'agit donc toujours d'une méthode de macro-vérification mais elle ne devrait pas être appelée méthode des agrégats. Lorsqu'on peut connaître les extrémités de la distribution de la fonction de contrôle grâce à des tracés en boîte (ce qui est recommandé), on parle de méthode du tracé en boîte.

En conclusion, nous pouvons dire qu'il n'est pas souhaitable d'utiliser la méthode des agrégats pour de petites enquêtes; la méthode du tracé en boîte convient mieux dans ces conditions.

6.2 Méthode descendante

6.2.1 Documentation

La méthode descendante est décrite dans Granquist (1987) et Lindblom (1990). Elle a été utilisée sur un gros ordinateur pour les besoins de l'enquête sur les livraisons et les soldes des carnets de commandes (Survey of Delivery and Orderbook Situation -- DOS). Le programme de production est écrit en langage APL. Dans son rapport, Lindblom (1990) élabore un prototype en PC-SAS conçu pour les micro-ordinateurs.

6.2.2 Description de la méthode

Essentiellement, la méthode descendante consiste à trier les valeurs des fonctions de contrôle (qui sont des fonctions des données d'entrée pondérées), puis à faire l'examen manuel des observations à partir du haut ou du bas de la liste jusqu'à ce que ce processus n'ait plus d'effet notable sur les estimations.

Cette méthode est décrite ici du point de vue de son application dans le programme de production de l'enquête DOS. Cependant, la généralisation est évidente.

La procédure est déterminée par un programme interactif de menus qui est rédigé en langage APL. On constitue progressivement un fichier de passage à l'aide des enregistrements tirés des trois lots formés à l'étape de la saisie des données. Les enregistrements à étudier sont choisis suivant trois fonctions, à savoir

i) les 15 plus fortes variations positives
ii) les 15 plus fortes variations négatives
iii) les 15 plus grandes contributions.

Ces fonctions peuvent être appliquées au total ainsi qu'aux 38 grands domaines d'études pour chaque variable. Pour une fonction et un domaine d'études donnés, l'opérateur peut voir à l'écran les rubriques suivantes pour les 15 enregistrements du fichier de passage triés selon un ordre descendant:

IDENTITÉ DONNÉES POIDS VALEUR PONDÉRÉE TOTAL

L'opérateur sélectionne un enregistrement et aussitôt, le contenu de cet enregistrement apparaît à l'écran. Si une erreur est détectée, l'opérateur peut mettre à jour l'enregistrement à l'écran et constater aussitôt les effets de la correction. L'enregistrement disparaît ensuite de la liste des 15 premiers et le total est modifié. Cette opération se poursuit jusqu'à ce que les mises à jour n'aient plus d'effet sur le total.

6.2.3 Applications

La méthode descendante avait été conçue à l'origine comme le complément ou le substitut d'une méthode de micro-vérification classique. Les deux méthodes ont été élaborées en même temps mais la seconde, qui devait servir de méthode de base pour l'enquête DOS, a tout de suite fait place à la première.

La méthode descendante peut servir à de la vérification durant le traitement des données d'une enquête sans que la qualité ou les délais de production en souffrent. Elle peut aussi réduire le travail de vérification dans des proportions de 50 à 75% et elle renseigne constamment les préposés sur le sujet et les problèmes de l'enquête. Les commis en tirent beaucoup de satisfaction parce qu'ils ont le plein contrôle des opérations et qu'ils peuvent voir les effets de leur travail.

Depuis le moment où on a utilisé pour la première fois les techniques de macro-vérification dans la production de données, le nombre d'enregistrements à réviser manuellement diminue petit à petit. Les statisticiens en charge de l'enquête DOS croient fermement que *la vérification devrait se faire uniquement au niveau du "secteur de la fabrication"*.

Bien que la méthode descendante semble donner lieu à une certaine sur-vérification, elle est, sans conteste, la méthode de vérification la plus logique qui soit à Statistique Suède.

Selon Anderson (1989a), cette méthode sert aussi au "contrôle des résultats" et est considérée, à ce titre, comme la plus efficace des méthodes en usage au Australian Bureau of Statistics.

6.3 Méthode de Hidiroglou-Berthelot (contrôle statistique)

6.3.1 Documentation

La méthode de Hidiroglou-Berthelot (méthode HB) est décrite comme une méthode de micro-vérification dans Hidiroglou-Berthelot (1986). Inspirée des méthodes d'analyse préliminaire de données de Tukey (Exploratory Data Analysis -- EDA) (voir Tukey, 1977), elle consiste en un contrôle de rapports et ses concepteurs la présentent dans leur article comme un moyen de résoudre certains problèmes liés à la méthode classique du contrôle de rapports. En usage à Statistique Canada, elle est connue sous l'appellation "contrôle statistique".

Höglund (1989) en parle comme d'une méthode de macro-vérification. À Statistics Sweden, on l'a étudiée dans le cadre de diverses enquêtes. Seule l'étude relative à l'enquête sur les livraisons et les soldes des carnets de commandes (DOS) existe en version anglaise (Höglund, 1989).

6.3.2 Description de la méthode

La méthode HB est conçue pour le contrôle de rapports. Les bornes sont calculées automatiquement à l'aide des données qui doivent être contrôlées. La méthode utilise des paramètres robustes comme la médiane, les quartiles et l'écart interquartile car les bornes ne sauraient être influencées par des valeurs aberrantes. La méthode utilisée ordinairement pour le contrôle de rapports présente deux inconvénients: -- *elle ne permet pas toujours de déceler des observations aberrantes à l'extrémité gauche de la distribution* et -- *elle ne tient pas compte de ce que la variabilité des rapports pour de petites valeurs est plus grande que la variabilité des rapports pour de grandes valeurs*. La méthode HB élimine ces inconvénients par une transformation symétrique suivie d'une transformation de dimension.

$$S_i = \begin{cases} 1 - R_{\text{MEDIAN}} / R_i \ , \ O < R_i < R_{\text{MEDIAN}} \\ R_i / R_{\text{MEDIAN}} - 1 \ , \ R_i \geq R_{\text{MEDIAN}} \end{cases}$$

$$E_i = S_i * (\text{MAX } (X_i(t) \ , \ X_i(t+1)))^U, \ 0 \leq U \leq 1$$

$X_i(t)$ et $X_i(t+1)$ représentent les valeurs pondérées de l'élément pour les périodes t et $t+1$ respectivement. E_{Q1}, E_{Q3} désignent les premier et troisième quartiles de la transformation E.

$$D_{Q1} = \text{MAX } (E_{\text{MEDIAN}} - E_{Q1} \ , \ | \ A * E_{\text{MEDIAN}} \ | \ )$$

$$D_{Q3} = \text{MAX } (E_{Q3} - E_{\text{MEDIAN}} \ , \ | \ A * E_{\text{MEDIAN}} \ | \ ),$$

ce qui donne les bornes inférieure et supérieure des contrôles:

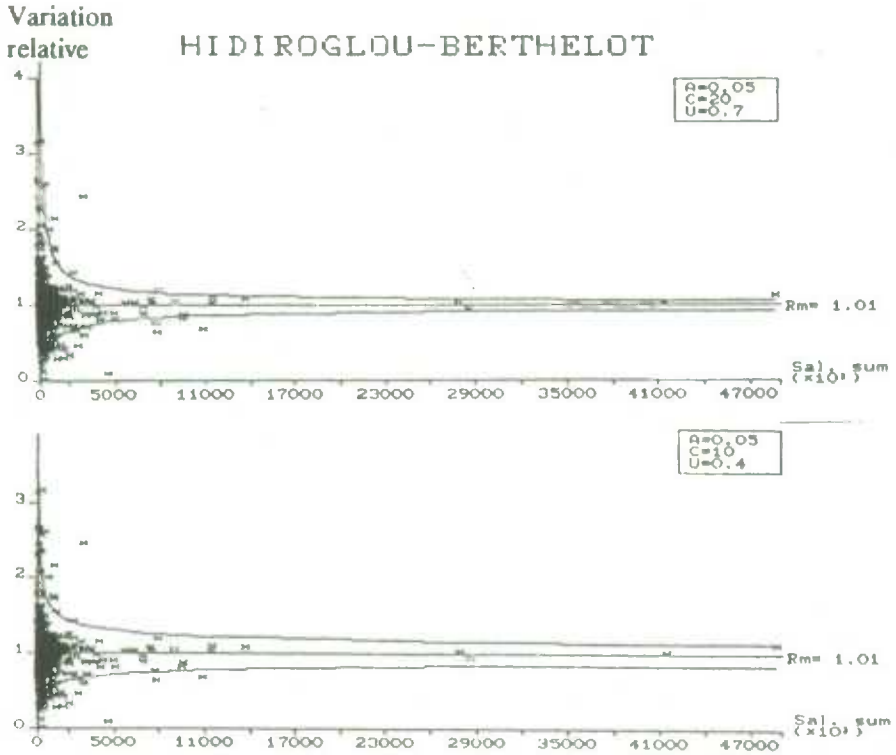
$$1 = E_{\text{MEDIAN}} - C * D_{Q1}$$

$$u = E_{\text{MEDIAN}} + C * D_{Q3}$$

"A" est tenu pour constant et sa valeur égale 0.05; il ne reste plus donc que deux paramètres, U et C, qu'il faut déterminer au préalable si l'on veut rendre la méthode opérationnelle.

La figure 1 peut faire mieux comprendre les transformations et la méthode. Cette figure est le résultat de l'application du prototype de la méthode de la boîte à des données de l'enquête sur l'emploi et la rémunération. Les limites d'acceptation ont été calculées pour deux couples de valeurs des paramètres *U* et *C*. On peut appliquer cette méthode lorsque la méthode HB s'inscrit dans les opérations du processus de production. Dans ces circonstances, les limites d'acceptation sont déterminées entièrement à l'aide des données qui doivent être contrôlées.

Figure 1



6.3.3 Applications

Il s'agit là d'une excellente méthode qui peut s'avérer un bon substitut pour la méthode descendante. Une comparaison faite avec des données de l'enquête sur l'emploi et la rémunération a permis de constater que la méthode HB était supérieure à la méthode des agrégats. Nous avons de plus observé que les paramètres *U* et *C* ne sont pas très sensibles. On peut se servir des mêmes valeurs pour un grand nombre de variables d'une enquête, ce qui rend facile l'utilisation de cette méthode.

6.4 Méthode du tracé en boîte

6.4.1 Introduction

C'est à Tukey (1977) que l'on doit la notion de tracé en boîte. Anderson (1989b) parle de la méthode du tracé en boîte comme d'une méthode de micro-vérification.

Anderson (1989a) relate une expérience réalisée avec des données de l'enquête du Australian Bureau of Statistics (ABS) sur les gains hebdomadaires moyens (Average Weekly Earnings -- AWE). Le ABS a porté les bornes de chaque contrôle de rapports utilisé dans cette enquête à ($Q1-3*IQR$, $Q3+3*IQR$), où $Q1$, $Q3$ et IQR sont, respectivement, les premier et troisième quartiles et l'écart interquartile. L'étude a révélé qu'on avait pu épargner 75% des ressources normalement affectées à l'examen manuel des données signalées dans l'ensemble du processus de vérification en limitant cet examen aux "valeurs aberrantes extrêmes". Anderson s'est servi de

la technique d'évaluation que nous avons utilisée pour nos études à Statistics Sweden. Les erreurs restantes n'avaient aucune incidence sur les estimations.

Dans la seconde version du rapport, Anderson propose de fixer les bornes inférieure et supérieure des contrôles d'après une analyse manuelle des tracés en boîte des fonctions de contrôle. On peut ainsi modifier les bornes en tenant compte des valeurs aberrantes qui sont voisines des bornes établies pour les valeurs aberrantes extrêmes. Par-dessus tout, les préposés de l'enquête auraient le plein contrôle sur les opérations de vérification.

6.4.2 Description de la méthode

Les distributions des fonctions de contrôle des valeurs pondérées (selon le plan de sondage) sont représentées sous forme de tracé en boîte. Les préposés se fondent sur ces tracés pour définir les intervalles d'acceptation des contrôles; ces intervalles sont ensuite introduits dans le programme de détection d'erreurs. D'après Anderson (1989a) et les études sur la méthode des agrégats mentionnées dans cette communication, la définition des valeurs aberrantes extrêmes peut servir de repère pour l'établissement de limites efficaces.

6.5 Méthode de la boîte

6.5.1 Introduction

La méthode de la boîte est une méthode graphique de macro-vérification qui est en voie d'élaboration à Statistics Sweden. La première version d'un prototype pour l'enquête sur l'emploi et la rémunération doit sortir en janvier 1991. Essentiellement, cette méthode consiste à utiliser l'infographie pour visualiser la distribution de la fonction de contrôle des données pondérées et à exploiter l'interactivité de l'ordinateur pour savoir quand interrompre le travail de vérification manuelle. Elle peut être vue comme la combinaison d'une méthode du tracé en boîte généralisée et de la méthode descendante.

6.5.2 Description de la méthode

Les valeurs des éléments sont pondérées, puis intégrées à la fonction de contrôle. N'importe quelle expression mathématique peut servir de fonction de contrôle. Les valeurs de la fonction sont représentées graphiquement à l'écran et on peut aussi avoir des régions d'acceptation de forme variée. L'opérateur dessine une boîte autour des observations qu'il veut examiner. Il voit alors apparaître à l'écran les données relatives à des éléments présélectionnés des enregistrements qui se rattachent aux observations encadrées. Pour chaque fonction de contrôle, l'utilisateur peut choisir les éléments d'enregistrements qu'il veut faire afficher. Il introduit une modification de façon interactive et aussitôt, il peut voir l'effet de cette modification à l'aide de données (statistiques).

La méthode de la boîte peut aussi servir à déterminer des régions d'acceptation appropriées pour d'autres méthodes de vérification (par ex.: méthode HB).

7. RÉSUMÉ

7.1 Description des méthodes

La caractéristique fondamentale de toutes les méthodes de macro-vérification que nous venons d'exposer est que les régions d'acceptation sont déterminées uniquement à l'aide des distributions des observations des fonctions de contrôle. Les valeurs d'entrée de la variable à contrôler sont pondérées au moyen d'un facteur de gonflement avant d'être introduites dans la fonction de contrôle.

En ce qui concerne la méthode des agrégats, la méthode du tracé en boîte, la méthode HB et la méthode descendante, toutes les valeurs de la fonction de contrôle sont triées selon un ordre de grandeur.

Dans la méthode des agrégats et la méthode du tracé en boîte, les queues des distributions sont affichées et analysées afin d'établir les limites d'acceptation des contrôles. Dans la méthode HB, cette opération se fait automatiquement.

Dans la méthode descendante et la méthode de la boîte, l'étendue de l'examen manuel est "déterminée" par l'incidence des erreurs détectées sur les estimations. Dans la méthode descendante, l'examen manuel débute avec les valeurs extrêmes et se poursuit vers la valeur médiane, tandis que dans la méthode de la boîte, l'opérateur choisit les enregistrements qui doivent faire l'objet d'un examen manuel à partir d'un graphique des valeurs des fonctions de contrôle. Cette sélection peut être facilitée par des régions d'acceptation figurant sur le même graphique.

Le choix de la méthode dépendra du nombre de variables à contrôler par la macro-vérification et du mode de travail que voudront adopter les préposés.

7.2 Macro-vérification VS micro-vérification

La macro-vérification n'est pas un concept nouveau. Elle est utilisée depuis toujours pour le contrôle des données, mais seulement comme une forme de contrôle final. On dit aussi "contrôle des résultats". Ce qu'il y a de nouveau, c'est que les méthodes de contrôle des résultats peuvent être utilisées de la même manière que les méthodes de micro-vérification pour le contrôle des données et qu'elles s'avèrent beaucoup plus efficaces que les méthodes de micro-vérification classiques.

Les méthodes de macro-vérification que nous venons de décrire peuvent être vues comme un moyen statistique de faire de la micro-vérification avec des limites d'acceptation efficaces. Ces limites sont établies uniquement en fonction des données à contrôler. Les méthodes de macro-vérification redéfinissent le travail de vérification en fonction de priorités, c'est-à-dire que les données sont contrôlées suivant leur incidence sur les estimations. Ces méthodes éliminent le problème fondamental des méthodes de micro-vérification, à savoir que celles-ci produisent un trop grand nombre de signaux d'erreur sans donner d'indications sur la manière de répartir les ressources affectées à la vérification. Même les erreurs très importantes échappent parfois aux méthodes de micro-vérification à cause du grand nombre de données signalées.

Toutes les études qui ont été publiées sur l'incidence de la vérification montrent que seul un petit nombre des erreurs détectées influent sur les estimations. Par exemple, Greenberg et coll. (1982) constatent que pour tous les éléments étudiés et tous les domaines d'études, environ 5% des cas expliquaient plus de 90% de la variation totale. Bon nombre des variations les plus notables étaient dues au fait que des valeurs avaient été exprimées en unités plutôt qu'en milliers. On trouve des résultats semblables dans une étude concernant l'enquête de 1987 sur les comptes financiers en Suède et dans les études dont font état Linacre et Trewin et qui traitent du processus de vérification utilisé dans trois enquêtes du Australian Bureau of Statistics.

Par conséquent, dans les cas où seulement quelques-unes des erreurs influent sur les estimations, les méthodes de macro-vérification permettent de réduire considérablement les ressources affectées à l'examen des données. On parle d'une réduction du volume de travail de l'ordre de 35 à 80%.

Cependant, il n'existe aucune limite quant au nombre de cas qui peuvent être choisis pour un examen manuel. Le préposé peut sélectionner tous les cas qu'il juge nécessaires. À la différence des méthodes de micro-vérification, les méthodes de macro-vérification prévoient la sélection des cas selon un ordre de priorité, c'est-à-dire suivant l'effet que ces cas pourraient avoir sur les estimations. Cette sélection est faite essentiellement par le préposé, ce qui signifie que celui-ci est entièrement maître du processus d'examen. En micro-vérification, le préposé doit tout laisser faire par l'ordinateur et il ne peut voir les résultats de son travail.

Les deux types de méthodes ont plus particulièrement pour objet les erreurs aléatoires dues à la négligence et elles signalent toutes deux les valeurs aberrantes ou les observations extrêmes suivant le même principe. Les méthodes de micro-vérification signalent les données selon des critères préétablis, fondés sur des données chronologiques, tandis que les méthodes de macro-vérification signalent des données qui, au moment même de l'opération, représentent des valeurs extrêmes par rapport aux estimations.

Ni l'une ni l'autre des méthodes ne permet (en théorie) de détecter les erreurs systématiques, c'est-à-dire les erreurs qui se produisent lorsqu'un bloc de répondants comprennent mal une question ou donnent délibérément une mauvaise réponse.

Il convient de souligner que ni les méthodes classiques de micro-vérification ni aucune des méthodes décrites dans cette communication n'amélioreront vraiment la qualité des estimations d'enquête. Il est vrai que ces méthodes contribuent à éliminer certains types d'erreurs mais il n'est pas sûr que les estimations soient de meilleure qualité pour autant. L'étude qui a été faite sur la vérification automatique des données de l'Enquête mondiale sur la fécondité a révélé que l'opération n'avait pas amélioré les estimations mais avait plutôt retardé d'un an la publication des résultats (voir Pullum et coll., 1986 ou Granquist, 1988a).

Nous donnons ci-dessous quelques ouvrages traitant des méthodes qui ont principalement pour objet les erreurs de compréhension. Ces ouvrages montrent que même des données contrôlées soigneusement à l'aide d'une méthode de micro-vérification peuvent renfermer des erreurs qui influent considérablement sur les estimations. Ces erreurs ont même une incidence plus forte que les erreurs "restantes" des études de simulation qui ont porté sur les méthodes de macro-vérification analysées ici.

Werking et coll. (1988) parlent d'une enquête pour l'analyse de réponses (Response Analysis Survey): dans le cadre d'une enquête permanente, on réalise une enquête spéciale visant à vérifier de quelle manière les participants ont répondu à certaines questions. Cette enquête se fait au moyen d'un téléphone à clavier; le participant enregistre ses réponses par simple pression d'un bouton-poussoir. Cette technique permet de déceler les erreurs qui échappent aux méthodes de vérification classiques, ce qui réduit le biais dans les estimations.

Mazur (1990) présente une méthode qui permet de déceler les valeurs dites "immuables" parmi des données d'enquête. Dans des enquêtes à passages répétés, un certain nombre de participants donnent systématiquement la même réponse d'une fois à l'autre à des questions qui, normalement, appellent une réponse différente d'une période à l'autre. Ces valeurs sont dites "immuables" parce qu'elles se situent toujours entre les bornes des contrôles classiques. Dans les expériences qu'elle a réalisées, Mazur constate que ces "immuables" peuvent introduire un biais appréciable dans les estimations.

En conclusion, nous ne pouvons dire si l'un ou l'autre des types de méthode améliore vraiment la qualité des estimations mais ce dont nous sommes sûrs, c'est que la macro-vérification permet plus efficacement d'obtenir les mêmes normes de "qualité" et peut libérer des ressources aux fins de la détection des erreurs de compréhension.

BIBLIOGRAPHIE

Anderson, K. (1989a). *Draft, Output Edit Study, Average Weekly Earnings,* Statistical Services Branch, Australian Bureau of Statistics.

Anderson, K. (1989b). *Enhancing Clerical Cost-Effectiveness in the Average Weekly Earnings*, Draft, Australian Bureau of Statistics, Statistical Services Branch.

Ferguson, D.P. (1989). *Review of Methods and Software Used in Data Editing*, SCP2/DE/WP.33 (U.S. Department of Agriculture, National Agricultural Statistics Service).

Granquist, L. (1984a). *On the Role of Editing*, Statistisk Tidskrift 1984:2

Granquist, L. (1987). *Macroediting - The Top-Down Method*, Statistics Sweden, Report 1987-04-09.

Granquist, L. (1988a). *On the Need for Generalized Numeric and Imputation Systems*, The Seminar on Statistical Methodology, Geneva.

Granquist, L. (1988b). *Macroediting - The Aggregate Method*, Statistics Sweden, Report 1988-08-18.

Greenberg, B., et Petkunas, T. (1982). *An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in Business Division*, 1982 Economic Censuses and Census of Governments, Evaluation Studies.

Hidiroglou, M.A., et Berthelot, J.-M. (1986). Contrôle statistique et imputation dans les enquêtes-entreprises périodiques. *Techniques d'enquête*, 12, 1, 79-89.

Höglund D.E. (1989). *Macroediting - The Hidiroglou-Berthelot Method (Statistical Edits)*, Statistics Sweden, Report 1989-03-28.

Linacre, S.J., et Trewin, D.J. *Evaluation of Errors and Appropriate Resource Allocation in Economic Collections*, Undated, Internal paper, Australian Bureau of Statistics.

Lindblom, A. (1990). *A review of the macro-editing procedure Top-Down*, Data Editing Joint Group Product NR SCP2/D.12/f.

Lindström, K. (1990). *A macroediting method application developed in PC-SAS*, Data Editing Joint Group Product NR SCP2/D.11/f.

Mazur, C. (1990). *A Statistical Edit for Livestock Slaughter Data*, USDA-NASS, Washington DC 202 50.

Pierzchala, M. (1988). *A Review of the State of the Art in Automated Data Editing and Imputation*, Staff report for National Agricultural Statistics Service, U.S.D.A. No. SRB-88-10.

Pullum, T.W., Harpham, T., et Ozsever, N. (1986). *The Machine Editing of Large Sample Surveys: The Experience of the World Fertility Survey*, International Statistical Review, 54, 3.

Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley Publishing Company.

Werking G., Tupek A., et Clayton R. (1988). CATI and Touchtone Self-Response Applications for Establishment Surveys, *Journal of Official Statistics*, 1988-4.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

POSSIBILITÉS INTERACTIVES DU SPEER

(Programme structuré pour la vérification et l'étude de cas complexes dans les enquêtes économiques)

L. Draper, B. Greenberg et T. Petkunas[1]

RÉSUMÉ

Le système SPEER est un système polyvalent de contrôle et d'imputation élaboré par le personnel du Bureau of the Census pour traiter les données élémentaires continues faisant l'objet de contrôles de rapports. Le système a été utilisé pour contrôler "l'Auxiliary Establishment Report" et "l'Enterprise Summary Report" des recensements de l'économie de 1982 et de 1987, ainsi que les formules de déclaration du recensement des industries de la construction de 1987. En outre, on a utilisé une version modifiée du SPEER pour contrôler les questionnaires du recensement des manufactures de 1987 ainsi que ceux des enquêtes annuelles subséquentes sur les manufactures. Le présent article porte sur les possibilités interactives du système SPEER qui ont été utilisées pour l'examen par les analystes des cas de renvoi au contrôle ainsi que pour la saisie des données recueillies à l'aide de questionnaires envoyés par la poste.

MOTS CLÉS: Contrôle; imputation; SPEER.

1. INTRODUCTION

Toutes les données d'enquête doivent faire l'objet d'un contrôle visant à déceler les combinaisons de réponses incohérentes sur les questionnaires, à apporter des modifications aux données déclarées introduites au clavier et à imputer les éléments d'information manquants. Cette opération a donc pour double objectif de déceler et de corriger les erreurs causées par une mauvaise compréhension des questions, une déclaration erronée ou des problèmes relatifs la saisie des données. Au cours de ce processus de détection et de correction des erreurs, l'enregistrement de données est soumis à un ensemble de procédures automatisées, de procédures manuelles et de combinaisons interactives des deux procédures. Nous allons décrire dans cet article la séquence d'activités mises en oeuvre au Census Bureau dans le cadre des enquêtes de grande envergure ou recensements sur les entreprises.

Les données sont recueillies à l'aide d'un questionnaire imprimé que l'on poste aux membres de l'univers de l'enquête. Une fois le questionnaire rempli, les répondants le retournent à la succursale du Census Bureau de Jeffersonville (Indiana) où les données sont saisies. Au cours de cette opération, on effectue des vérifications rudimentaires afin de déceler les erreurs de frappe susceptibles de survenir. Une fois les données saisies, elles sont transmises au bureau central de Suitland (Maryland), où elles sont soumises à un programme automatisé de contrôle par lots visant à déceler les incohérences, à apporter les modifications nécessaires et à imputer les réponses manquantes.

Au cours de l'exécution des programmes de contrôle et d'imputation, certains enregistrements sont prélevés afin d'être examinés par les analystes. Les critères types régissant le prélèvement des enregistrements sont: 1) des changements importants aux données déclarées; 2) la nécessité d'imputer des données pour des établissements de grande taille; enfin, 3) l'impossibilité d'imputer une valeur susceptible de passer avec succès les contrôles de

[1] L. Draper et T. Petkunas, Statistical Research Division; B. Greenberg, Business Division, Bureau of the Census, Department of Commerce, Washington, D.C. 20233 U.S.A..

tolérance. Au cours du processus d'examen, l'analyste a le questionnaire du répondant en main, il est en mesure de joindre le répondant au téléphone et il dispose d'autres sources de données afin d'imputer une valeur raisonnable pour une non-réponse ou une valeur supposée erronée. L'examen des cas de renvoi par les analystes nécessite l'exécution des tâches suivantes: 1) traitement des enregistrements au bureau central; 2) transmission de la liste des cas de renvoi à Jeffersonville; 3) correction manuelle des documents de renvoi; 4) saisie des corrections; 5) transmission des corrections au bureau central; enfin, 6) cycle de traitement ultérieur au bureau central. Les possibilités de retard et de nouvelles erreurs sont donc nombreuses et ne font que se multiplier à mesure que le nombre de cycles de traitement augmente.

Il était devenu clair, depuis un certain temps, qu'il était nécessaire d'offrir aux analystes la possibilité d'effectuer leur examen des cas de renvoi en mode interactif. L'objectif visé est de permettre à l'analyste d'apporter les corrections directement sur l'enregistrement de données et de faire en sorte que les modifications soient contrôlées au fur et à mesure. Bien que le système SPEER ait été initialement conçu pour assurer le traitement des enregistrements par lots, il a été perfectionné depuis et permet maintenant d'examiner les cas de renvoi et d'introduire les données en mode interactif. En outre, le système est assez souple pour pouvoir satisfaire à un large éventail d'exigences spécifiques propres à chaque utilisateur.

Nous décrivons à la section 2 les fonctions exécutées par le système SPEER, la structure du système et les méthodes fondamentales auxquelles il fait appel. Toutefois, cet article n'étudie pas ces sujets en profondeur et le lecteur est invité à consulter Greenberg (1981, 1982, 1987) ainsi que Greenberg et Surdi (1984) pour obtenir plus de renseignements. On trouve dans Greenberg et Petkunas (1990) une description des applications réelles du système SPEER et de l'incidence qu'a eu l'interaction avec les utilisateurs sur l'évolution du système. Dans un certain sens, cet article et celui-ci sont complémentaires. À la section 3, nous décrivons les possibilités interactives du système SPEER et étudions l'utilisation qui en a été faite pour le contrôle des cas de renvoi et la saisie des données. Le lecteur y trouvera aussi deux exemples destinés à illustrer la souplesse du système ainsi que les options qu'il offre.

Les versions du SPEER assurant un traitement par lots et celles assurant un traitement en mode interactif sont toutes les deux pourvues du même programme de base. Toutefois, la version du système assurant un traitement en mode interactif marque une pause en certains points du code pour permettre la saisie de données au clavier. Le système est piloté par menu et permet à l'utilisateur de dialoguer avec tous les sous-programmes de SPEER. Les possibilités interactives et les images-écrans illustrées dans le présent article ont été choisies par les agents spécialisés de "l'Economic Surveys Division" et de "l'Industry Division" avec lesquels nous avons travaillé en étroite collaboration.

2. MODE DE FONCTIONNEMENT DU SYSTÈME SPEER

Le système SPEER est un système polyvalent de contrôle et d'imputation des données numériques faisant l'objet de contrôles de rapports. Chaque enregistrement sur les établissements consiste en un vecteur de zones de données numériques et un contrôle de rapport est une règle stipulant que le quotient de deux zones doit tomber entre deux limites prédéterminées qui sont mises en mémoire sous forme de paramètres. Le contrôle de rapport type prend la forme de $L_{ij} \le x_i/x_j \le U_{ij}$, où L_{ij} et U_{ij} sont les limites inférieure et supérieure fixées pour le rapport de x_i à x_j. Ainsi, le rapport du montant total des salaires versés aux travailleurs de la construction pendant l'année au nombre total d'heures travaillées par ces travailleurs pendant l'année doit se situer à l'intérieur de limites raisonnables. Soit $L_{12} \le x_1/x_2 \le U_{12}$ et $L_{23} \le x_2/x_3 \le U_{23}$ deux contrôles de rapports, nous avons alors le contrôle implicite $L_{12}L_{23} \le x_1/x_3 \le U_{12}U_{23}$. Au départ, un sous-programme de génération de contrôles établit tous les contrôles implicites qu'on peut obtenir à partir d'un ensemble de contrôles explicites fournis par l'utilisateur.

Ensuite, des sous-programmes de vérification de contrôle déterminent quels sont (parmi l'ensemble de contrôles implicites) les contrôles auxquels un enregistrement donné est accepté ou rejeté. Si l'enregistrement subit tous les contrôles avec succès et qu'aucun élément d'information n'est manquant, l'enregistrement est considéré comme acceptable. Non seulement le système permet-il d'utiliser les données déclarées pour la période courante pour la vérification de contrôle, mais il permet également de comparer les données aux données de l'année précédente ou à d'autres données comparables (voir la section 3). Si l'enregistrement subit tous les contrôles

avec succès mais que certains éléments d'information sont manquants, il fait l'objet d'une autre opération visant à imputer les données manquantes. Si l'enregistrement est rejeté à un ou plusieurs contrôles, il fait l'objet d'une opération de repérage des erreurs visant à déterminer un ensemble de zones devant être supprimées afin que les autres zones soient mutuellement compatibles. L'objectif poursuivi consiste à supprimer un ensemble minimal pondéré de zones.

Avant d'imputer les données manquantes (données déclarées et supprimées ou données non déclarées), le système détermine une plage d'imputation. Supposons que l'enregistrement comporte n zones et que (après un reclassement, au besoin) les données des zones $x_1,...,x_k$ où $k \leq n$ ont été déclarées et n'ont pas été ciblées en vue d'une modification ultérieure (ces zones sont mutuellement compatibles). En particulier, les rapports $L_{ij} \leq x_i/x_j \leq U_{ij}$ se vérifient pour tous les $i,j \leq kj$. Lorsque $k=n$, l'enregistrement est cohérent.

Si $k<n$, nous établissons une plage d'imputation pour x_{k+1}. On notera que pour tous les $j<k$, nous obtenons le rapport $L_{k+1,j} \leq x_{k+1}/x_j \leq U_{k+1,j}$.; en multipliant ce rapport par x_j, nous obtenons également la paire d'inégalités $x_j L_{k+1,j} \leq x_{k+1} \leq x_j U_{k+1,j}$ où $x_j, L_{k+1,j}$ et $U_{k+1,j}$ sont connus. Chaque $j=1, ..., k$ détermine un intervalle à l'intérieur duquel x_{k+1} doit se situer pour être compatible avec x_j. Si x_{k+1} se situe à l'intersection des intervalles k, il sera compatible avec chacune des zones x_j. Lorsque les contrôles sont compatibles, l'intersection n'est pas vide et elle constitue la plage admissible pour la zone x_{k+1}.

Il existe pour chaque zone un module d'imputation qui contient une séquence de règles d'imputation fournies par le personnel affecté à l'enquête. Lorsqu'une zone est sélectionnée en vue de l'imputation, la plage admissible pour cette zone est calculée et le programme sollicite le module d'imputation pour obtenir une imputation admissible. Les variables du SPEER sont typiquement classées en deux catégories: les éléments d'information de base et les éléments d'information secondaires. Les variables de base, qui sont les variables fondamentales pour l'exploitation d'une entreprise, sont contrôlées conjointement par un programme résidant en mémoire centrale, comme il est décrit ci-devant. Les éléments d'information secondaires sont groupés en ensembles auxiliaires d'éléments connexes qui sont comparés les uns aux autres. Les variables de base sont d'abord comparées les unes aux autres, puis le programme vérifie la cohérence des variables de base et des variables secondaires ainsi que la cohérence des variables secondaires faisant partie du même ensemble auxiliaire.

3. UTILISATION DU SYSTÈME SPEER EN MODE INTERACTIF

3.1 Examen des cas de renvoi

Le "Enterprise Statistics Program" assure la diffusion d'une série de publications à l'échelle de l'entreprise, au nombre desquelles figurent les publications "Large Companies" et "Auxiliary Establishments".

La publication "Large Companies" est établie à partie des réponses aux questionnaires remplis par les entreprises comptant au moins 500 employés et les tableaux y figurant font état de certaines statistiques financières sur les grandes entreprises. La publication "Auxiliary Establishments" présente les données relatives aux unités auxiliaires des entreprises à établissements multiples. Les établissements auxiliaires jouent un rôle de soutien par rapport aux autres établissements de l'entreprise. Il peut s'agir de centres de recherche et de développement, d'entrepôts et de bureaux administratifs. On recueille les données figurant dans la publication "Large Companies" à l'aide de la formule "Enterprise Summary Report" (ES-9100) et les données figurant dans la publication "Auxiliary Establishments", à l'aide de la formule "Auxiliary Establishment Report" (ES-9200).

Les premières versions du système SPEER ont été élaborées pour contrôler les "Enterprise Summary Report" et "Auxiliary Establishment Report" de 1982. Le système a de nouveau été utilisé pour contrôler ces deux formules de déclaration au cours des recensements de l'économie de 1987. À cette occasion, nous avons également tiré parti des nouvelles possibilités interactives du système aux fins de l'examen des cas de renvoi.

Lorsqu'il examine un cas de renvoi à l'aide de la version du système SPEER assurant le traitement en mode interactif, l'analyste doit solliciter un enregistrement, puis indiquer quelle zone il souhaite examiner et

éventuellement réviser. Si l'analyste doit examiner une deuxième zone sur le même enregistrement, le programme commande l'affichage de renseignements de la même façon que pour la première zone. Les valeurs relatives à la deuxième zone sont fondées en partie sur la nouvelle valeur inscrite dans la première zone révisée. Lorsque toutes les zones ont été examinées et corrigées, l'examen est terminé et on peut mettre fin au traitement par lots. L'analyste peut dialoguer avec le programme automatisé afin d'accroître l'expertise du système et d'annuler l'effet des règles de traitement par lots. L'écran a été conçu et le système a été modifié en fonction des demandes du personnel affecté à l'enquête.

Lors de l'examen des cas de renvoi relatifs à la formule ES-9100, un message de guidage apparaît à l'écran pour inviter l'utilisateur à solliciter un enregistrement. L'utilisateur peut 1) solliciter l'enregistrement suivant, 2) solliciter un enregistrement spécifique, 3) solliciter le prochain cas non résolu, 4) retourner au premier enregistrement du fichier, 5) rappeler l'enregistrement en cours d'examen, 6) introduire en mémoire un enregistrement sur une entreprise, 7) introduire en mémoire un enregistrement sur un établissement auxiliaire, enfin, 8) sortir du système SPEER.

L'écran n° 1 permet de traiter les éléments d'information de base recueillis sur la formule de déclaration ES-9100, tandis que les éléments d'information secondaires sont traités à partir des écrans subséquents. Les programmes de contrôle de la formule ES-9100 comprennent 3 écrans, 9 éléments d'information de base et 46 éléments d'information secondaires. L'en-tête indique le numéro de fichier de recensement, le nom de l'établissement, le code de catégorie de 1982, le code de catégorie de 1987 et le numéro de référence du microfilm.

Figure 1: Image-écran relative à l'écran n° 1 du programme ES-9100

9999999901 The American Weigh CAT82:999A CAT87:9999 99991

Mnem	Current	Reported	ST	Lower	Upper
EMP	1 000	1 000	R	971	2 963
APR	32 000	32 000	R	17 066	32 947
QPR	7 111	3 000	X	4 267	11 250
FBR	2 843	0		1 545	5 525
SLS	120 000	120 000	R	74 866	180 415
AET	66 945	100 000	X	65 632	69 804
TOT	60 000	60 000	R	57 543	203 031
RPT	1 200	0	I	64	2 160
ADE	50 500	50 500	R	14 202	51 510

Action taken: Mult: 1.0 Analyst: TFP 2/14/88 Rank:17
Flags: AETDET AETDIMP TOTDET TOTDIMP ABTDET ABTDIMP FGCET

ACTIONS: 0.Accept 1.Delete 2.Run SPEER 3.Restore reported
5.Impute 6.Restore complx 7.Next screen 8.Return
9.View reported C.View complex M.Change mult

La première colonne indique le symbole mnémonique de chaque élément d'information de base: nombre d'employés (EMP), charges de personnel annuelles (APR), charges de personnel au premier trimestre (QPR), charges sociales à absorber (FBR), chiffre d'affaires (SLS), actif total de clôture (AET), actif total (TOT), charges de location totales (RPT), et montant cumulé des amortissements à la clôture de l'exercice (ADE). Les deux colonnes suivantes indiquent les valeurs des données pour chacun des éléments d'information de base, les montants d'argent étant exprimés en milliers de dollars. La colonne deux indique les valeurs obtenues à la suite du contrôle par lots (valeur courante), tandis que la colonne trois indique les valeurs déclarées pour chaque élément d'information de base. Lorsque la valeur déclarée diffère de la valeur courante, cette dernière est mise en évidence.

La quatrième colonne indique le code d'état courant pour chaque élément d'information de base: valeur déclarée plus grande que zéro et acceptée aux contrôles (R), valeur déclarée plus grande que zéro et modifiée par SPEER (X), valeur déclarée plus grande que zéro et forcée à zéro (Z), valeur positive imputée pour une valeur déclarée égale à zéro (I), et non-réponse forcée à zéro (N). Les deux dernières colonnes indiquent les limites inférieure et supérieure de la plage admissible pour chaque élément d'information de base. On trouve sur la première ligne suivant l'affichage des données le multiplicateur d'enregistrement, l'identification de l'analyste, la date et le rang de l'entreprise. Généralement, l'analyste accordera une plus grande attention aux entreprises occupant un rang élevé.

Les multiplicateurs d'enregistrement permettent à l'utilisateur d'accroître l'amplitude de la plage admissible afin que des valeurs de données tombant à l'extérieur de la plage habituelle puissent être acceptées. Supposons que les contrôles de rapports se présentent sous la forme $(1/m)L_{ij} \le x_i/x_j \le mU_{ij}$ où $m \ge 1$ est le multiplicateur. Lorsque $m=1$, nous obtenons les rapports explicites initiaux; lorsque $m=1$, les limites inférieure et supérieure du rapport sont repoussées. En conséquence, la limite inférieure de la plage admissible est divisée par m, la limite supérieure est multipliée par m et l'amplitude de la plage s'en trouve accrue.

Les codes figurant à la ligne suivante indiquent les modifications apportées à l'ensemble de l'enregistrement et permettent à l'analyste de disposer d'un instantané de l'enregistrement sans avoir à parcourir tous les écrans. Ainsi, le code FGCET indique que le montant des dépenses en immobilisations, qui est indiqué sur l'écran n° 2, a fait l'objet d'une modification substantielle. C'est l'utilisateur qui définit la valeur du paramètre "modification substantielle". Le menu donne la liste des options que les agents spécialisés ont choisi d'offrir à l'utilisateur.

0. Accept: Cette option indique que l'état actuel de l'enregistrement est acceptable - soit après le traitement par lots, soit après les modifications apportées par l'analyste.

1. Delete: Cette option permet d'effacer un enregistrement dans la base de données.

2. Run SPEER: En appelant SPEER, l'analyste peut voir immédiatement quel sera l'effet des modifications apportées sur le reste de l'enregistrement.

3. Restore reported: Cette option permet à l'analyste de commander la récupération des données initialement déclarées en appuyant sur une touche.

5. Impute: Cette option permet à l'analyste de remplacer toutes les valeurs par des blancs afin d'imputer toutes les valeurs de l'enregistrement à partir de seulement quelques données. L'analyste doit utiliser cette option pour imputer les données à partir des dossiers administratifs.

6. Restore complex: Cette option permet à l'analyste de commander la récupération des valeurs ayant figuré dans l'enregistrement au début de la séance.

7. Next Screen: Cette option commande l'affichage de l'écran suivant, qui contient d'autres éléments d'information, généralement des éléments secondaires.

8. Return: Cette option permet à l'analyste de réintégrer l'enregistrement dans la base de données en vue d'un examen ultérieur.

9. View reported: Cette option permet à l'analyste de commander l'affichage de toutes les valeurs initialement déclarées sur un même écran afin d'obtenir une vue d'ensemble de l'établissement sans avoir à parcourir tous les écrans.

C. View complex: Cette option permet à l'analyste de commander l'affichage de toutes les valeurs en cours de contrôle sur un même écran.

M. Change mult: Cette option permet à l'analyste de modifier le multiplicateur de l'enregistrement en cours de traitement pour annuler l'effet du multiplicateur en place.

Les options 1, 3, 5 et 6 sont dotées de dispositifs de protection et il faut appuyer à deux reprises sur la touche appropriée pour les appeler. Chaque fois qu'une de ces options est sélectionnée, le programme lance un avertissement pour éviter que l'analyste superpose par erreur des données sur les données courantes ou efface un enregistrement de la base de données par inadvertance.

Les programmes du système SPEER peuvent générer une grande quantité de messages de diagnostic et il revient au personnel affecté à l'enquête de déterminer quels messages seront affichés à l'écran. Non seulement le système peut-il générer un grand nombre de messages de diagnostic pendant l'examen de l'enregistrement, mais il est capable de fournir les renseignements nécessaires pour suivre le déroulement du processus d'examen. Ainsi, il est possible de savoir combien de fois les analystes utilisent un multiplicateur, acceptent les mesures prises par le système automatisé, annulent l'effet des programmes de traitement par lots, et ainsi de suite. Ces possibilités peuvent faciliter la surveillance des activités des analystes et l'évaluation de leur performance. En outre, les chefs de programme peuvent utiliser ces renseignements pour mieux comprendre et peut-être même améliorer ce processus hautement subjectif.

Lorsque la valeur introduite par un analyste n'est pas compatible avec les autres éléments d'information de base, une sonnerie se fait entendre, le menu disparaît et un message est affiché à l'écran. Par exemple, si, dans la figure 1, l'analyste introduisait la valeur $ 72 000 pour la variable AET, le message suivant apparaîtrait à l'écran: "Lo = 65 632, AET = 72 000, Up = 69 804; 72 000 is not within current bounds. Should this value be accepted (Y/N)? = = >".

Si l'analyste souhaite que cette valeur soit acceptée, il répond "YES" et un nouveau message est affiché pour lui demander s'il souhaite modifier le multiplicateur de l'enregistrement. On notera qu'un nouveau multiplicateur est affiché et que les limites supérieure et inférieure pour chaque élément d'information de base ont été modifiées, le nouveau multiplicateur ayant été respectivement utilisé pour multiplier les limites supérieures et diviser les limites inférieures.

Figure 2: Effet de la modification du multiplicateur (écran abrégé)

EMP	1 000	1 000	R	898	3 204
APR	32 000	32 000	R	15 781	35 631
QPR	7 111	3 000	X	3 945	12 166
FBR	2 843	0		1 429	5 965
SLS	120 000	120 000	R	74 454	195 112
AET	72 000	100 000	X	60 689	75 490
TOT	60 000	60 000	R	57 226	219 570
RPT	1 200	0	I	59	2 336
ADE	50 500	50 500	R	14 124	58 085

Action taken: Mult: 1.1 Analyst: TFP 2/14/88 Rank: 17

Nous recueillons également des données sur les entreprises en agrégeant les données sur les établissements déclarées dans d'autres segments des recensements de l'économie. Les données déclarées pour tous les établissements appartenant à une même entreprise sont recueillies à partir d'autres sources comme les recensements du commerce de gros, du commerce de détail, des industries de services, des manufactures, des industries minières, des industries de la construction et de certaines industries des transports. Les données obtenues de ces diverses sources sont agrégées afin d'établir un enregistrement sommaire sur l'entreprise étudiée. Ces données sommaires servent ensuite au contrôle et à l'imputation des données sur l'entreprise recueillies sur la formule ES-9100. Pour utiliser les données sommaires afin d'imputer les données manquantes ou erronées de la formule ES-9100, il suffit de sélectionner l'option 4 sur le menu ci-après, comme pour l'élément d'information FBR.

Figure 3: Écran n° 1 avec données sommaires pour la formule ES-9100 (écran abrégé)

9999999901 The American Weigh CAT82:999A CAT87:9999 99991

Mnem	Current	Summary	Reported	ST	Lower	Upper
EMP	1 000	987	1 000	R	971	2 963
APR	32 000	30 117	32 000	R	17 066	32 947
QPR	7 111	7 111	3 000	X	4 267	11 250
FBR	2 843	2 843	0		1 545	5 525
SLS	120 000	101 056	120 000	R	74 866	180 415
AET	66 945	66 945	100 000	X	65 632	69 804
TOT	60 000	0	60 000	R	57 543	203 031
RPT	1 200	1 200	0	I	64	2 160
ADE	50 500	0	50 500	R	14 202	51 510

ACTIONS: 0.Accept 1.Delete 2.Run SPEER 3.Restore reported
4.Sum replace 5.Impute 6.Restore complx 7.Next screen
8.Return 9.View reported C.View complex M. Change mult

Les programmes du système SPEER sont rédigés en FORTRAN et le système peut facilement être transféré d'un système d'exploitation à un autre. De même, les programmes ont pu être adaptés aux micro-ordinateurs sans difficulté. Les versions du système SPEER assurant le traitement par lots des formules ES-9100 et ES-9200 ont été exécutées sur le système d'exploitation UNISYS en raison surtout des liaisons d'intercommunication établies entre Suitland et Jeffersonville. Après que les enregistrements aient été traités par lots sur l'unité centrale UNISYS, les cas de renvoi ont été téléchargés et les analystes ont examiné ces cas à l'aide de micro-ordinateurs IBM raccordés par l'intermédiaire du réseau local et partageant une même base de données.

3.2 Saisie des données

L'enquête annuelle sur les manufactures (EAM) permet d'établir des estimations intercensitaires annuelles de certains indicateurs clés de l'activité manufacturière. Les données relatives à ces indicateurs clés, ainsi que d'autres statistiques détaillées sur les industries manufacturières, sont recueillies au moyen des recensements des manufactures. Depuis 1949, on réalise une enquête au cours de chaque année intercensitaire afin de disposer d'une série continue de statistiques de base ainsi que de données repères pour l'établissement d'indicateurs de l'activité économique courante et de mesures de la production industrielle et de la productivité. "L'Industry Division" a conçu, en s'inspirant du système SPEER, un programme de contrôle qui a été utilisé pour le recensement des manufactures de 1987 ainsi que pour chaque enquête annuelle sur les manufactures depuis 1986.

L'enquête annuelle sur les manufactures est réalisée auprès d'un échantillon d'établissements manufacturiers prélevé dans l'univers du recensement des manufactures. Les établissements sélectionnés participent à l'enquête pendant quatre années intercensitaires consécutives et reçoivent à cet effet des questionnaires qu'ils doivent retourner par la poste. Les formules de réponse reçues après la date de fermeture des installations de saisie de Jeffersonville sont appelées "ajouts tardifs" et les données qui y figurent sont introduites dans la base de données par les analystes du bureau central. Généralement, on reçoit de 2 000 à 2 500 ajouts tardifs chaque année pour l'EAM et jusqu'à 15 000 pour le recensement des manufactures. Dans le passé, les ajouts tardifs étaient introduits dans la base de données sans être soumis aux programmes de contrôle automatisé par lots et certains enregistrements sur les établissements de grande taille ont été introduits dans la base sans avoir fait l'objet d'un contrôle machine.

Le personnel affecté à l'enquête annuelle sur les manufactures a donc demandé d'être pourvu d'une version du système SPEER assurant le traitement en mode interactif afin d'introduire les données relatives aux ajouts tardifs. Ce système permet de contrôler les données au moment de leur introduction sans qu'il soit nécessaire de les soumettre à un contrôle par lots. Les programmes sont pilotés par menu et respectent la structure de base du système SPEER, les écrans et les options spécialisés étant conçus pour répondre aux besoins du personnel affecté à l'enquête annuelle sur les manufactures.

Nous allons maintenant décrire le système de saisie en mode interactif des données relatives aux ajouts tardifs pour l'enquête annuelle sur les manufactures de 1988. Ce système était exécuté sur des micro-ordinateurs, puis les données ainsi introduites étaient ultérieurement téléchargées dans l'unité centrale UNISYS. L'exécution du programme commence par l'affichage d'un message invitant l'analyste à introduire ses initiales. Les valeurs des paramètres de contrôle, constituées d'une limite inférieure, d'une limite supérieure et de la moyenne à l'échelle de la branche d'activité pour chaque rapport explicite, sont introduits pour 529 codes de la Classification Type des Industries (CTI). Ces valeurs comprennent, pour chaque code CTI, un ensemble distinct de limites de contrôle pour l'année en cours et pour l'année précédente. Tant les règles de contrôle que les règles d'imputation utilisent les données relatives à l'année précédente lorsqu'on en dispose. Les éléments d'information sont comparés aux données de l'année en cours et aux données de l'année précédente, et chaque élément doit subir les deux ensembles de contrôles avec succès pour être considéré comme acceptable.

Le programme utilise les données relatives à l'année précédente pour contrôler les données déclarées pour l'année en cours au moyen d'un contrôle de rapport interannuel. Soit $x_i x_j y_i$ et y_i les valeurs respectives des données pour l'année en cours et l'année précédente, le contrôle de rapport interannuel est $l_{ij} \le [x_i/x_j]/[y_i/y_i] \le u_{ij}$ où l_{ij} et u_{ij} sont les limites inférieure et supérieure du rapport [le lecteur trouvera un exposé détaillé des contrôles de rapports interannuels dans Greenberg (1981)]. Le programme calcule des plages admissibles à partir des valeurs relatives à l'année en cours et des valeurs relatives à l'année précédente, puis c'est l'intersection de ces deux plages qui constitue la plage admissible définitive. Lorsqu'on ne doit pas employer les données relatives à l'année précédente pour un enregistrement donné, le programme fonctionne en utilisant uniquement les données relatives à l'année en cours.

Le premier écran sert à introduire les données relatives aux éléments d'information de base. Au départ, l'analyste doit introduire un numéro d'établissement permanent (NÉP) de 10 chiffres, suivi d'un code CTI de 6 chiffres correspondant à un des 529 codes CTI figurant dans une liste distincte. Si le NÉP ne contient pas 10 chiffres ou si le code CTI ne correspond à aucun des codes de la liste, les messages d'erreur appropriés sont affichés à l'écran et l'utilisateur doit introduire les données de nouveau. Lorsque l'utilisateur introduit un code CTI valide, les paramètres relatifs à l'année en cours et à l'année précédente sont introduits pour ce code à partir d'un fichier à accès direct. Cette caractéristique permet de réduire au minimum la capacité de la mémoire de travail du programme.

Figure 4: Écran d'introduction des données en mode interactif pour l'EAM.

PPN: 1234567890 SIC: 201112

Mnem	Reported	Prior	ST	Lower	Upper
SW	1 000	21 000	R	600	3 076
VS	25 000	30 000	R	11 029	54 546
OW	600	700	R	35	1 000
WW	500	1 300	R	533	1 000
PW	50	100	R	46	70
PH	125	175	R	53	117
TIE	2 500	1 700	R	0	5 666
TE	-1	-1	N	50	200
TIB	1700	-1	R	250	4 999 995
LE	-1	-1	N	10	243
OE	-1	-1	N	10	120
CM	-1	-1	N	1 250	15 000
VP	-1	-1	N	10	400

Mult: 1.0 Editor: Lisa Date: DEC 3, 1990

La première colonne indique les symboles mnémoniques de chacun des éléments d'information de base. Ces éléments et symboles sont: salaires et traitements (SW); valeur des expéditions (VS); salaires des autres employés

(ON) (membres du personnel auxiliaire); salaires des travailleurs de la production (WW); nombre de travailleurs de la production (PW); heures travaillées à l'usine par les travailleurs de la production (PH); stock total à la clôture de l'exercice (TIE); emploi total (TE); stock total au début de l'exercice (TIB); dépenses exigées par la loi (LE) (comme l'assurance sociale, etc.); autres employés (OE) (membres du personnel auxiliaire); coût des matières (CM); enfin, charges facultatives (VP) (assurance maladie, fonds de retraite, etc.).

La deuxième colonne indique les données introduites à partir de la formule de déclaration. La colonne suivante renferme les données relatives à l'année précédente, que l'analyste peut saisir au choix. La quatrième colonne (ST) réferre aux codes d'état, lesquels indiquent la provenance des chiffres. Enfin, les deux dernières colonnes indiquent la plage admissible de chaque zone. Pour qu'un enregistrement soit cohérent, chacune de ses zones doit se situer à l'intérieur de sa plage admissible. La valeur "-1" est utilisée comme substitut pour une zone vierge.

Lorsqu'un analyste introduit un enregistrement en mémoire, il peut voir les limites se modifier à l'écran au fur et à mesure qu'il introduit les éléments d'information au clavier, puisque le programme calcule les nouvelles limites inférieure et supérieure de la plage admissible chaque fois qu'un nouvel élément est introduit. En effet, bien que le programme évalue la cohérence de chaque donnée relative à l'année en cours avec les autres éléments d'information lorsque cette donnée est introduite, aucune modification n'est apportée à ce moment.

Une fois que tous les renseignements figurant sur la formule ont été saisis, le programme de contrôle SPEER est exécuté sur les données et un nouvel écran est affiché. Cet écran indique les valeurs déclarées, les valeurs courantes (contrôlées) et les valeurs relatives à l'année précédente. Par ailleurs, le code CTI est fractionné en code de branche d'activité économique et en code de sous-branche, le numéro de l'enregistrement figure à côté de la date et le menu indiquant les options offertes est affiché au bas de l'écran. L'analyste a alors la possibilité d'apporter d'autres modifications à l'enregistrement ou de l'accepter tel quel. Les options figurant dans le menu sont assez semblables aux options offertes pour les formules ES-9100.

Figure 5: Écran n° 1 des programmes de contrôle en mode interactif des données de l'EAM.

PPN: 1234567890 IND: 2011 SUB: 12

Mnem	Current	Prior	Reported	ST	Lower	Upper
SW	1 100	2 000	1 000	X	600	3 076
VS	25 000	30 000	25 000	R	14 583	60 000
OW	600	700	600	R	115	1 100
WW	500	1 300	500	R	375	1 100
PW	50	100	50	R	33	70
PH	88	175	125	X	53	117
TIE	2 500	1 700	2 500	R	0	5 666
TE	73	-1	-1	I	50	220
TIB	1 700	-1	1 700	R	250	2 299 997
LE	121	-1	-1	I	11	267
OE	23	-1	-1	I	10	73
CM	8 750	-1	-1	I	1 250	15 000
VP	101	-1	-1	I	11	440
LC	222				N/A	N/A

Action taken: Mult: 1.0 Editor: Lisa Date: DEC 3, 1990 1
ACTIONS: 1.Accept record 2.Run thru SPEER 3.Set multiplier
5.Change status flag 6.Next screen 7. Print audit trail
8.Start record over 9.Print record

La version du système SPEER assurant le traitement en mode interactif s'est révélée d'une grande efficacité pour introduire les données relatives aux ajouts tardifs de l'enquête annuelle sur les manufactures de 1988. Par suite de l'utilisation réussie des programmes décrits ci-devant, on nous a demandé d'élaborer une version interactive

de ces programmes destinée à être installée sur le micro-ordinateur VAX qui sollicitera la base de données de l'EAM en direct. Nous travaillons actuellement à élaborer la version 2.0 du système, qui comprend pour l'instant 3 écrans: 2 écrans de données secondaires et 1 écran d'éléments d'information de base. Pour l'essentiel, ces écrans comportent les mêmes renseignements que les écrans de la version 1.0, mais ils font état d'un plus grand nombre de données secondaires et comportent des menus offrant des options différentes à l'analyste.

Comme la version 2.0 du système permettra de solliciter la base de données en direct, l'analyste n'aura plus besoin d'introduire les données relatives à l'année précédente au clavier puisque ces données figureront déjà dans la base. Grâce à cette caractéristique, l'utilisateur n'aura qu'à taper un numéro d'onglet à l'aide duquel l'enregistrement sera apparié à l'enregistrement correspondant de la base de données. L'enregistrement apparié sera alors affiché à l'écran. Le programme exécutera ensuite une vérification interne visant à déterminer, à partir du rapport entre le montant des traitements et salaires de l'année courante et celui des traitements et salaires de l'année précédente, si le contrôle doit tenir compte des données de l'année précédente. Si ce rapport ne se situe pas à l'intérieur d'une plage admissible, les données de l'année précédente ne seront pas prises en considération. En pareil cas, ou si aucun des enregistrements de la base de données ne peut être apparié au numéro d'onglet introduit au clavier, le programme contrôlera l'enregistrement en utilisant uniquement les données relatives à l'année courante.

Si un enregistrement subit avec succès tous les contrôles portant sur les données de l'année courante mais est rejeté à certains des contrôles portant sur les données de l'année précédente, les règles de contrôle relatives aux données de l'année précédente ne seront appliquées ni à l'étape du contrôle ni à l'étape de l'imputation. Si un enregistrement subit avec succès tous les contrôles portant sur les données de l'année précédente mais est rejeté à certains des contrôles relatifs aux données de l'année courante, les règles relatives à la cohérence interne des données de l'année courante ne seront appliquées ni à l'étape du contrôle ni à celle de l'imputation. Dans le premier cas, nous acceptons qu'un enregistrement intrinsèquement cohérent diffère d'une année à l'autre; dans le second, nous acceptons que l'établissement offre une performance atypique pour l'année courante. Dans les deux cas, notre objectif est de permettre à un plus grand nombre de données déclarées de subir les contrôles avec succès et de tenter d'éviter les contrôles excessifs. Après le repérage des erreurs, la version 2.0 du programme marque une pause avant d'effacer les zones et met en évidence les zones susceptibles d'être effacées. Cette nouvelle caractéristique permet à l'utilisateur de choisir les zones à modifier sans utiliser les sous-programmes de repérage d'erreurs automatique. Toutefois, il a toujours l'option d'utiliser ces sous-programmes.

Pour l'instant, tout exposé de la version 2.0 du système est condamné à être très incomplet puisque les programmes sont toujours en cours d'élaboration. Cette version du système SPEER assurant le traitement en mode interactif servira à introduire et à contrôler les données relatives aux ajouts tardifs de l'enquête annuelle sur les manufactures de 1989 et des années subséquentes ainsi que du recensement des manufactures de 1992.

4. CONCLUSION

L'élaboration des programmes qui allaient mener à la création du système SPEER a débuté en 1980. Notre objectif initial était de concevoir des programmes de contrôle des données numériques faisant l'objet de contrôles de rapports qui intégraient les nouvelles méthodes élaborées à Statistique Canada par Fellegi et Holt (1976) ainsi que Gordon Sande (1976, 1981). Au départ, nous avons travaillé en étroite collaboration avec le personnel de "l'Industry Division" pour concevoir un prototype destiné à être utilisé pour l'enquête annuelle sur les manufactures et nous n'avions nullement l'intention d'élaborer un système à utilisateurs multiples. Le système SPEER a cependant évolué à partir de ces travaux.

On peut se représenter l'actuelle version du système comme un squelette que nous personnalisons pour répondre aux besoins des diverses équipes d'enquête. Chaque nouvelle mise en application du système s'est traduite par l'apport de perfectionnements aux programmes de base. Nous n'avons pas attendu que le système dispose de toutes les fonctions souhaitables avant de le mettre en application et on peut considérer qu'il fait l'objet d'un développement continu. Les exemples des sections précédentes ont bien illustré la souplesse du système et notre étude de l'utilisation des données sommaires et des données relatives à l'année précédente a fait ressortir ses possibilités d'adaptation. Ce sont les besoins et les demandes des utilisateurs qui sont la source des modifications

apportées au système SPEER, et les travaux décrits dans le présent article ont été mis en oeuvre pour répondre à la demande de programmes permettant le traitement des données en mode interactif.

REMERCIEMENTS

Il nous fait plaisir de souligner l'excellence du travail de John Monaco et de Cindy Schott de l'Economic Surveys Division, qui ont conçu les programmes de contrôle et d'imputation pour les formules ES-9100 et ES-9200. John et Cindy ont assuré la direction des travaux en décrivant la configuration exigée, en déterminant les besoins à satisfaire et en demandant l'intégration des fonctions nécessaires. Il n'existe aucun élément de ces programmes pour lequel nous ne leur sommes redevables. Bob Rosati et Rich Sterner de l'Industry Division étaient responsables de la conception des programmes d'introduction des données de l'EAM pour 1988 et assurent actuellement la direction des travaux relatifs à la version 2.0 du programme pour l'EAM de 1989. Encore une fois, il n'existe aucun élément de ces programmes pour lequel nous ne leur sommes redevables.

BIBLIOGRAPHIE

Fellegi, I.P., et Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, 71, 17-35.

Greenberg, B. (1981). Developing an Edit System for Industry Statistics, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, Springer-Verlag, New York, 11-16.

Greenberg, B. (1982). Using an Edit System to Develop Editing Specifications, Proceedings of the Section on Survey Research Methods, *Journal of the American Statistical Association*, 366-371.

Greenberg, B. (1987). Edit and Imputation as an Expert System, in *Statistical Policy Working Paper Number 14: Statistical Uses of Microcomputers in Federal Agencies*, Statistical Policy Office, Office of Information and Regulatory Affairs, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, D.C., 85-92.

Greenberg, B., et Petkunas, T. (1990). SPEER (Structured Programs for Economic Editing and Referrals), *Proceedings of the Section on Survey Research Methods,* American Statistical Association, to appear.

Greenberg, B., et Surdi, R. (1984). A Flexible and Interactive Edit and Imputation System for Ratio Edits, Proceedings of the Section on Survey Research Methods, *Journal of the American Statistical Association*, 421-426.

Sande, G. (1976). Numerical Edit and Imputation, invited paper presented at the International Association for Statistical Computing, 42nd session of the International Statistical Institute, Manila, Philippines.

Sande, G. (1981). Descriptive Statistics Used in Monitoring Edit and Imputation Processes, presented at Workshop on Automated Edit and Imputation, *Computer Science and Statistics; Proceedings of the 13th Symposium on the Interface*, Springer-Verlag, New York.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

ESTIMATEURS-M ET ESTIMATEURS À L'ÉPREUVE DES VALEURS ABERRANTES EN REMPLACEMENT DE L'ESTIMATEUR PAR QUOTIENT

L.-P. Rivest et E. Rouillard[1]

RÉSUMÉ

Dans cet article, nous proposons des estimateurs à l'épreuve des valeurs aberrantes en remplacement de l'estimateur par quotient. Ceux-là sont construits à l'aide d'estimateurs-M robustes de la pente du modèle de régression à ordonnée à l'origine nulle, qui est à la base de l'estimateur par quotient. Les estimateurs à l'épreuve des valeurs aberrantes (EEVA) réduisent l'importance relative des unités d'échantillon qui sont considérées comme des unités à valeurs aberrantes par rapport au modèle de régression. Au moyen d'une étude de Monte Carlo, nous comparons plusieurs EEVA à l'estimateur par quotient. Nous en retenons un, que nous analysons plus en profondeur et dont nous étudions les propriétés par rapport au plan de sondage. L'estimateur retenu n'est pas convergent; nous donnons les formules pour son biais et sa variance dans le cas d'une taille d'échantillon élevée. Nous proposons aussi des estimateurs de l'erreur quadratique moyenne de l'EEVA. Les niveaux de confiance réels des intervalles produits à l'aide de ces estimateurs sont analysés au moyen d'une étude de Monte Carlo. Enfin, nous comparons le rendement de l'estimateur par quotient et celui de l'EEVA retenu, étant donné un nombre déterminé de valeurs aberrantes dans l'échantillon.

MOTS CLÉS: Analyse conditionnelle; intervalles de confiance; estimateurs-M; simulations de Monte Carlo; régression; plan de sondage.

1. INTRODUCTION

Le contrôle (des données) est une étape importante dans le traitement des données d'enquête. Il permet d'en vérifier la qualité et d'établir des estimations qui soient représentatives des populations étudiées. Une fonction importante du contrôle, du point de vue statistique, est l'examen des réponses anormales, c'est-à-dire des réponses qui diffèrent largement de ce que l'on pouvait attendre d'une unité particulière. Nous appellerons ce genre de réponses des valeurs aberrantes. Dans certains cas, les valeurs aberrantes sont attribuables à des erreurs de saisie, qui peuvent être corrigées à l'étape du contrôle. La réponse anormale est alors remplacée par la réponse juste. Ce genre de valeur aberrante est dite non représentative. Dans d'autres cas, les valeurs aberrantes reflètent l'état réel d'unités de l'échantillon. Par exemple, des unités de la population qui subissent de grands bouleversements en peu de temps peuvent devenir des unités à valeurs aberrantes. On parlera alors de valeur aberrante représentative. Ce dernier type de valeur fait partie intégrante des populations étudiées; on doit en tenir compte dans l'estimation des caractéristiques d'intérêt. Autrement dit, on doit pouvoir rattacher une partie de la variance d'échantillonnage des valeurs estimées des caractéristiques au traitement réservé aux valeurs aberrantes représentatives. Le fait de remplacer ces valeurs par des valeurs imputées, à l'étape du contrôle, et de considérer ces dernières comme des valeurs réelles à l'étape de l'estimation entraîne une sous-estimation de la variance d'échantillonnage des estimateurs d'enquête.

[1] L.-P. Rivest et E. Rouillard, Département de mathématiques et de statistique, Université Laval, Cité universitaire, Québec, G1K 7P4, Canada.

Les méthodes statistiques utilisées pour les valeurs aberrantes contenues dans les données d'enquête comportent normalement deux étapes. La première consiste à détecter ces valeurs selon des critères fondés sur les données. La seconde consiste à soumettre les valeurs aberrantes détectées à un traitement particulier. En règle générale, on réduit la valeur y ou le poids d'échantillonnage correspondants. On recourt souvent à la winsorisation, c'est-à-dire au remplacement d'une valeur y par une constante déterminée ou une constante qui repose sur des données. Ces méthodes ont été analysées par plusieurs auteurs (Searl, 1966; Fuller, 1970; Rao, 1971; Ernest, 1980; Tambay, 1988). Ceux-ci montrent que le traitement des valeurs aberrantes peut réduire de façon notable l'erreur quadratique moyenne des estimateurs d'enquête.

Dans cet article, nous examinons des estimateurs conçus pour résister aux valeurs aberrantes et incorporant des procédures spéciales lorsque de l'information supplémentaire est utilisée pour l'estimation de caractéristiques de la population. Nous étudions divers estimateurs susceptibles de remplacer l'estimateur par quotient. Les estimateurs proposés reposent sur des estimateurs-M de "statistiques robustes". Chambers (1986), Gwet et Rivest (1990) et Lee (1990) ont étudié ce genre d'estimateur dans le contexte des enquêtes par sondage. Dans la section 2, nous construisons les estimateurs à l'épreuve des valeurs aberrantes. À l'aide d'une simulation de Monte Carlo, nous comparons ces estimateurs à l'estimateur par quotient, puis nous en retenons un afin de l'étudier plus à fond. Dans la section 3, il est question de l'estimation de l'erreur quadratique moyenne de cet estimateur et de la production d'intervalles de confiance à l'épreuve des valeurs aberrantes pour la moyenne de population. Enfin, dans la section 4, nous faisons une analyse conditionnelle -- étant donné un pourcentage de valeurs aberrantes dans l'échantillon -- des propriétés de l'estimateur à l'épreuve des valeurs aberrantes et de l'estimateur par quotient.

2. CONSTRUCTION DES ESTIMATEURS À L'ÉPREUVE DES VALEURS ABERRANTES

Nous proposons ici une méthode pour construire des estimateurs à l'épreuve des valeurs aberrantes (EEVA). Mais avant, revoyons les principes de base de l'estimateur par quotient. Un échantillon s est tiré d'une population U de taille N suivant un échantillonnage aléatoire simple sans remise. La taille de l'échantillon est n. Nous cherchons à estimer $\bar{Y}$, la moyenne de la variable Y pour la population, à l'aide des données de l'échantillon et de la variable auxiliaire X, qui est connue pour toutes les unités de U. Le modèle probabiliste (M), que l'on définit par l'expression

$$Y_i = \beta X_i + \varepsilon_i$$

où $E(\varepsilon_i) = 0$ et $V(\varepsilon_i) = \sigma^2 X_i$, décrit souvent avec justesse la relation entre Y et X. L'estimateur par quotient, $\bar{y}_r = \bar{X}\bar{y}/\bar{x}$, où $\bar{y}$ et $\bar{x}$ sont les moyennes empiriques des variables y et x respectivement, est alors un bon estimateur de $\bar{Y}$. Suivant le modèle (M), l'estimateur $\bar{y}_r$ peut être défini comme le produit de $\bar{X}$ par l'estimateur des moindres carrés de β. Cela nous amène à construire des EEVA qui pourraient être substitués à $\bar{y}_r$, notamment

$$\bar{y}_g = \bar{X}\hat{\beta}_g \qquad (1)$$

où $\hat{\beta}_g$ est l'EEVA du paramètre β du modèle (M).

Les EEVA de la pente du modèle (M) sont définis normalement au moyen d'équations implicites. Plusieurs solutions sont proposées à cette fin. Elles supposent toutes une fonction croissante bornée impaire, notée ψ, et une constante de robustesse, appelée k. Les versions les plus courantes de ψ sont la fonction de Huber, qui est définie

$$\psi_H(t) = sgn(t)\min(|t|,1),$$

où sgn(.) désigne la fonction signe, la fonction de Huber modifiée, qui s'écrit

$$\psi_{Hm}(t) = \begin{cases} \sin\left(\frac{\pi}{2}\, t\right) \text{ for } |t| \leq 1 \\ \text{sgn}(t) \text{ for } |t| > 1), \end{cases}$$

où sin(.) désigne la fonction sinus, et la fonction "fair", qui est définie

$$\psi_f(t) = \frac{t}{1+|t|}.$$

Rey (1983, section 6.4.10) compare ces trois fonctions entre elles. Elles sont de forme semblable et donnent des estimateurs qui ont les mêmes propriétés. Un avantage des deux dernières par rapport à la première est que leurs dérivées premières sont continues. Nous considérons trois types d'estimateur robuste pour le paramètre β du modèle (M). Il y a d'abord les estimateurs-M, qui sont définis comme la solution de

$$\sum_s \sqrt{x_i}\ \psi\ \left(\frac{y_i - \beta x_i}{k\sqrt{x_i}}\right) = 0,$$

où k est une constante de robustesse. Avec ce type d'estimateur, les unités à valeurs aberrantes pour lesquelles la valeur x est élevée ont encore une incidence appréciable. Cela a amené les chercheurs à élaborer des estimateurs-GM pour les paramètres de régression, ce qui a permis de réduire davantage l'incidence des unités à valeur x_i élevée. Les estimateurs-GM se divisent en deux catégories (voir Hampel, Ronchetti, Rousseeuw et Stahel, 1986). Ceux de la catégorie de Hampel-Krasker sont définis comme la solution de

$$\sum_s \psi\ \left(\frac{y_i - \beta x_i}{k}\right) = 0,$$

et ceux de la catégorie de Mallows sont obtenus par la résolution de l'équation

$$\sum_s \psi_1\ \left(\frac{\sqrt{x_i}}{k_1}\right)\ \psi_2\left(\frac{y_i - \beta x_i}{k_2\sqrt{x_i}}\right) = 0.$$

L'estimateur des moindres carrés $\bar{y}/\bar{x}$ appartient aux trois catégories définies ci-dessus. On peut l'obtenir en choisissant une valeur de k, k_1 ou k_2 arbitrairement élevée ou en utilisant une fonction $\psi(t)$, $\psi_1(t)$ ou $\psi_2(t)$, égale à t.

2.1 Description de la simulation de Monte Carlo

Au moyen d'une simulation de Monte Carlo, nous comparons le rendement de trois estimateurs qui représentent chacun une catégorie parmi celles définies ci-dessus. Pour cela, nous utilisons cinq populations, les quatre premières comptant 500 unités. Ces populations ont été construites selon la méthode proposée par Robinson (1987). La population 1 est construite suivant le modèle (M). L'équation de régression de la population 2 renferme un terme quadratique. La population 3 est semblable à la population 1, avec 7% de valeurs aberrantes aléatoires. Pour ce qui est de la population 4, la pente de l'équation de régression est égale à 2 pour la plupart des données simples; elle est égale à 1 pour 7% des unités. La population 5 renferme 235 unités. Nous l'avons formée à partir de la population de Kish (1965) en supprimant de celle-ci toutes les unités pour lesquelles $x \leq 2$ et en perturbant les valeurs y de 13 unités afin de créer des valeurs aberrantes. Les graphiques des cinq populations figurent à l'annexe 2.

Tableau 1: Estimateurs visés par l'étude de Monte Carlo

Type d'estimateur	Symbole	Fonction ψ	Constante de robustesse
Estimateur par quotient	RE	$\psi(t) = 1$	*sans objet*
Estimateur-M	ME	$\psi(t) = \psi_{Hm}(t)$	$k = 1.2107\sigma$
Estimateur-GM (Mallows)	GMM	$\psi_1(t) = \psi_H(t)$ $\psi_2(t) = \psi_{Hm}(t)$	$k_1 = 1.46(med(\sqrt{X_i})$ $k_2 = 1.2107\sigma$
Estimateur-GM (Hampel Krasker)	GMH	$\psi(t) = \psi_f(t)$	$k = \frac{med(\sqrt{X_i})}{.6745}\ 2.56\ \sigma$

La forme générale des trois estimateurs de $\bar{Y}$ étudiés est définie par l'équation (1); quant aux estimateurs $\hat{\beta}_g$ qui font l'objet de la présente simulation, ils sont définis dans le tableau 1. Nous avons analysé les propriétés de ces estimateurs pour de petits échantillons au moyen d'échantillons de 20 et de 30 unités pour chacune des cinq populations représentées à l'annexe 2. Chaque simulation comprenait 2000 répétitions.

Certaines constantes de robustesse dépendent de σ, l'écart type résiduel du modèle (M). Pour pouvoir utiliser l'EEVA dans la pratique, il nous faut une estimation de σ qui pourrait être établie à l'aide des données d'enquêtes précédentes. Rey (1983, section 6.4.5) souligne une propriété intéressante des estimateurs qui reposent sur $\psi_f(t)$: ils ne sont pas sensibles au changement de constante de robustesse (k). Par conséquent, dans le cas de ces estimateurs, un estimateur approximatif de σ suffit pour obtenir des résultats significatifs. Les valeurs constantes qui figurent dans les formules de la constante de robustesse ont été calculées de telle manière que, pour un modèle normal, l'efficacité de l'estimateur robuste par rapport à l'estimateur des moindres carrés soit de 95% (voir Rey, 1983, p. 105 et Hampel, Ronchetti, Rousseeuw et Stahel, 1986, p. 333).

Au point de vue du calcul, il est utile de considérer les EEVA $\hat{\beta}_g$ comme des rapports pondérés:

$$\hat{\beta}_g = \frac{\sum w_i y_i}{\sum w_i x_i} \tag{2}$$

où les poids dépendent du type d'estimateur. Pour l'estimateur GMH, par exemple, nous avons

$$w_i = \left(1 + \frac{|y_i - \hat{\beta}_g x_i|}{k}\right)^{-1}, \quad \text{pour } i = 1, \ldots, n. \tag{3}$$

L'équation (2) montre clairement qu'avec des EEVA, la détection et le traitement des valeurs aberrantes se font simultanément. Les poids servent à identifier les valeurs aberrantes: celles-ci ont un résidu élevé et sont affectées d'un petit poids. En ce qui concerne le traitement, les poids entrent dans le calcul de $\hat{\beta}_g$.

2.2 Résultats de la simulation

Dix situations sont analysées. Une situation est définie comme le "croisement" d'une taille d'échantillon (n = 20 ou 30) et d'une population (1 à 5). Les résultats de simulation pour ces dix situations figurent dans le tableau 2. On y retrouve deux statistiques: le biais relatif (c'est-à-dire le biais divisé par $\bar{Y}$, le paramètre d'intérêt) et le coefficient de variation (qui est égal au quotient de la racine carré de l'erreur quadratique moyenne par $\bar{Y}$).

Pour les cinq populations, les EEVA sont au moins aussi performants que l'estimateur par quotient. Dans le cas des populations 1 et 2, qui ne comptent pas d'unités à valeurs aberrantes, les EEVA sont aussi précis que

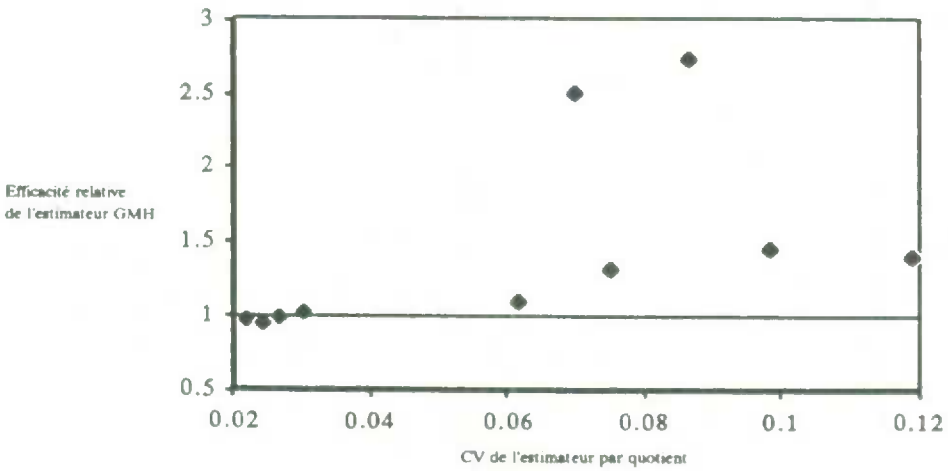
l'estimateur par quotient. Lorsque les populations renferment des unités à valeurs aberrantes, comme c'est le cas des populations 3, 4 et 5, on observe des gains de précision intéressants avec les EEVA. Ces résultats rejoignent ceux de Lee (1990), qui a montré que les EEVA donnaient des totaux plus exacts pour certaines enquêtes-entreprises. En ce qui concerne les populations 3 et 4, le biais relatif de EM et de GMM est presqu'égal au coefficient de variation correspondant, ce qui montre que la majeure partie de l'erreur quatratique moyenne est liée au biais. De tous les EEVA étudiés, c'est l'estimateur de la catégorie de Hampel-Krasker (GMH), fondé sur la fonction "fair", qui donne les meilleurs résultats au point de vue du biais et de l'erreur quadratique moyenne. Nous analysons plus en détail cet estimateur dans les sections qui suivent.

Tableau 2: Coefficient de variation (CV) et biais relatif (b. rel.) des quatre estimateurs du tableau 1 pour 5 populations et 2 tailles d'échantillon

Est.	Population 1		Population 2		Population 3		Population 4		Population 5	
n = 20	CV	b.rel	CV	b.rel.	CV	b.rel.	CV	b.rel.	CV	b.rel.
EQ	0.027	0.000	0.031	-0.002	0.087	-0.004	0.075	-0.001	0.119	0.002
EM	0.027	0.000	0.031	-0.009	0.054	-0.044	0.072	0.058	0.108	0.028
GMM	0.027	0.000	0.030	-0.011	0.054	-0.045	0.072	0.058	0.098	0.019
GMH	0.027	0.000	0.030	-0.011	0.052	-0.034	0.066	0.039	0.100	-0.005
Est.	Population 1		Population 2		Population 3		Population 4		Population 5	
n = 30	CV	b.rel.	CV	b.rel.	CV	b.rel.	CV	b.rel.	CV	b.rel.
EQ	0.022	0.000	0.024	-0.002	0.070	0.000	0.062	0.002	0.099	0.000
EM	0.023	-0.001	0.025	-0.009	0.049	-0.042	0.069	0.060	0.090	0.031
GMM	0.023	-0.001	0.025	-0.011	0.049	-0.043	0.069	0.061	0.080	0.022
GMH	0.022	-0.001	0.025	-0.011	0.044	-0.032	0.059	0.042	0.082	0.008

La figure 1 contient un graphique qui met en relation l'efficacité de l'estimateur GMH par rapport à l'estimateur par quotient et le coefficient de variation de ce dernier. Nous pouvons en déduire que les EEVA peuvent engendrer des gains d'efficacité intéressants lorsque le coefficient de variation de l'estimateur par quotient est supérieur à 5%.

Figure 1: Graphique de l'efficacité relative de l'estimateur GMH, définie comme le rapport de l'EQM de l'estimateur par quotient à l'EQM de l'EEVA, par rapport au coefficient de variation de l'estimateur par quotient



3. ESTIMATION DE L'ERREUR QUADRATIQUE MOYENNE DE L'ESTIMATEUR GMH

Gwet et Rivest (1990) ont examiné les propriétés (par rapport au plan de sondage) d'estimateurs à l'épreuve des valeurs aberrantes (EEVA) qui peuvent être substitués à l'estimateur par quotient. Nous allons nous servir ici des résultats de leur analyse pour calculer l'erreur quadratique moyenne asymptotique de l'estimateur GMH. Nous comparons plusieurs estimateurs de cette erreur quadratique moyenne au moyen d'une simulation de Monte Carlo. Dans cette section, $\hat{\beta}_g$ désigne l'estimateur-GM de la catégorie de Hampel-Krasker (GMH), fondé sur la fonction "fair", ψ_f. À compter de maintenant, cette fonction sera désignée simplement par ψ.

En règle générale, les EEVA ne sont pas convergents pour $\bar{Y}$. On peut exprimer leur biais asymptotique sous la forme $\bar{Y} - \bar{X}\beta_g$, où β_g est le paramètre de population estimé par β_g. Le paramètre β_g est défini comme la solution de l'équation suivante,

$$\frac{1}{n}\sum_U \psi\left(\frac{y_i - \beta x_i}{k}\right) = 0.$$

Une autre expression servant à définir le biais asymptotique est $\bar{X}(\beta_r - \beta_g)$ où β_r est le rapport de moyennes $\bar{Y} / \bar{X}$. Gwet et Rivest (1990) calculent la variance asymptotique de l'EEVA par des techniques de linéarisation. Dans le cas de l'estimateur GMH, un estimateur asymptotiquement non biaisé de la variance est

$$v(\bar{y}_g) = \frac{1-f}{n}\left(\frac{n\bar{X}}{\sum_s w_i^2 x_i}\right)^2 \frac{1}{n-1}\sum_s w_i^2 e_i^2$$

où e_i est le résidu, $y_i - \hat{\beta}_g x_i$, et w_i représente les poids définis par l'équation (3). Notons que lorsque tous les poids valent 1, $v(\bar{y}_g)$ équivaut à l'estimateur de la variance (v_2) pour l'estimateur par quotient. L'estimateur de l'erreur quadratique moyenne est défini par l'expression $v(\bar{y}_g) + b^2(\bar{y}_g)$, où $b(\bar{y}_g)$ est l'estimateur du biais $\bar{Y} - \bar{X}\beta_g$.

3.1 Estimation du carré du biais

L'estimateur fondamental du biais est $\hat{b}_0(\bar{y}_g) = \bar{y}_r - \bar{y}_g$, ou $\hat{b}_0(\bar{y}_g) = \bar{X}(\hat{\beta}_r - \hat{\beta}_g)$ où $\hat{\beta}_r = \bar{y}/\bar{x}$. Dans des estimateurs de l'erreur quadratique moyenne, le carré du biais est lui-même estimé. L'inconvénient d'un estimateur du carré du biais comme $\hat{b}_0(\bar{y}_g)^2$ est qu'il produit une surestimation du carré du biais réel. Soit l'équation

$$\bar{X}^2 E[(\hat{\beta}_g - \hat{\beta}_r)^2] = \bar{X}^2[(\beta_g - \beta_r)^2 + c_1 + c_2] \qquad (4)$$

où $c_1 = 2(\beta_g - \beta_r)(b(\hat{\beta}_g) - b(\hat{\beta}_r))$, $c_2 = V(\hat{\beta}_g - \hat{\beta}_r)$, et $b(\hat{\theta})$ est le biais de l'estimateur du paramètre de population θ. Dans beaucoup de cas, $c_1 + c_2$ est positif et $\hat{b}_0(\bar{y}_g)^2$ a un biais positif. Afin d'obtenir des estimateurs quasi non biaisés du carré du biais, nous avons examiné deux approches.

Premièrement, la correction de $b_0(\bar{y}_g)^2$ par la soustraction d'un estimateur asymptotique du biais. Les quantités c_1 et c_2 de l'équation (4) peuvent être estimées facilement. Par la linéarisation asymptotique pour EEVA décrite dans Gwet et Rivest (1990), nous obtenons l'estimateur suivant pour c_2:

$$\hat{c}_2 = \frac{1-f}{n(n-1)}\sum_s \left[\frac{y_i - \hat{\beta}_r x_i}{\bar{x}} - \frac{k\psi((y_i - \hat{\beta}_g x_i)/k)}{\sum_s w_i^2 x_i / n}\right]^2.$$

En ce qui concerne l'estimation de c_1, nous devons connaître l'estimateur du biais asymptotique de $\hat{\beta}_r$ et de $\hat{\beta}_g$.

Le biais asymptotique de $\hat{\beta}_r$ est défini par l'équation (Cochran, 1977, p. 161)

$$\hat{b}(\hat{\beta}_r) = -\frac{1-f}{n}\frac{(s_{xy} - \hat{\beta}_r^2 s_x^2)}{\bar{x}^2}$$

où s_{xy} est la covariance empirique de x et de y et s_x^2 est la variance empirique de x. Pour ce qui est du biais asymptotique de l'estimateur de $\hat{\beta}_g$, une expression appropriée a été construite dans l'annexe 1, soit

$$\hat{b}(\hat{\beta}_g) = -\frac{1-f}{n}\left(\frac{\Sigma_s x_i e_i w_i^3/n}{(\Sigma_s x_i w_i^2/n)^2} - \frac{1}{k}\frac{\Sigma_s x_i^2 sgn(e_i) w_i^3}{\Sigma_s x_i w_i^2}\, v(\bar{y}_g)\right).$$

En construisant des estimateurs du carré du biais, nous avons pensé qu'il était important que ces estimateurs ne soient jamais négatifs. Par conséquent, voici une version corrigée de l'estimateur du carré du biais, inspirée de l'équation (4),

$$\hat{b}_1(\bar{y}_g)^2 = \max(0,\ \hat{b}_0(\bar{y}_g)^2 - \bar{X}^2(\hat{c}_1 + \hat{c}_2)).$$

Le terme de variance, c_2, semblait représenter la majeure partie du biais; nous avons donc examiné une seconde version corrigée de l'estimateur du carré du biais:

$$\hat{b}_2(\bar{y}_g)^2 = \max(0,\ \hat{b}_0(\bar{y}_g)^2 - \bar{X}^2\hat{c}_2).$$

La seconde approche consistait à utiliser un estimateur jackknife du carré du biais. Cet estimateur est défini par l'équation

$$\hat{b}_j(\bar{y}_g)^2 = n\hat{b}_0(\bar{y}_g)^2 - \frac{n-1}{n}\sum_s \hat{b}_{(i)}(\bar{y}_g)^2,$$

où $\hat{b}_{(i)}(\bar{y}_g)$ désigne l'estimateur $\hat{b}_0(\bar{y}_g)$ calculé sans l'observation i. L'estimateur jackknife nécessite un très grand nombre de calculs; il faut résoudre n équations implicites afin de déterminer les n pseudo-valeurs nécessaires pour calculer une valeur de $\hat{b}_j(\bar{y}_g)^2$.

3.2 Résultats de la simulation de Monte Carlo

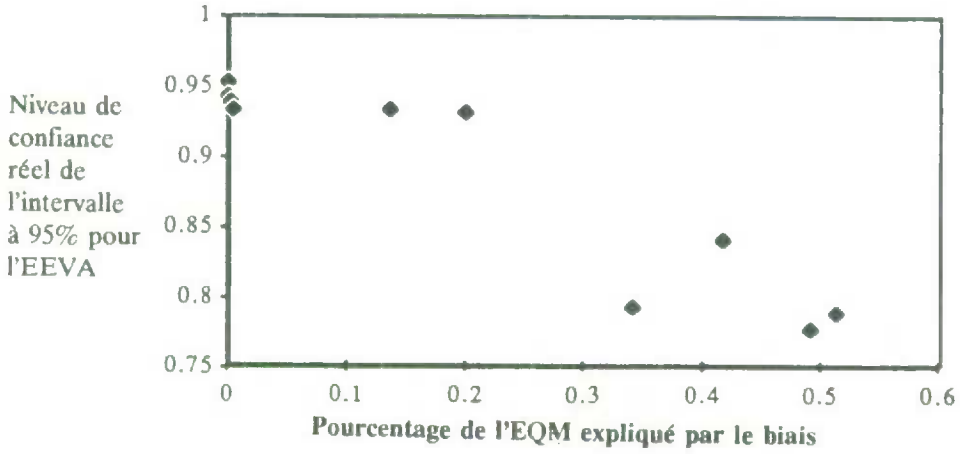
Le tableau 3 donne les résultats de la comparaison de quatre estimateurs de l'erreur quadratique moyenne, pour des EEVA, et de l'estimateur de la variance, v_2, pour l'estimateur par quotient (les quatre premiers sont formés par l'addition de l'estimateur de variance $v(\bar{y}_g)$ et de l'un ou l'autre des estimateurs $\hat{b}_0(\bar{y}_g)^2$, $\hat{b}_1(\bar{y}_g)^2\hat{b}_2(\bar{y}_g)^2$, $\hat{b}_j(\bar{y}_g)^2$ selon le cas). Ces résultats portent sur le biais relatif, le coefficient de variation et le niveau de confiance réel d'un intervalle à 95%.

Tableau 3: Biais relatif (exprimé en % du paramètre de la population cible; pour v_2, il s'agit de l'EQM de l'estimateur par quotient et pour les autres, de l'EQM de l'estimateur à l'épreuve des valeurs aberrantes), coefficient de variation (CV) et niveau de confiance réel (NC) d'un intervalle à 95% (pour v_2, l'intervalle de confiance est défini $\bar{y}_r \pm t_{.025}(v_2)^{1/2}$ et pour les autres, $\bar{y}_g \pm t_{.025}(v(\bar{y}_g) + b^2)^{1/2}$

Est. EQM	Population 1			Population 2			Population 3			Population 4			Population 5		
n=20	BR	CV	NC	BR	CV	NC	BR	CV	NC	BR	CV	NC	BR	CV	NC
v_2	−1	0.4	93	−8	0.46	93	−12	0.83	82	−8	0.59	83	−12	0.57	89
$v(\bar{y}_g)+b_1^2$	6	0.47	95	2	0.57	93	106	2.93	84	29	1.18	83	24	0.78	94
$v(\bar{y}_g)+b_2^2$	5	0.46	95	1	0.59	93	62	2.86	82	29	1.33	81	23	0.8	94
$v(\bar{y}_g)+b_0^2$	5	0.46	95	3	0.51	93	36	2.28	82	12	1.07	80	15	0.69	94
$v(\bar{y}_g)+b_j^2$	3	0.45	94	5	0.52	93	12	2.72	69	15	1.22	79	10	0.71	94
Est. EQM	Population 1			Population 2			Population 3			Population 4			Population 5		
n=30	BR	CV	NC	BR	CV	NC	BR	CV	NC	BR	CV	NC	BR	CV	NC
v_2	−4	0.31	95	−4	0.4	93	−8	0.71	85	−9	0.51	86	−13	0.5	88
$v(\bar{y}_g)+b_0^2$	0	0.32	95	3	1.04	93	111	2.92	84	22	1.07	82	21	0.67	94
$v(\bar{y}_g)+b_1^2$	−1	0.32	95	1	1.04	93	60	2.64	80	18	1.15	78	16	0.67	94
$v(\bar{y}_g)+b_2^2$	−1	0.28	95	0	0.94	92	43	2.28	79	5	0.97	78	10	0.58	93
$v(\bar{y}_g)+b_j^2$	−5	0.28	95	0	0.87	92	5	2.53	59	2	1.02	75	−1	0.6	92

Le tableau 3 indique que les estimateurs de l'EQM pour $\bar{y}_g$ ont un biais positif. Les trois variantes de l'estimateur fondamental du carré du biais, b_0^2, réduisent ce biais. Les coefficients de variation des estimateurs de l'EQM pour les EEVA sont assez élevés. Si on les compare au coefficient de variation de v_2, il faut se rappeler que, pour les populations 3, 4 et 5, v_2 sert à estimer un paramètre beaucoup plus grand que l'erreur quadratique moyenne des EEVA. Dans la population 3, les variances de v_2 et de $v(y_g) + b_2^2$ sont comparables; dans la population 5, v_2 a une variance plus élevée que $v(\bar{y}_g)+b_2^2$. Parmi les trois variantes de l'estimateur fondamental du carré du biais, b_2^2 est celui qui donne les meilleurs résultats. L'estimateur de l'EQM correspondant est celui qui, par rapport aux autres estimateurs de l'EQM pour $\bar{y}_g$, a un coefficient de variation et un biais peu élevés et produit des intervalles dont le niveau de confiance réel se rapproche de 95%. Dans le cas des populations 3 et 4, le niveau de confiance réel des intervalles à 95% pour l'EEVA laisse à désirer. Pour ces populations, l'EQM de $\bar{y}_g$ est inférieure à celle de $\bar{y}_r$; en revanche, $\bar{y}_r$ produit des intervalles dont le niveau de confiance réel se rapproche plus de 95%, surtout lorsque n=30. La figure 2 nous permet de mieux comprendre le phénomène. Elle décrit une relation étroite entre le niveau de confiance réel des intervalles pour l'EEVA et la proportion de l'EQM qui est expliquée par le biais. Le niveau de confiance réel est adéquat lorsque le carré du biais de $\bar{y}_g$ est faible (moins de 15% par ex.) par rapport à l'erreur quadratique moyenne.

Figure 2: Niveau de confiance réel de l'intervalle à 95% pour l'EEVA (calculé avec l'estimateur de biais $\hat{b}_2$) par rapport au pourcentage de l'EQM de l'EEVA expliqué par le biais



4. PROPRIÉTÉS CONDITIONNELLES DES EEVA

Le nombre réel d'unités à valeurs aberrantes dans un échantillon est un facteur déterminant pour l'efficacité des estimateurs. Par exemple, on peut penser que l'estimateur par quotient sera efficace lorsqu'il y a peu d'unités à valeurs aberrantes. Ces considérations ont amené les chercheurs à étudier les propriétés conditionnelles des estimateurs, étant donné un nombre déterminé d'unités à valeurs aberrantes dans l'échantillon (Hidiroglou et Srinath, 1981; Rao, 1985). Si on peut diviser des populations, comme les populations 4 et 5, en deux strates, l'une contenant des valeurs aberrantes (U_1) et l'autre non (U_2), il est utile de comparer des estimateurs lorsqu'un nombre déterminé d'unités à valeurs aberrantes sont échantillonnées. Cela peut se faire au moyen d'un échantillonnage stratifié de n_1 unités dans U_1 et n_2 unités dans $U_2 (n_1+n_2=n)$. Cette section renferme des analyses conditionnelles portant sur les populations de l'annexe 2 qui ont une strate d'unités à valeurs aberrantes facilement identifiable.

4.1 Biais et niveau de confiance conditionnels pour des échantillons de 20 unités tirés de la population 4

En ce qui concerne la population 4, la strate des valeurs aberrantes contient les 35 unités dont la pente est égale à 1. On reconnaît facilement cette strate dans le graphique de l'annexe 2. Le nombre prévu d'unités à valeurs aberrantes dans des échantillons de 20 unités tirés de la population 4 est 20•07% = 1.4. De tels échantillons devraient contenir de 0 à 4 unités à valeurs aberrantes; les probabilités inconditionnelles, calculées au moyen d'une approximation de Poisson de la distribution du nombre d'unités à valeurs aberrantes, sont indiquées dans le tableau 4. Pour chaque valeur de n_1, nous avons effectué entre 0 et 4, 2 000 simulations de Monte Carlo d'un échantillonnage aléatoire stratifié de n_1 unités dans U_1 et 20-n_1 unités dans U_2 pour calculer le biais conditionnel de $\bar{y}_g$ et $\bar{y}_r$ de même que le niveau de confiance réel conditionnel des intervalles pour $\bar{Y}$ obtenus au moyen de $\bar{y}_g$ et de $\bar{y}_r$. Nous nous sommes servis de l'estimateur de l'erreur quadratique moyenne $v(\bar{y}_g) + \hat{b}_2^2$ pour construire le premier intervalle et de l'estimateur de la variance v_2 pour construire le second.

Tableau 4: Biais relatif conditionnel (BR) et niveau de confiance réel conditionnel (NC) des intervalles à 95% construits avec les estimateurs $\bar{y}_g$ et $\bar{y}_r$ pour des échantillons de 20 unités tirés de la population 4

No. d'unités à valeurs aberrantes	0	1	2	3	4
Probabilité	.247	.352	.242	.113	.039
$BR(\bar{y}_g)$	.071	.046	.019	−.012	−.047
$BR(\bar{y}_r)$	.071	.019	−.029	−.075	−.118
$NC(\bar{y}_g)$	.311	.871	.999	1.000	1.000
$NC(\bar{y}_r)$	.280	.994	.999	.995	.925

Comme on peut le voir dans la case de la population 4 du tableau 2 pour n=20, le biais inconditionnel de $\bar{y}_g$ est 4%, ce qui est beaucoup plus que celui de $\bar{y}_r$, qui est à peu près nul. L'analyse conditionnelle du tableau 4 révèle quelque chose de tout à fait différent: en l'absence d'unités à valeurs aberrantes, $\bar{y}_r$ et $\bar{y}_g$ ont le même biais; lorsqu'il y a une unité (p=.352), le biais de $\bar{y}_g$ est plus élevé et lorsqu'on compte plus d'une unité (p=.394), c'est le biais de $\bar{y}_r$ qui est supérieur. Nous pouvons donc dire, du point du vue de l'analyse conditionnelle, que $\bar{y}_r$ est plus biaisé que $\bar{y}_g$. Les niveaux de confiance conditionnels des intervalles à 95% nous indiquent que les niveaux inconditionnels peu élevés observés dans la section précédente sont attribuables à la présence d'échantillons qui ne renferment pas d'unité à valeurs aberrantes et pour lesquels il y a surestimation de $\bar{Y}$ et sous-estimation de la variabilité des estimateurs.

4.2 Biais et niveau de confiance conditionnels pour des échantillons de 30 unités tirés de la population 5

Pour ce qui a trait à la population 5, U_1 est définie comme l'ensemble des unités auxquelles correspondent les 13 plus grands résidus de valeur absolue. On reconnaît facilement ces unités dans le graphique de l'annexe 2; ce sont celles dont la valeur y est inférieure à ce à quoi on pouvait s'attendre pour x. Le nombre prévu d'unités à valeurs aberrantes dans des échantillons de 30 unités est 1.1; de tels échantillons devraient contenir de 0 à 5 unités à valeurs aberrantes. Les probabilités inconditionnelles correspondantes, calculées à l'aide d'une approximation de Poisson, sont indiquées dans le tableau 5. Pour chaque valeur de n_1 entre 0 et 5, nous avons effectué 2,000 simulations de Monte Carlo pour calculer les biais et les niveaux de confiance conditionnels. Les résultats figurent dans le tableau 5.

Nous constatons que le biais conditionnel de $\bar{y}_g$ est toujours plus petit que celui de $\bar{y}_r$. De même, le niveau de confiance réel de l'intervalle pour l'EEVA est plus près du niveau théorique de 95% que ne l'est le niveau réel de l'intervalle pour l'estimateur par quotient. Dans cet exemple, l'EEVA produit des résultats plus précis que ceux obtenus par l'approche classique.

Tableau 5: Biais relatif conditionnel (BR) et niveau de confiance réel conditionnel (NC) des intervalles à 95% construits avec les estimateurs $\bar{y}_g$ et $\bar{y}_r$ pour des échantillons de 30 unités tirés de la population 5

Nombre d'unités à valeurs aberrantes	0	1	2	3	4	5
Probabilité	.190	.316	.262	.145	.060	.020
$BR(\bar{y}_g)$	.067	.031	-.005	-.044	-.091	-.131
$BR(\bar{y}_r)$	.093	.031	-.024	-.075	-.129	-.172
$NC(\bar{y}_g)$	.824	.949	.979	.978	.968	.912
$NC(\bar{y}_r)$	.640	.939	.932	.964	.907	.782

5. CONCLUSIONS

Les estimateurs à l'épreuve des valeurs aberrantes offrent une solution de remplacement intéressante vis-à-vis de l'estimateur par quotient lorsqu'un modèle de régression à ordonnée à l'origine nulle peut décrire avec justesse la plupart des unités d'une population. Le choix de la constante de robustesse k semble être un aspect déterminant de l'élaboration d'un EEVA. Dans notre analyse, ce choix s'est fait de manière à respecter des critères d'efficacité. Lorsqu'on recherche des EEVA pour une catégorie particulière de données, il est recommandé de tester plusieurs valeurs de k. Si on choisit une valeur trop petite, on risque d'obtenir un estimateur dont le biais serait trop élevé par rapport à la variance; les résultats de la section 3 donnent à penser que cette disproportion est à l'origine d'intervalles de confiance qui laissent à désirer. Si on choisit une valeur trop grande, on obtient l'estimateur par quotient habituel, caractérisé par un biais à peu près nul et une variance possiblement élevée.

BIBLIOGRAPHIE

Chambers, R.L. (1986). Outlier robust finite population estimation, *Journal of the American Statistical Association*, 81, 1063-1069.

Cochran, W.G. (1977). *Sampling Techniques. Third edition*. New York: John Wiley & Sons.

Ernest, L.R. (1980). Comparison of estimators of the mean which adjust for large observations, *Sankhya C*, 42, 1-16.

Fuller, W.A. (1970). Simple estimators for the mean of skewed populations. Technical Report. Iowa State University. Department of Statistics.

Gwet, J.-P., et Rivest, L.-P. (1990). Outlier resistant alternatives to the ratio estimator. Prépublication, Département de mathématiques et de statistique, Université Laval.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., et Stahel, W.E. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Hidiroglou, M.H., et Srinath, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.

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

Lee, H. (1990). Outlier resistant regression estimators. Présenté au 1990 Annual Meeting of the Statistical Society of Canada à St-John's, Newfoundland.

Rao, C.R. (1971). Some aspects of statistical inference in problems of sampling from finite populations in *Foundations of Statistical Inference*, Godambe and Sprott (Eds.). Toronto: Holt, Rinehart & Winston.

Rao, J.N.K. (1985). Conditional inferences in surveys sampling. *Survey Methodology*, 11, 15-31.

Rey, W.J.J. (1983). *Introduction to Robust and Quasi-Robust Statistical Methods.* New York: Springer Verlag.

Robinson, J. (1987). Conditioning ratio estimates under simple random sampling, *Journal of the American Statistical Association*, 82, 826-831.

Searl, D.T. (1966). An estimator which reduces large true observations, *Journal of the American Statistical Association*, 61, 1200-1204.

Tambay, J.L. (1988). An integrated approach for the treatment of outliers in sub-annual economic surveys, dans *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

ANNEXE 1: CALCUL DU BIAIS ASYMPTOTIQUE DE β_g.

Définissons $\hat{g}(\beta)$ comme suit:

$$\hat{g}(\beta) = \frac{1}{n} \sum_s \psi(\frac{y_i - \beta x_i}{k})$$

et posons $g(\beta)$ comme l'espérance de $\hat{g}(\beta)$. Par un développement de Taylor du deuxième ordre, nous obtenons l'expression suivante:

$$\hat{g}(\hat{\beta}g) - \hat{g}(\beta_g) = (\hat{\beta}_g - \beta_g)\hat{g}'(\beta_g) + \frac{1}{2}(\hat{\beta}_g - \beta_g)^2\hat{g}''(\beta_g) + o_p(1/n), \qquad (A.1)$$

où $\hat{g}'(\beta)$ et $\hat{g}''(\beta)$ sont les dérivées première et seconde par rapport à β de $\hat{g}(\beta)$. Par exemple,

$$\hat{g}'(\beta) = \frac{-1}{n} \sum_s \frac{x_i}{k} \psi'(\frac{y_i - \beta x_i}{k}).$$

En résolvant (A.1) en fonction de $\hat{\beta}_g - \beta_g$ et en choisissant la solution qui converge vers 0, nous obtenons

$$\hat{\beta}_g - \beta_g = \frac{-\hat{g}(\beta_g) - \sqrt{[\hat{g}'(\beta_g)]^2 - 2\hat{g}''(\beta_g)\hat{g}(\beta_g)}}{\hat{g}''(\beta_g)} + o_p(1/n).$$

Comme $\hat{g}(\beta_g)$ est $o_p(1/\sqrt{n})$ et que tous les autres termes sont $o_p(1)$, nous pouvons développer la racine carré du membre de droite de l'équation ci-dessus, ce qui donne

$$\hat{\beta}_g - \beta_g = -\frac{\hat{g}(\beta_g)}{\hat{g}'(\beta_g)} - \frac{\hat{g}''(\beta_g)[\hat{g}(\beta_g)]^2}{2[\hat{g}'(\beta_g)]^3} + o_p(1/n).$$

Considérons le membre de droite de cette équation. Dans le deuxième terme, $[\hat{g}(\beta_g)]^2$ est $o_p(1/n)$; par conséquent, jusqu'à $o_p(1/n)$, nous pouvons remplacer $\hat{g}''(\beta_g)$ et $\hat{g}'(\beta_g)$ par leurs espérances respectives, $g''(\beta_g)$ et $g'(\beta_g)$. Le premier terme du membre de gauche de l'équation est égal à

$$-\frac{\hat{g}(\beta_g)}{g'(\beta_g)} \quad \frac{1}{1 + (\hat{g}'(\beta_g) - g'(\beta_g))/g'(\beta_g)} \qquad (A.2)$$

Comme $(\hat{g}'(\beta_g) - g'(\beta_g))/g'(\beta_g))$ est $o_p(1/\sqrt{n})$, nous pouvons obtenir une approximation $o_p(1/n)$ de (A.2) en utilisant un développement de Taylor à un terme. En introduisant cette approximation dans la formule du développement de $\hat{\beta}_g - \beta_g$, nous obtenons

$$\hat{\beta}_g - \beta_g = -\frac{\hat{g}(\beta_g)}{g'(\beta_g)} + \frac{\hat{g}(\beta_g)(\hat{g}'(\beta_g) - g'(\beta_g))}{[g'(\beta_g)]^2} - \frac{g''(\beta_g)[\hat{g}(\beta_g)]^2}{2[g'(\beta_g)]^2} + o_p(1/n).$$

Ainsi, jusqu'à $o(1/n)$, l'espérance de $\hat{\beta}_g$ est égale à

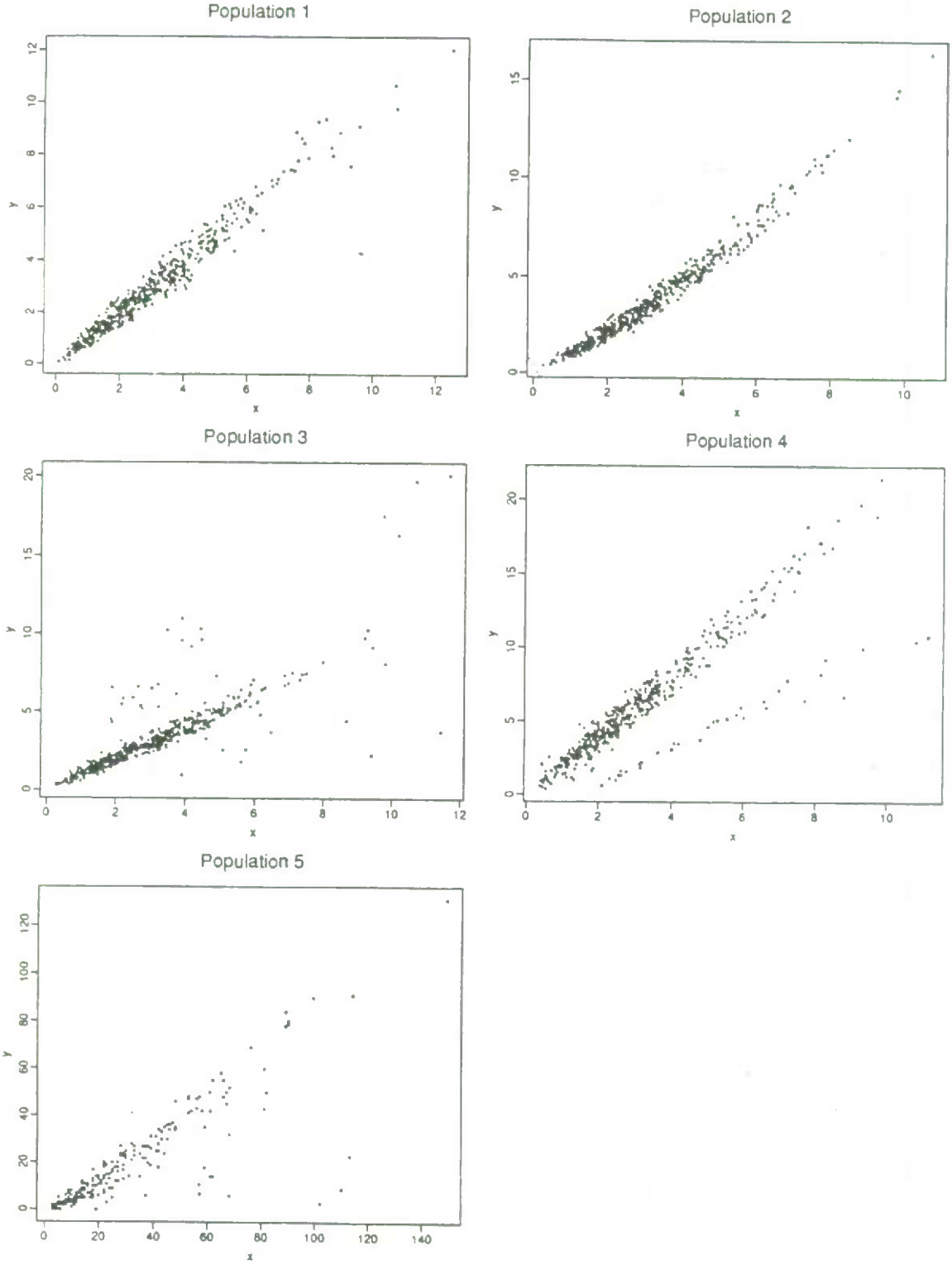
$$\beta_g + \frac{cov(\hat{g}(\beta_g),\hat{g}'(\beta_g))}{[g'(\beta_g))^2} - \frac{g''(\beta_g)(\hat{g}(\beta_g))^2}{2g'(\beta_g)[g'(\beta_g)]^2}.$$

Pour ce qui a trait à l'estimateur GMH, les deux premières dérivées de g peuvent être estimées par

$$\hat{g}'(\hat{\beta}_g) = \frac{-1}{n}\sum_s \frac{x_i}{k} w_i^2 \quad and \quad \hat{g}''(\hat{\beta}_g) = \frac{-1}{n}\sum_s \left(\frac{s_i}{k}\right)^2 sgn(ei)\, w_i^3.$$

Cela nous amène à l'équation (5), qui définit l'estimateur du biais asymptotique de $\hat{\beta}_g$.

Annexe 2. Populations utilisées pour l'étude de Monte Carlo



SESSION 9

Nos produits statistiques sont-ils utilisables?

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

LE PASSAGE DES STATISTIQUES DESCRIPTIVES À L'INFÉRENCE

K. O'Conor, B.K. Atrostic et R. Gillette[1]

RÉSUMÉ

Les organismes statistiques d'État produisent principalement des statistiques descriptives. Cependant, avec l'évolution des besoins des utilisateurs, ils ont modifié leurs produits. Ils ont commencé à changer leurs plans de sondage, leurs fichiers à grande diffusion et leurs publications.

Dans cette communication, nous soutenons que ces organismes devraient avoir parmi leurs buts celui de produire des ensembles de données qui permettent aux utilisateurs de mettre à profit, dans l'analyse des données qu'ils leurs fournissent, tout le potentiel de la statistique moderne, au lieu de produire simplement une plus grande quantité de statistiques descriptives. Il est clair qu'il faut également permettre à davantage de gens d'avoir accès aux données descriptives existantes. Celles-ci devraient, notamment, être distribuées de diverses façons: sur disquette, disque optique ou panneau d'affichage électronique. Il est nécessaire que les fournisseurs répondent à la demande des utilisateurs en leur fournissant, avec le mode d'emploi, un mélange équilibré de statistiques descriptives et inférentielles de grande qualité.

MOTS CLÉS: Microdonnées; plans de sondage; fichiers à grande diffusion.

1. INTRODUCTION

Au cours des dernières années, la révolution en matière de qualité amorcée par Deming (1986) et Juran (1988) a eu une incidence considérable sur la production des statistiques fédérales. Tant aux États-Unis qu'au Canada, des améliorations importantes sont apportées dans tous les domaines des méthodes statistiques. Le présent article décrit quatre grandes enquêtes fédérales réalisées aux États-Unis et fait état des efforts déployés dans le cadre de ces enquêtes pour améliorer la qualité et l'aptitude à l'utilisation des fichiers à grande diffusion, particulièrement aux fins des travaux de recherche en matière de politique fiscale.

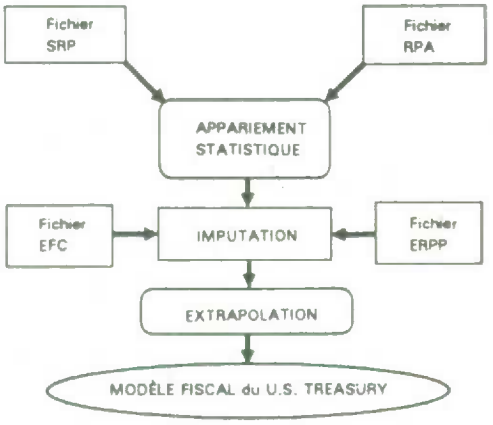
Chacun des programmes étudiés dans cet article a adopté une approche différente à l'égard de la gestion de la qualité. Ainsi, la Statistics of Income Division (SOI) du Internal Revenue Service a récemment mis en oeuvre un processus d'interaction entre les clients et les fournisseurs de données en vue de remanier son fichier de statistiques sur le revenu des particuliers (SRP), établi à partir d'un échantillon de déclarations d'impôt sur le revenu. Dans une autre veine, l'équipe de la Current Population Survey [recensement de la population actuelle (RPA)], réalisée par le U.S. Census Bureau, a calculé et diffusé des coefficients de pondération répétés, qui permettent de calculer l'erreur d'échantillonnage presque tous les types d'estimations. De son côté, l'équipe de la Survey of Income and Program Participation [enquête sur le revenu et la participation aux programmes (ERPP)] a mis sur pied une base interactive de données relationnelles afin d'être mieux en mesure de répondre aux multiples besoins des utilisateurs. Enfin, l'équipe de la Survey of Consumer Finances [enquête sur les finances des consommateurs (EFC)], enquête mise en oeuvre par le U.S. Board of Governors of the Federal Reserve, a utilisé la relaxation stochastique pour corriger les données de l'effet de la non-réponse à certaines questions.

---

[1] K. O'Conor, Internal Revenue Service, P.O. Box 2608, Washington D.C., U.S.A. 20013, B.K. Atrostic et R. Gillette, Office of Tax Analysis.

Les efforts déployés dans le cadre de ces enquêtes ne sont que quelques exemples des travaux en cours dans le secteur fédéral américain. Nous avons retenu ces exemples parce que nous nous en inspirons pour élaborer un des modèles de microsimulation utilisés par le U.S. Treasury Department aux fins des travaux de recherche en matière de politique fiscale (le modèle fiscal du U.S. Treasury) (voir la figure 1). Après avoir décrit ce modèle pour donner au lecteur un bref aperçu d'un exemple d'utilisation intensive des fichiers à grande diffusion et une idée générale des besoins des modélisateurs, nous décrivons aux sections suivantes les échantillons des quatre enquêtes sur lesquelles se fonde le modèle ainsi que les efforts déployés dans le cadre de ces enquêtes pour améliorer la qualité et l'aptitude à l'utilisation des fichiers à grande diffusion, afin de permettre la construction de meilleurs modèles pour l'élaboration de la politique fiscale.

Figure 1 - Modèle fiscal du U.S. Department of the Treasury



1.1 Brève description du modèle fiscal du U.S. Treasury

Au Canada et aux États-Unis, la politique fiscale est élaborée à l'aide de modèles de microsimulation qui permettent de combiner les données de l'administration fiscale avec les données d'enquête et d'autres renseignements. Les données de la base de données résultante sont ensuite redressées en fonction des projections économiques et démographiques relatives aux années futures afin de calculer les effets sur le revenu et les conséquences sur la répartition des richesses des modifications qu'on propose d'apporter au droit fiscal. La précision de ces estimations varie en fonction de la justesse et de l'accessibilité des données complémentaires. Le défi consiste donc à élargir l'éventail et à accroître la justesse des données fournies aux stratèges, afin qu'ils puissent prendre des décisions plus éclairées. À cette fin, il est nécessaire de disposer de meilleures données, de faire une meilleure utilisation de ces données et de veiller à ce que les producteurs de données comprennent mieux comment les utilisateurs prennent réellement leurs décisions.

L'élaboration du modèle fiscal du U.S. Treasury s'effectue en deux étapes: une phase d'appariement et d'imputation suivie d'une phase d'extrapolation. La première étape consiste à apparier ou à coupler les particuliers statistiquement «similaires» (c.-à-d., appariement statistique) - plutôt que de procéder au couplage des enregistrements relatifs à des particuliers identiques - en utilisant l'âge et le revenu comme variables déterminantes. L'appariement est conditionnel au maintien des distributions marginales des deux populations (Barr et coll., 1987). Une fois l'appariement terminé, on utilise les résultats du RPA, de l'enquête sur le revenu et la participation aux programmes (ERPP) et de l'enquête sur les finances des consommateurs (EFC) à des fins d'imputation. Après l'étape de l'appariement et de l'imputation, on établit par extrapolation des prévisions permettant de brosser un tableau de la conjoncture économique pour l'administration. Le lecteur trouvera un aperçu général du modèle fiscal du U.S. Treasury dans Cilke et Wyscarver (1987) et un exposé plus détaillé sur le même modèle dans Cilke et Wyscarver (1990).

Au Canada, les modèles comme celui de Wolfson permettent un appariement statistique plus précis que le modèle fiscal du U.S. Treasury, en partie du fait que la structure de l'appareil statistique canadien facilite les approches du genre. En particulier, les organismes fédéraux canadiens sont plus libres d'échanger des données entre eux.

La comparabilité et la fiabilité des mesures du revenu établies par les diverses enquêtes sont des facteurs déterminants pour le choix des enquêtes devant être utilisées pour construire le modèle fiscal. Comme le fichier de statistiques sur le revenu des particuliers renferme peu de renseignements autres que certains éléments d'information sur le revenu, ce dernier doit constituer une des variables clés pour tout processus d'appariement statistique ou d'imputation entre ce fichier et une autre enquête. Bien que les définitions du revenu utilisées pour le RPA, l'ERPP et l'EFC varient légèrement, il existe entre elles de nombreux recoupements.

De fait, une grande partie des travaux que nous nous proposons d'étudier dans cet article ont eu pour résultat d'améliorer la comparabilité des données sur le revenu. Vous trouverez aux sections suivantes une brève description des systèmes d'enquête utilisés pour élaborer le modèle fiscal du U.S. Treasury et un exposé de certains des projets innovateurs mis en oeuvre par l'équipe de chacune des enquêtes choisies pour accroître l'aptitude à l'utilisation des données.

2. FICHIER DE STATISTIQUES SUR LE REVENU DES PARTICULIERS

2.1 L'échantillon du fichier de statistiques sur le revenu des particuliers

Le U.S. Internal Revenue Service a commencé à élaborer le fichier de statistiques sur le revenu des particuliers en 1918 et a produit son premier fichier à grande diffusion en 1960. L'échantillon à partir duquel le fichier est actuellement établi comprend alternativement 80 000 et 120 000 déclarations d'impôt sur le revenu. Suivant l'actuel plan d'échantillonnage, les données recueillies sur les formules 1040, 1040A et 1040EZ sont stratifiées selon le montant le plus élevé du revenu net total ou des pertes nettes totales et selon la somme du revenu d'entreprise et du revenu d'agriculture. En outre, les strates sont fondées sur la présence ou l'absence d'un revenu de source étrangère (formule 2555), d'un crédit pour impôt étranger (formule 1116), de bénéfices ou de pertes provenant d'une entreprise ou d'une profession (annexe C) et d'un revenu et de dépenses d'agriculture (annexe F). À compter de juin 1991, les données seront recueillies à l'aide d'un nouveau plan d'échantillonnage, dont vous trouverez une brève description à la section 2.2.

Tant le plan actuel que le nouveau prévoient deux méthodes de prélèvement des déclarations au sein de chaque strate. La première méthode utilise certains des derniers chiffres du numéro de sécurité sociale, tandis que la seconde utilise les derniers chiffres de nombres aléatoires obtenus à la suite de transformations des numéros de sécurité sociale. Ce processus en deux étapes a pour objet d'assurer un certain chevauchement entre l'échantillon du fichier de statistiques sur le revenu des particuliers et le Continuous Work History Sample de l'administration de la sécurité sociale, un des panels longitudinaux les plus anciens au monde, utilisé pour recueillir des données sur les gains des employés couverts par le programme de sécurité sociale. Le chevauchement intervient pour environ 10,000 déclarations lorsque l'échantillon en comprend 80,000 (ou pour 20,000 lorsque l'échantillon en comprend 120,000). Selon le nouveau plan d'échantillonnage, il interviendra pour environ 20,000 déclarations (Smith et coll., 1989).

2.2 Remaniement de l'échantillon transversal

2.2.1 Données de base

Comme nous l'avons vu, la Statistics of Income (SOI) Division du Internal Revenue Service procède actuellement à un remaniement de son échantillon de déclarations d'impôt sur le revenu des particuliers afin de fournir aux modélisateurs de meilleures données pour estimer les effets des modifications qu'on propose d'apporter à la politique fiscale. L'échantillon remanié comporte trois composantes de base:

1. un panel longitudinal d'environ 83 000 déclarations utilisant 1987 comme année de base;

2. les déclarations des personnes à la charge des familles sélectionnées, composante ayant pour objet de permettre la définition d'un «impôt familial»; enfin,

3. un échantillon transversal annuel remanié.

Ces trois composantes seront sélectionnées et traitées chaque année. La première, à savoir le panel longitudinal, permettra de mesurer les changements réels dans les composantes du revenu déclaré en se reportant aux déclarations produites par les particuliers au fil des ans plutôt qu'en ayant recours aux comparaisons transversales répétées actuellement utilisées. Ce panel constituera une base de sondage dans laquelle on pourra sélectionner rapidement des panels spécialisés. La deuxième composante, à savoir les déclarations des personnes à charge prélevées pour définir un impôt familial, permettra de mesurer le revenu de familles réelles plutôt que celui des familles synthétiques auparavant formées par appariement statistique du fichier de statistiques sur le revenu des particuliers avec d'autres fichiers. En outre, cette composante facilitera la conciliation du fichier de statistiques sur le revenu des particuliers avec les autres sources de données. La troisième composante, à savoir le nouvel échantillon transversal annuel remanié, permettra, d'une part, de rationaliser l'échantillon de composantes du revenu visées par la politique fiscale et, d'autre part, d'assurer une meilleure couverture de certains groupes démographiques (Hostetter et coll., 1990; Czajka et coll., 1990).

Ce remaniement s'est révélé être une tâche de longue haleine: sa planification a débuté il y a deux ans et sa mise en oeuvre intégrale est prévue pour 1991. Le panel longitudinal a déjà permis de recueillir des données pour les années 1987, 1988 et 1989. Nous procédons actuellement à l'élaboration et à l'examen de définitions qui devraient nous permettre de produire des données provisoires sur les familles pour les premiers cas à l'étude à la fin de 1991. Enfin, le système de sélection de l'échantillon transversal est en cours de programmation et on pourra l'utiliser pour prélever un premier échantillon en juin 1991 (Bates, 1991). La suite de cette section est consacrée à l'étude du processus d'interaction qui a donné au remaniement son caractère exceptionnel.

2.2.2 Participants et rôles

En 1987, un comité formé d'utilisateurs et de producteurs de données s'est réuni pour définir les objectifs généraux du remaniement, déterminer les composantes et les besoins du projet, ainsi qu'élaborer un cadre d'entente. Ce comité était formé de représentants de l'Office of Tax Analysis du Treasury Department (principaux utilisateurs des données), du secteur des systèmes informatiques de l'IRS, des centres de service de l'IRS (chargés du contrôle et du codage des statistiques sur le revenu) et de la Statistics of Income Division (chargée de la gestion de l'enquête), ainsi que d'experts indépendants. Sa réalisation de loin la plus importante a été l'élaboration d'un cadre d'entente et l'atteinte d'un consensus quant aux objectifs visés. Comme il arrive souvent en pareilles circonstances, ni les utilisateurs ni les producteurs de données n'avaient jusqu'alors pris le temps nécessaire pour tenter de comprendre les besoins de l'autre partie. (Pour plus de renseignements sur l'issue de ce processus, se reporter au mémoire que Susan Hostetter et Karen O'Conor présenteront aux 1991 Joint Statistical Meetings d'Atlanta.)

L'étape suivante de la planification, entreprise à l'automne 1989, a consisté à recueillir et à évaluer des idées en vue du remaniement de l'échantillon transversal annuel ainsi qu'à définir le processus de traitement des données recueillies auprès des membres du panel et des données sur les familles. Comme les travaux de recherche relatifs au remaniement nécessitaient une interaction plus étroite entre les parties, ils ont donné lieu à la mise sur pied d'une nouvelle équipe. Cette équipe, dont le mandat était beaucoup plus étroitement défini que celui du comité de 1987, était formée de trois ou quatre utilisateurs de données de l'Office of Tax Analysis (OTA), de deux spécialistes de l'échantillonnage du Mathematical Policy Research (MPR) et d'un groupe d'économistes et de spécialistes de la statistique mathématique de la SOI Division.

2.2.3 Processus interactif de remaniement

Afin d'atteindre les multiples objectifs concurrents qui lui étaient fixés, le comité de remaniement devait résoudre de nombreuses questions, dont la plus difficile, peut-être, consistait à élaborer une définition des déclarations portant sur un revenu à structure complexe, soit des déclarations les plus susceptibles d'être visées par les propositions de modification de la politique fiscale. En raison de leur relative rareté, les déclarations complexes sont susceptibles d'être sous-représentées au sein de l'échantillon. Toutefois, comme certaines caractéristiques relativement rares des déclarants présentent un intérêt pour l'élaboration de la politique, il était nécessaire de les isoler afin de pouvoir sélectionner un nombre suffisant de déclarations y afférentes.

Après de longues discussions, les membres du comité en sont arrivés à un consensus initial en vertu duquel on tenait la déclaration de l'une ou l'autre des 10 composantes suivantes du revenu comme un indice possible de

complexité et on déterminait 3 indicateurs du statut du déclarant permettant de repérer les déclarations susceptibles d'être sous-représentées dans l'échantillon.

Indices de revenu à structure complexe

- Gains ou pertes en capital
- Revenu ou pertes d'une société de personnes ou d'une petite société
- Déductions détaillées (annexe A)
- Déduction pour intérêt hypothécaire
- Prestations de sécurité sociale
- Prestations de pension ou revenu de rentes
- Crédit pour frais de garde d'enfants
- Prestations d'assurance-chômage
- Pension alimentaire
- Impôt minimum de remplacement

Indicateurs du statut du déclarant

- Exemption en raison de l'âge
- Chef de ménage non marié
- Exemption pour enfant à charge vivant à la maison ou pour parents à charge

Ensuite, le MPR a élaboré des indices permettant de déterminer la fréquence d'occurrence des diverses combinaisons possibles d'éléments connexes. Cette recherche a porté sur les déclarations relatives aux tranches inférieures de revenu positif, puisque le plan d'échantillonnage prévoyait déjà la sélection d'une proportion élevée des déclarations faisant état d'un revenu de plus de $250,000. Il a été démontré à l'aide d'une série de tableaux que les déclarations complexes portant sur des revenus de moins de $250,000 étaient réparties de façon relativement égale entre les autres indicateurs.

Deux autres objectifs venaient en conflit avec l'objectif de sélection de «déclarations complexes». Le premier de ces objectifs était de sélectionner un nombre de déclarations non complexes assez élevé pour assurer une couverture suffisante aux fins de la modélisation et de l'élaboration de statistiques descriptives. Le deuxième était de maintenir le plus haut degré de simplicité possible, en utilisant uniquement deux niveaux de stratification: le revenu et le type de formule. Cette considération est importante quand on sait que les phases d'appariement statistique, d'imputation et d'extrapolation rendent déjà la construction du modèle fiscal du U.S. Treasury très complexe - les modélisateurs voulaient également que le plan d'échantillonnage soit le plus simple possible, avec uniquement deux niveaux de stratification: le revenu et le type de formule. L'élaboration de plans d'échantillonnage d'essais a permis de maintenir un certain équilibre entre ces objectifs concurrents.

L'équipe a aussi déterminé qu'il importait de définir les déclarations *non* complexes en se fondant sur le revenu et sur d'autres caractéristiques *peu* susceptibles d'être visées par les propositions de modification de la politique fiscale. Il a fallu prélever plusieurs échantillons avant de pouvoir classer les déclarations d'une façon qui réponde aux besoins des chercheurs en matière de politique fiscale et assure le maintien de la fiabilité des estimations publiées sur le revenu des particuliers.

Une proposition initiale définissait les déclarations susceptibles d'être sous-représentées dans l'échantillon comme celles où une proportion substantielle du revenu positif total (75% dans le cas des déclarations portant sur un revenu positif allant de $60 000 à $250 000 et 90% dans celui des déclarations portant sur un revenu positif allant de $0 à $60 000) provenait des traitements et salaires ou du revenu de retraite (voir la figure 2). Cette définition s'est révélée trop restrictive, du fait qu'elle permettait de sélectionner trop peu de déclarations dans les 13 principales catégories. On a alors élaboré une définition élargie englobant trois autres types de déclarations non complexes: les déclarations faisant état d'éléments déduits du revenu au titre de l'impôt minimum de remplacement mais n'indiquant pas d'impôt minimum de remplacement; les déclarations classées comme complexes uniquement parce qu'elles font état d'un substantiel revenu d'intérêt exempt d'impôt; enfin, les déclarations faisant état d'un revenu constitué en majeure partie d'un revenu d'entreprise individuelle (annexe C) et d'un revenu en intérêts et en dividendes.

Cependant, la définition élargie ne permettait de résoudre une difficulté qu'en en soulevant une autre. Si on s'en était tenu à cette définition, le nombre de déclarations faisant état d'un revenu d'entreprise individuelle sélectionnées dans l'échantillon aurait été réduit d'environ 2 300, concentrées dans les tranches de revenu allant de $0 à $30 000 et de $30 000 à $60 000. L'échantillon résultant n'aurait pas permis d'assurer une stabilité suffisante des estimations du revenu d'entreprise individuelle net ou brut total. Comme la majeure partie du revenu net ou brut total déclaré au sein de la population l'est par les unités appartenant aux strates de faibles revenus positifs, qui auraient été sous-représentées dans l'échantillon, une telle modification aurait été inacceptable pour nos utilisateurs. Le Bureau of Economic Analysis (BEA) utilise ces données pour établir des estimations des comptes du revenu national et du produit national.

En conséquence, on a complété les deux premières définitions en les assortissants d'une troisième condition: les déclarations non complexes devaient être reclassées comme complexes si le revenu négatif total intervenait pour plus de 40% du revenu positif total. Grâce à cette mesure, on estime que 520 déclarations additionnelles faisant état d'un revenu d'entreprise individuelle ont été sélectionnées dans l'échantillon pour la tranche de revenu allant de $0 à $60 000 (Czajka, 1988; Hostetter, 1990).

2.3 Avantages

Des totalisations d'essai ont été exécutées pour déterminer l'incidence sur les données-échantillon de chacune des contraintes de classification relatives au revenu et au type de formules. Ces totalisations nous ont permis de recueillir des renseignements utiles en vue des décisions ultérieures. De plus, comme l'échantillon final devait répondre aux besoins de plusieurs groupes ayant des intérêts divergents, les études de simulation nous ont permis d'établir une base plus factuelle en vue de l'atteinte d'une solution de compromis. À l'évidence, le processus de négociation qui a suivi a permis à tous les participants de mieux comprendre les besoins des autres parties et a aidé l'équipe à se concentrer sur les faits plutôt que de se cantonner dans la critique des opinions. Lors de l'élaboration par le comité d'une définition opérationnelle de la complexité, des divergences de vues se sont fait jour quant à la valeur de diverses totalisations, du fait du coût assez élevé de certaines d'entre elles. D'ordinaire, on effectue peu de ces totalisations en raison du coût élevé des calculs y afférents. Cependant, toutes les totalisations étudiées ont été effectuées et le résultat obtenu a justifié chacun des sous investis car il a permis d'asseoir les décisions prises sur une base factuelle. Ce processus d'interaction s'est soldé par la mise au point d'un plan d'échantillonnage qui répond tant aux besoins des constructeurs de modèles fiscaux (OTA) qu'à ceux des utilisateurs de statistiques descriptives (BEA et IRS).

3. RECENSEMENT DE LA POPULATION ACTUELLE

3.1 L'échantillon du recensement de la population actuelle

Le Treasury Department utilise les RPA et d'autres sources pour obtenir les données démographiques et certaines données financières ne figurant pas dans le fichier de statistiques sur le revenu des particuliers. Réalisée pour la première fois en 1942 lorsque la responsabilité de la tenue de la Survey of Unemployment (enquête sur le chômage) a été transférée au Bureau of the Census, le RPA est peut-être la mieux connue des enquêtes dont les fichiers sont étudiés dans cet article. Le premier fichier à grande diffusion du RPA a été produit en 1968.

Enquête-ménage mensuelle réalisée par le U.S. Census Bureau pour le compte du Bureau of Labor Statistics, le RPA permet de recueillir des données sur un large éventail de sujets. Bien qu'il ait été initialement conçu pour recueillir des données démographiques et des données sur la population active, le RPA recueille maintenant aussi (surtout par l'intermédiaire d'un supplément à l'enquête annuelle de mars) des données sur des sujets comme les heures travaillées, la profession, la branche d'activité, le revenu périodique, le revenu personnel et le revenu familial, la migration, la scolarité, etc. C'est le fichier du RPA de mars qu'on utilise pour élaborer le modèle fiscal du U.S. Treasury.

Le recensement de la population actuelle porte sur l'ensemble de la population civile des États-Unis, à l'exclusion des pensionnaires d'établissements institutionnels. L'échantillon du RPA, qui compte 60 000 ménages ou environ 113 000 personnes de 16 ans ou plus, est un échantillon en grappes d'unités de logement stratifié

géographiquement. Il s'agit en outre d'un échantillon avec renouvellement selon lequel chaque ménage est interviewé pendant quatre mois, non interviewé pendant huit mois, interviewé pendant quatre autres mois, puis supprimé de l'enquête (Bureau of the Census, 1978). Pour les fins du présent article, nous nous intéressons surtout aux coefficients de pondération répétés élaborés pour le RPA afin de permettre aux chercheurs de calculer la variance d'un large éventail d'estimations.

3.2 Fonctions de variance généralisée

Dans le passé, le RPA utilisait des fonctions de variance généralisée pour calculer les erreurs d'échantillonnage. Cette pratique présentait plusieurs inconvénients. Ainsi, le Bureau of Labor Statistics (BLS), qui est particulièrement intéressé à utiliser les données du RPA pour estimer le taux de chômage dans chaque État, devait utiliser cette approximation pour calculer une mesure de la fiabilité du taux estimé pour chaque État. Ce faisant, il était aux prises avec deux sérieuses difficultés:

1) l'effet national du plan de sondage à l'intérieur de l'unité primaire d'échantillonnage est supposé constant d'un État à l'autre, puis corrigé pour tenir compte des unités d'échantillonnage dont on a déterminé qu'elles ne faisaient pas partie du champ d'observation de l'enquête; enfin,

2) le rapport de la variance entre les unités primaires d'échantillonnage à la variance totale est supposé constant dans le temps (Lent, 1991).

Le BLS utilise maintenant un fichier de coefficients de pondération répétés élaborés par le Census Bureau pour calculer des estimations de la variance et de la corrélation pour chaque État.

Une autre difficulté soulevée par l'ancienne procédure découlait du fait que, pour des raisons de confidentialité, la majorité des utilisateurs des données du RPA ne pouvaient obtenir les renseignements relatifs à la stratification (géographique) nécessaires pour calculer leurs propres estimations de la variance. Ces utilisateurs devaient se rabattre sur les tables de la variance généralisée, qui ne sont utiles que dans le cas des pourcentages et des totaux. Il leur était souvent impossible de calculer la variance des estimateurs très complexes.

La nouvelle technique générale de répétition tente de surmonter ces deux difficultés. Elle a permis au Census Bureau de fournir au BLS un fichier de coefficients de pondération répétés pour calculer les estimations de la variance et de la corrélation par État, tout en assurant le maintien d'une souplesse suffisante pour répondre aux besoins des petits utilisateurs, et ce sans compromettre la confidentialité des données. Le Census Bureau continuera de calculer la fonction de variance généralisée et de fournir les tables correspondantes aux utilisateurs qui préfèrent s'en servir.

3.3 Aperçu de la théorie générale de la répétition

Selon la nouvelle approche, l'équipe du RPA utilise la théorie générale de la répétition élaborée par Robert Fay pour calculer les coefficients de pondération répétés (Fay, 1984 et 1989; Wolter, 1985). Cette technique est très semblable à la méthode BRR (Balanced Repeated Replication), procédure permettant de calculer des estimations répétées à partir d'un demi-échantillon, de telle sorte que chaque estimation tienne compte de la moitié des observations relatives à chaque strate. Selon cette procédure, chaque observation fait partie de la moitié des échantillons.

Suivant la théorie générale de la répétition, toutes les observations sont utilisées à chaque répétition et on applique des facteurs de répétition pour déterminer la contribution de chaque demi-échantillon au coefficient de pondération répété. Ainsi, dans le cas de la méthode BRR classique, les facteurs de répétition pour un demi-échantillon prélevé dans une strate sont soit 0 ou 2; dans le cas de la théorie générale de la répétition, ces facteurs pourraient être égaux à 1/2 ou à 1 1/2. Comme le RPA ne compte qu'une unité primaire d'échantillonnage par strate, on crée des pseudo-strates en regroupant les strates d'échantillonnage. Certaines des pseudo-strates créées par le système de répétition du RPA sont des portions de strates autoreprésentatives plutôt que des groupes de strates regroupées. La théorie générale de la répétition utilise une méthode permettant d'attribuer des facteurs qui varient en fonction des pseudo-strates.

Selon des essais de simulation réalisés pour évaluer l'estimateur de la variance généralisée de Fay, ce dernier est utile lorsqu'il faut établir des estimations de la variance pour des statistiques lissées et non lissées ou lorsqu'on ne dispose que de très peu de degrés de liberté pour l'estimation de la variance (Judkins, 1990). Or, il s'agit précisément du genres d'estimations que de nombreux utilisateurs des données du RPA établissent.

3.4 Calcul de l'incertitude à l'aide des coefficients de pondération répétés

Les coefficients de pondération répétés ainsi obtenus ont facilité le calcul de l'incertitude. Il existe 48 coefficients de pondération répétés pour chaque enregistrement. Les estimations sont calculées à l'aide du coefficient de pondération répété plutôt que du coefficient de pondération du ménage, qui était auparavant la seule mesure disponible. On calcule une estimation pour chaque échantillon répété et on obtient la variance en multipliant la différence par un facteur de 4, afin de tenir compte du fait qu'on a utilisé l'ensemble de l'échantillon pour chaque estimation.

3.5 Avantages

Cette méthode permet au chercheur de calculer la variance presque toute estimation et tout sous-ensemble de la population. Bien que le calcul des estimations de la variance pour de petits sous-ensembles géographiques pose certains problèmes (on obtient pour les petits États un coefficient de variation de la variance aussi élevé que 50%), ces estimations représentent une amélioration considérable par rapport aux variances généralisées.

4. ENQUÊTE SUR LE REVENU ET LA PARTICIPATION AUX PROGRAMMES (ERPP)

4.1 L'échantillon de l'enquête sur le revenu et la participation aux programmes

L'enquête sur le revenu et la participation aux programmes a été réalisée pour la première fois en 1983. L'ERPP est une enquête longitudinale par panels multiples, portant sur les personnes de 15 ans ou plus, qui vise à mesurer les caractéristiques économiques et démographiques de ces personnes pendant une période de deux ans et demi. Les questions de base de l'enquête portent sur les caractéristiques démographiques, l'activité, la participation aux programmes, le montant et le type de revenu gagné et non gagné reçu, l'actif, l'assurance-maladie privée et les prestations de pension. Par suite de contraintes budgétaires et du chevauchement des panels, la taille de l'échantillon a variée de 11 500 à 37 000 ménages. L'enquête porte sur trois panels des mois de février à juillet et sur deux panels seulement, le reste de l'année. Les données de l'ERPP sont recueillies par le Census Bureau (Census, SIPP User's Guide). La nouvelle base de données relationnelles créée à l'intention des utilisateurs des données de l'ERPP permettra d'améliorer considérablement la qualité et l'aptitude à l'utilisation des données.

4.2 Premiers fichiers à grande diffusion de l'ERPP

En mettant en oeuvre l'enquête sur le revenu et la participation aux programmes, le Census Bureau s'écartait sensiblement de sa ligne de conduite relative aux enquêtes en cours. L'ERPP était beaucoup plus complexe et ambitieuse, et les données recueillies étaient riches de promesses pour les démographes et les chercheurs en matière de politique fiscale et de programmes de transfert. Malheureusement, les producteurs de l'ERPP ont initialement reçu des commentaires très négatifs de la part des utilisateurs du fichier: le cliché d'enregistrement portait à confusion; l'enregistrement comportait des zones redondantes (p. ex., âge initial, âge contrôlé et âge imputé); l'emplacement des variables changeait d'un ficher à l'autre; le «zéro» avait différentes significations, comme manquant, non applicable et zéro; enfin, aucun document ne faisait état des procédures de contrôle et d'imputation et les fichiers de données épurées n'étaient pas diffusés dans des délais raisonnables (ADPU, 1989).

Le Census Bureau s'est fondé sur ces commentaires pour remanier les fichiers à grande diffusion de l'ERPP. Ainsi, les données recueillies auprès des panels de 1990 seront diffusées selon un format personne-mois, c.-à-d. qu'un enregistrement sera établi pour chaque mois au cours duquel une personne fait partie de l'échantillon. On assure ainsi l'existence d'un rapport biunivoque entre la personne interviewée et le mois sur lequel portent les données déclarées. Le remaniement a aussi permis de résoudre un des problèmes les plus inquiétants posés par les fichiers à grande diffusion, à savoir le fait qu'un même enregistrement contenait des données relatives

à plusieurs cycles d'interview différents. En outre, les données de 1990 ont été réorganisées: les données redondantes ont été éliminées, on a élaboré un cliché d'enregistrement unique pour tous les ensembles de données de base, et de nouveaux codes ont été ajoutés pour aider l'utilisateur à manipuler certains des concepts les plus complexes utilisés dans l'enquête.

Après avoir corrigé nombre des lacunes signalées par les utilisateurs, le Census Bureau a étudié d'autres façons d'améliorer la qualité des données de l'ERPP et d'en accroître l'aptitude à l'utilisation. Un projet de recherche de grande envergure, financé par la National Science Foundation, a été mis en oeuvre à l'University of Wisconsin pour étudier diverses façons de rendre les données de l'ERPP plus accessibles aux chercheurs universitaires (David, 1989). L'équipe de ce projet a élaboré à l'aide des données recueillies auprès des panels de l'ERPP de 1984 une base de données relationnelles qui est maintenant soutenue par le Census Bureau et à laquelle on peut avoir accès par l'intermédiaire de cet organisme. Le Census Bureau est également en voie d'ajouter à la base de données les données recueillies auprès des panels de 1985.

4.3 Base de données relationnelles de l'ERPP

La base de données relationnelles permet à l'utilisateur d'établir des tableaux répondant à ses propres besoins. Le Census Bureau a élaboré la base actuelle en chargeant les données recueillies auprès des panels de l'ERPP de 1984 dans le progiciel de gestion de base de données relationnelles INGRES. Cette base de données comprend 210 tableaux que l'on peut classer en trois grands types: les tableaux relatifs aux cycles d'interview, les tableaux cumulatifs et les autres tableaux. Les tableaux relatifs aux cycles d'interview sont constitués de longs enregistrements comportant de nombreuses variables relatives aux données d'un cycle d'interview. Afin de réduire l'espace mémoire nécessaire au stockage des tableaux, on a restreint le nombre de lignes ou d'enregistrements pouvant figurer dans chacun d'entre eux. Il existe des tableaux distincts pour les personnes, les ménages et les familles, ainsi que pour différents types de personnes (p. ex., celles touchant un salaire ou un traitement). Les lignes sont constituées par les enregistrements relatifs aux personnes auprès desquelles on a recueilli des données valides pour le tableau, tandis que les colonnes portent sur des variables relatives au sujet du tableau.

Les tableaux cumulatifs sont constitués d'enregistrements courts comportant un petit nombre de variables relatives à l'ensemble des cycles d'interview. Ces tableaux sont spécialement conçus pour faire état de données recueillies auprès de personnes dont le nombre dans l'échantillon est moins élevé (p. ex., auprès des personnes touchant des prestations de Medicare ou des personnes touchant un revenu d'un travail indépendant).

Les autres tableaux sont des tableaux complémentaires comme:

- des tableaux d'appariement, qui permettent d'apparier les personnes, les familles et les ménages;
- divers tableaux portant sur des constantes, les couples, l'état matrimonial, la rétention et la modification du statut de bénéficiaire; enfin,
- des tableaux utilitaires faisant état des dates de mise en oeuvre de l'enquête et des dates de référence.

Ces autres tableaux sont couplés aux tableaux cumulatifs et aux tableaux relatifs aux cycles d'interview afin d'obtenir certains renseignements (p. ex., le nombre de membres d'un ménage qui touchent des prestations de Medicare). Le caractère relationnel de la base de données permet aux utilisateurs de coupler plusieurs fichiers afin d'obtenir les données dont ils ont besoin.

4.4 Avantages

Du fait qu'elle assure une souplesse d'utilisation accrue et une plus grande facilité d'accès, la base de données relationnelles présente de nombreux avantages pour les utilisateurs des données de l'ERPP. Il convient de noter que, avant la création de la base, les fichiers de l'ERPP n'étaient pas utilisés pour construire les modèles fiscaux du U.S. Treasury car le processus d'accès aux microdonnées était alors trop complexe. On utilisait plutôt les fichiers de l'ERPP pour obtenir des données complémentaires aux fins de l'analyse des dispositions fiscales lorsque ces fichiers renfermaient des données non accessibles à partir du RPA ou d'autres sources. Même avec la base de données relationnelles, compte tenu de la différence de taille entre l'échantillon de l'ERPP et celui

du fichier de statistiques sur le revenu des particuliers, les fichiers de l'ERPP sont plus susceptibles d'être utilisés à des fins d'imputation qu'à des fins d'appariement statistique.

5. ENQUÊTE SUR LES FINANCES DES CONSOMMATEURS (EFC)

5.1 L'échantillon de l'enquête sur les finances des consommateurs

La dernière composante du modèle fiscal du Treasury Department à avoir fait l'objet d'importantes modifications afin de mieux répondre aux besoins des utilisateurs est l'enquête sur les finances des consommateurs. Le Federal Reserve Board réalise l'enquête sur les finances des consommateurs tous les trois ans depuis 1983. L'EFC a pour objectif principal de recueillir des données complètes sur l'actif, le passif, les pensions et le revenu, ainsi que des données descriptives sur l'emploi, les antécédents matrimoniaux, la structure de la famille, la santé et d'autres variables démographiques. L'EFC de 1989 (comme celle de 1983, mais non celle de 1986) consiste de fait en deux enquêtes connexes: une enquête-ménage et une enquête sur les pourvoyeurs de pensions. Pour les fins de cet article, nous allons nous limiter à l'étude de l'enquête-ménage, en raison des travaux sur l'imputation multiple actuellement effectués par Arthur Kennickell du Board of Governors de la Federal Reserve.

L'échantillon de l'enquête-ménage a été sélectionné à l'aide d'une double base de sondage, composée d'une liste et d'une base aréolaire. La liste était stratifiée en fonction d'un indice approximatif de richesse, tandis que la base aréolaire l'était en fonction d'un critère géographique. À l'occasion de l'EFC de 1989, environ 870 ménages sélectionnés à partir de la liste et 1,130 ménages sélectionnés à partir de la base aréolaire ont répondu au questionnaire de l'enquête. Cet échantillon de ménages est important en raison du nombre élevé de ménages riches ayant répondu: 100 répondants détenaient chacun une richesse estimative valant de 10 à 250 millions de dollars.

5.2 Non-réponse au questionnaire

L'EFC est la seule source de données d'enquête sur la richesse dont l'échantillon de ménages appartenant aux tranches de revenu supérieures est assez important pour permettre une analyse distincte de ces ménages. (L'ERPP recueille également des données sur la richesse, de façon périodique, mais la taille de son échantillon ne contient pas assez d'observations pour réaliser des analyses distinctes sur les tranches supérieures de la répartition de la richesse.) Non seulement les contribuables à revenu élevé sont-ils très importants pour la modélisation des modifications apportées à la politique fiscale, mais ils présentent un intérêt considérable pour les autres chercheurs. De fait, les contribuables dont le revenu se situe dans le décile supérieur de la répartition des revenus paient 50% des impôts, et ceux dont le revenu se situe dans le centile supérieur en paient 26%.

Malgré l'importance de ces données, on estime en général que les taux de réponse diminuent à mesure que la valeur de la richesse augmente. Comme l'illustre la figure 2, les taux de réponse enregistrés pour l'EFC viennent corroborer cette hypothèse et font ressortir la difficulté qu'il y a à recueillir des données auprès de contribuables aussi «puissants».

Figure 2: Taux de réponse des ménages sélectionnés à partir de la liste (Riches)

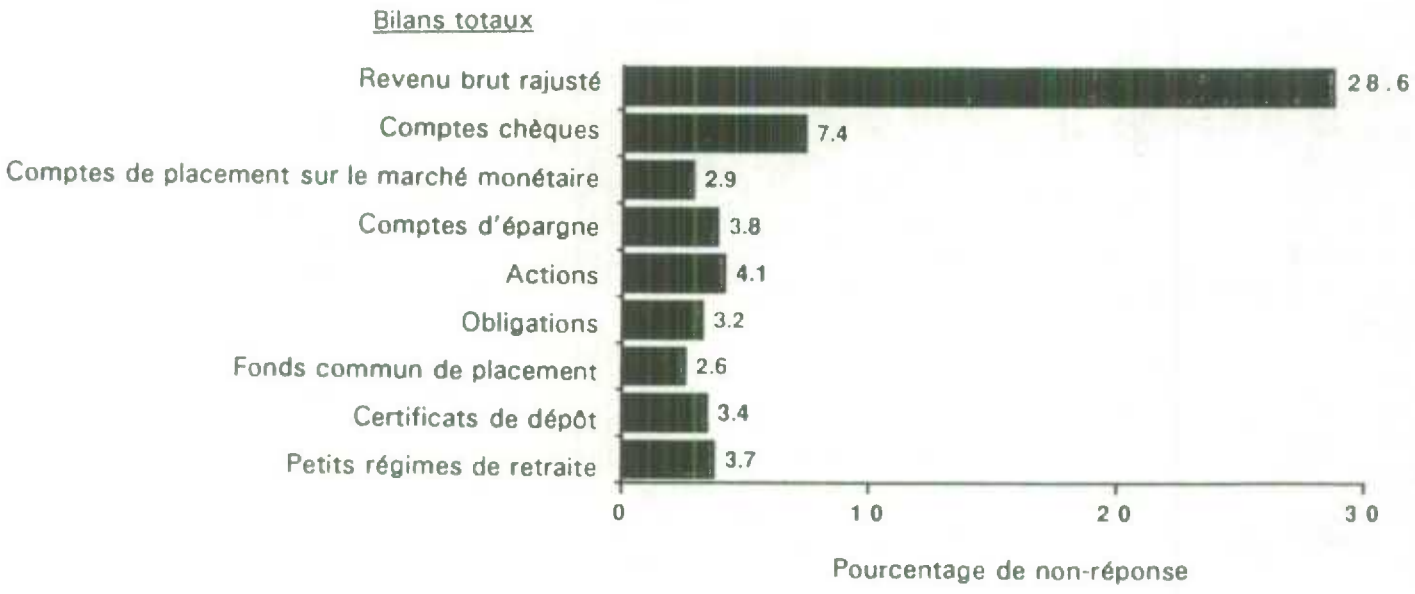
INDICE APPROXIMATIF DE RICHESSE	TAUX DE RÉPONSE	NOMBRE DE RÉPONDANTS
0 < 100 000	48.4%	45
100 000 < 500 000	43.3%	116
500 000 < 1 million	39.6%	158
1 million < 2.5 millions	39.4%	232
2.5 millions < 10 millions	30.6%	215
10 millions < 250 millions	20.1%	100
TOTAL	34.1%	866

La portion des ménages de l'échantillon de l'EFC touchant un revenu élevé constitue un sous-groupe dont le revenu et la richesse sont très difficiles à évaluer, en partie en raison de la non-réponse au questionnaire et de la non-réponse à certaines questions. Une étude portant exclusivement sur les ménages riches, déterminés à la lumière d'un indice de richesse, ne saurait justifier du taux de réponse prévu nécessaire (70%) pour obtenir l'approbation du U.S. Office of Management and Budget. (Les enquêtes fédérales doivent être approuvées par l'Office of Management and Budget avant qu'on puisse procéder à la collecte des données.) Cependant, ces ménages revêtent une importance capitale pour la modélisation fiscale. Afin d'établir des comparaisons et de calculer des estimations nationales, on a complété l'échantillon sélectionné à partir de la liste par un échantillon sélectionné à partir d'une base aréolaire comportant surtout des ménages à faible revenu. Cet échantillon a permis de faire passer le taux de réponse à environ 70%. Toutefois, toutes les enquêtes sont aux prises avec des sous-groupes affichant de faibles taux de réponse. Un des points forts de l'EFC tient au fait que la base de sondage comporte des variables robustes facilitant la modélisation des données manquantes. C'est donc sur cette modélisation qu'ont été concentrés les nouveaux efforts visant à améliorer la qualité des données pour les utilisateurs.

5.3 Correction des données des effets de la non-réponse à certaines questions

La non-réponse à certaines questions constitue également un grave problème pour l'EFC, bien qu'elle soit loin d'être aussi importante que la non-réponse au questionnaire: de fait, le taux de non-réponse à la question sur le revenu brut rajusté est de 28.6% (voir la figure 3). Pour surmonter cette difficulté, l'équipe de l'EFC corrige les données des effets de la non-réponse à certaines questions à l'aide de techniques de régression sophistiquées utilisées en conjonction avec des méthodes d'imputation multiple. Cette méthode, appelée relaxation stochastique, est aussi connue comme l'algorithme d'échantillonnage et de maximisation des espérances mathématiques de Gibbs (Rubin, 1987, 1990 et 1990; Geman, 1984).

Figure 3. Taux de non-réponse relatifs à l'enquête sur les finances des consommateurs



La théorie sur laquelle se fonde cette approche exige que toutes les variables imputées soient des variables continues qui puissent être transformées en variables normales. Ces hypothèses sont à peu près satisfaites. En utilisant une méthode itérative, on emploie un modèle de régression aléatoire pour estimer toutes les valeurs manquantes pour une question donnée - par exemple, la question sur le revenu brut rajusté (RBR). Une fois qu'une valeur a été imputée pour tous les enregistrements où le revenu brut rajusté n'était pas indiqué, la méthode est appliquée à la variable suivante. Quand une valeur a été imputée pour toutes les variables, le cycle d'imputation recommence.

L'imputation comme telle est faite à l'aide d'un modèle de régression aléatoire. L'imputation des valeurs manquantes pour chaque variable s'effectue en quatre grandes étapes.

1. On détermine un ensemble de variables de condition (dans le cas du RBR, ces variables sont au nombre de 300).

2. On génère la matrice des variances-covariances en utilisant les cas pour lesquels une valeur est indiquée pour au moins 75% des variables de condition.

3. Pour chaque enregistrement où aucune valeur n'est indiquée pour la variable visée par l'imputation, on effectue les deux étapes secondaires suivantes:

   a) on construit une matrice des covariances à partir de la matrice de l'étape 2, en se fondant sur les variables de condition pour lesquelles une valeur est indiquée dans cet enregistrement; enfin,

   b) on impute cinq valeurs indépendantes aléatoires pour la variable manquante, en utilisant la régression définie par la matrice des covariances relative à cet enregistrement.

4. On utilise la valeur imputée comme variable de condition dans la matrice des variances-covariances de la variable suivante.

Une fois toutes les variables imputées, on répète l'étape 1. De cette façon, on s'approche par itération de l'estimation du maximum de vraisemblance de la matrice des variances-covariances (Kennickell, 1991).

5.4 Avantages

Un des avantages de la relaxation stochastique tient au fait que les inférences faites à partir de la base de données imputées (à l'aide d'une technique d'imputation multiple) sont aussi valides, puisque la distribution des données imputées est identique à la distribution sous-jacente. En outre, l'imputation tire parti des relations existant entre les variables plutôt que d'imputer chaque variable indépendamment.

6. CONCLUSION

On a traditionnellement mesuré la qualité des données en se fondant sur des critères tenant compte des préoccupations des producteurs de données. Une mesure plus complète de la qualité des données se doit de prendre en considération les deux faces de la médaille, si on peut dire, à savoir la conformité aux normes et l'aptitude à l'utilisation.

La majeure partie des ressources affectées à la collecte des données sont consacrées à la production de statistiques descriptives. Or, des utilisateurs de plus en plus nombreux s'intéressent également aux microdonnées (c.-à-d. aux fichiers à grande diffusion). Malheureusement, au cours des dernières décennies, l'amélioration des produits et l'élaboration de produits plus souples, comme les fichiers à grande diffusion, ont marqué le pas, tandis que les utilisateurs de données ne cessaient de se sophistiquer et d'accroître leurs capacités techniques. Nous assistons actuellement à une inversion de cette tendance, mais, plus important encore, cette inversion découle du fait que les producteurs de données prêtent maintenant une oreille attentive aux commentaires des utilisateurs.

Le fait que les organismes statistiques fédéraux des États-Unis se soucient de plus en plus de la qualité a amené les producteurs de données à reconnaître l'importance de tenir compte du point de vue des utilisateurs. Les quatre enquêtes étudiées illustrent le large éventail d'efforts déployés pour rendre les données plus accessibles au public et pour en accroître l'aptitude à l'utilisation. L'équipe du projet de remaniement du fichier de statistiques sur le revenu des particuliers s'est assurée de mieux tenir compte des besoins des utilisateurs en invitant ces derniers à participer activement au processus de remaniement. Le fichier remanié devrait fournir des données de base plus fiables que le fichier antérieur ainsi que permettre de disposer d'un nombre suffisant

de cas-échantillon relatifs aux déclarations complexes et aux autres catégories de déclarations susceptibles de nécessiter la réalisation d'analyses distinctes pour les fins de la politique fiscale.

De même, la plus grande attention portée aux besoins des utilisateurs a fait du RPA et de l'EFC d'importantes sources de données pour la modélisation des composantes du revenu et pour l'estimation des effets de la politique fiscale. Cette prise en considération des besoins des utilisateurs accroît également la probabilité qu'on puisse maintenant utiliser le riche ensemble de données recueillies dans le cadre de l'ERPP. L'importance de ces sources de données non fiscales tient au fait que la source fondamentale de données fiscales, le fichier de statistiques sur le revenu des particuliers, porte uniquement sur les variables figurant sur les déclarations d'impôt sur le revenu. Or, l'étude de nombreuses questions de nature politique nécessite qu'on dispose de données complémentaires. Il en est ainsi, par exemple, du calcul des effets d'une modification de la politique fiscale à l'échelle de la famille plutôt qu'à celle de la déclaration, ou encore de la modélisation de l'effet de l'inclusion d'une source de revenu ne figurant pas actuellement sur les déclarations. L'accroissement de la facilité d'utilisation des données, le perfectionnement des plans d'échantillonnage et l'amélioration des ensembles de données à grande diffusion ont pour effet d'élargir l'éventail de données pouvant être utilisées aux fins de l'analyse des effets de la politique fiscale.

BIBLIOGRAPHIE

Association of Public Data Users (1989). *SIPP Supplement to the APDU Newsletter*, 1 supplements 1-6 and 2 supplements 1-2, NJ: Princeton University Computing Center.

Barr, R., et Turner, S.J. (1978). A New Linear Programming Approach To Micro Data File Merging, *1978 Compendium of Tax Research*, Department of the Treasury, . 131-149.

Bates, J. (1991). Creating a Database for Longitudinal Analysis of Families, *Proceedings of the Section on Survey Research Methods*.

Cilke, J.M., et Wyscarver, R.A. (1987). The Treasury Individual Income Tax Simulation Model, *Compendium of Tax Research 1987*, Washington, D. C.: Department of the Treasury, Office of Tax Analysis.

Cilke, J.M., et Wyscarver, R.A. (1990). The Treasury Individual Income Tax Simulation Model, Washington, D.C.: Department of the Treasury, Office of Tax Analysis.

Czajka, J., et Schirm, A. (1990). Overlapping Membership in Annual Samples of Individual Tax Returns, *Proceedings of the Section on Survey Research Methods*.

Czajka, J. (1988). Development of a New Income Classifier for a Sample of Individual Tax Returns, *Proceedings of the Section on Survey Research Methods*.

David, M., et Robbin, A. (1989). Database Design for Large Scale, Complex Data, U.S. Bureau of the Census, SIPP Working Paper 8923.

Deming, W. E. (1986). *Out of the Crisis*, Massachusetts Institute of Technology, Center for Advanced Engineering Study: Cambridge, Ma.

Fay, R.E. (1984). Some Properties of Estimates of Variance Based on Replication Methods, *Proceedings of the Section on Survey Research Methods*.

Fay, R.E. (1989). Theory and Application of Replicate Weighting for Variance Calculations, *Proceedings of the Section on Survey Research Methods*.

Geman, S., et Geman, G. (1984). Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Hostetter, S. *et coll.* (1990). Choosing The Appropriate Income Classifier For Economic Tax Modeling, *Proceedings of the Section on Survey Research Methods*.

Hostetter, S., et O'Conor, K. (1991). Satisfying the Needs of Income Tax Policy Modelers While Preserving The Reliability of Descriptive Statistics, *Proceedings of the Section on Survey Research Methods*.

Judkins, D.R. (1990). Fay's Method For Variance Estimation, *Journal of Official Statistics*, 223-239.

Juran, J.M. (1988). *Juran on Planning Quality*, The Free Press: New York, NY.

Lent, J. (1991). Variance Estimation for Current Population Survey State Labor Force Estimates, *Proceedings of the Section on Survey Research Methods*.

Schirm, A., et Czajka, J. (1990). Intertemporal Stability in Total Income and the Overlap in Annual Samples of Tax Returns, *Proceedings of the Section on Survey Research Methods*.

Smith, C., (1989). Social Security Administration Continuous Work History Sample, *Social Security Bulletin*. (L'échantillon du Continuous Work History fut commencé par le Social Security Administration dans les années 1930. Jusqu'en 1980, le Social Security Administration en an produit des fichiers à grande diffusion, mais en raison de préoccupations quant à une divulgation potentielle de ces données, il a cessé toute diffusion de celles-ci.)

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons: New York, NY.

Rubin, D. (1987). EM and Beyond, European Meeting of Biometric Association.

Rubin, D. (1990). Imputation Procedures and Inferential Versus Evaluative Statistical Statements, *Proc. Census Annual Research Conference VI*.

United States Bureau of the Census (1978). The Current Population Survey: Design and Methodology, Technical Paper 40, 2.

United States Bureau of the Census (1987). Survey of Income and Program Participation: Users' Guide, Chapter 1.

Wolter, K.M. (1985). *Introduction to Variance Estimation*, Springer Series in Statistics, New York: Springer-Verlag Inc.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

EXAMEN DES CRITÈRES DE STATISTIQUE CANADA RELATIFS À LA QUALITÉ DES DONNÉES DIFFUSÉES

R.D. Burgess[1]

RÉSUMÉ

Lorsqu'ils publient des données, les auteurs de Statistique Canada doivent aviser les utilisateurs de toute éventuelle limitation importante liée à la qualité des données en question. On recommande parfois aux utilisateurs de faire preuve "de prudence" dans l'utilisation de certaines données. Dans d'autres cas, des critères précis sont appliqués afin d'éviter la publication de données de moindre qualité. Il est possible d'empêcher la diffusion des données par l'intermédiaire des spécifications relatives aux tableaux ou encore en "supprimant les données" à l'intérieur des tableaux. Pour avoir accès à ces données de moindre qualité, l'utilisateur doit faire l'acquisition de tableaux personnalisés et même alors, il est possible qu'il doive composer avec certaines restrictions.

Certains statisticiens et utilisateurs estiment que la qualité des données ne devrait pas, en elle-même, servir de critère de diffusion. Selon eux, toutes les données sont susceptibles d'avoir une certaine valeur et seul l'utilisateur est en mesure de décider de leur pertinence.

Le présent article a pour objet de décrire les méthodes actuelles mises en oeuvre à Statistique Canada pour évaluer la qualité des données diffusées, d'examiner et de comparer les critères utilisés par divers programmes statistiques, ainsi que de déterminer les avantages et les inconvénients de l'utilisation des critères relatifs à la qualité des données diffusées.

MOTS CLÉS: Qualité des données; critères de diffusion; suppression des données.

1. INTRODUCTION

1.1 Objet et portée

Lorsqu'ils publient ou diffusent, par quelque autre méthode, des données, les auteurs de Statistique Canada sont tenus de donner aux utilisateurs des renseignements sur les concepts et les méthodes utilisés, ainsi que de leur indiquer quelle est la qualité de ces données (Statistique Canada, 1986). Ainsi, l'utilisateur est en mesure de déterminer si les données décrivent les caractéristiques qu'il souhaite mesurer et si elles sont suffisamment précises pour les fins auxquelles il veut les utiliser. On recommande parfois aux utilisateurs de faire preuve "de prudence" dans l'utilisation de certaines données. Cependant, la plupart du temps, en raison des objectifs de l'enquête et du coût de la totalisation et de la publication ou encore de contraintes d'espace, la majorité des données de moindre qualité ne sont pas incluses dans les spécifications relatives aux tableaux devant figurer dans les publications ou être diffusés.

Certains programmes statistiques sont déjà dotés de normes ou de lignes directrices qui restreignent la diffusion des données de moindre qualité. La diffusion de ces données peut être empêchée par l'intermédiaire des spécifications relatives aux tableaux devant être publiés, mais, à l'occasion, des données de piètre qualité englobées par une spécification sont éliminées ou "supprimées" à l'intérieur même d'un tableau diffusé. Pour

---

[1] R.D. Burgess, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6.

avoir accès à ces données de moindre qualité, l'utilisateur doit faire l'acquisition de tableaux non publiés ou personnalisés et même alors, il est possible qu'il doive composer avec certaines restrictions.

Certains utilisateurs estiment que la qualité des données ne devrait pas, en elle-même, servir de critère de diffusion. Selon eux, toutes les données sont susceptibles d'avoir une certaine valeur et seul l'utilisateur est en mesure de décider de leur pertinence. Certains statisticiens partagent ce point de vue, tandis que d'autres jugent qu'il incombe à un organisme statistique d'empêcher la diffusion des "données" qui ne sont pas aptes à être utilisées et qui pourraient autrement être considérées comme des "statistiques officielles".

Le présent article a pour objet de décrire les méthodes actuellement mises en oeuvre à Statistique Canada pour évaluer la qualité des données diffusées, d'examiner et de comparer les critères utilisés par divers programmes statistiques, ainsi que de déterminer les avantages et les inconvénients des critères relatifs à la qualité des données diffusées.

Les questions ayant trait aux critères relatifs à la qualité des données diffusées sont distinctes et largement complémentaires des questions relatives à la confidentialité. Ces dernières questions ne sont pas étudiées dans le présent article. Toutefois, il est entendu, sans que ce fait influe sur la présente discussion, qu'il est pratique courante d'éviter de divulguer des renseignements confidentiels dans tous les produits statistiques du Bureau.

1.2 Organisation

Après avoir donné, à la section 2, une description fondamentale de la nature des critères de diffusion et de la façon dont ils peuvent être appliqués, nous exposerons, à la section 3, le cadre dans lequel l'examen et l'étude des critères de diffusion s'effectuent à Statistique Canada et nous y présenterons les questions générales soulevées par ces critères. La section 4 porte sur l'utilisation qui est faite des critères de diffusion à Statistique Canada et dans quelques autres organismes statistiques. Nous décrivons, à la section 5, l'effet ou l'effet éventuel de certains critères de diffusion. Enfin, après avoir présenté et étudié, à la section 6, certains arguments militant en faveur et en défaveur de l'utilisation de critères de diffusion, nous exposerons à la section 7 les conclusions auxquelles nous en sommes arrivés.

2. DÉFINITIONS ET TERMINOLOGIE

2.1 Critères de diffusion et suppression

Pour les fins du présent article, un critère relatif à la qualité des données diffusées est d'abord une règle ou un ensemble de règles utilisé(e) pour faire une évaluation subjective ou objective de la qualité éventuelle ou réelle d'un ensemble de données, puis, deuxièmement, une règle ou un ensemble de règles permettant de déterminer, en fonction de la qualité des données, s'il faut empêcher la publication ou la diffusion de ces dernières, ou encore poser certaines restrictions à leur utilisation. Comme ces critères sont destinés à être appliqués pour déterminer quelles données sont suffisamment fiables pour être offertes aux utilisateurs sans restriction, ils constituent des normes relatives à la qualité des produits statistiques. On dit des données qui ne satisfont pas à ces normes et qui ne sont pas diffusées qu'elles ont été "supprimées".

Dans la suite de cet article, nous distinguerons trois formes de suppression de données: la "suppression de cases", la suppression par l'intermédiaire des spécifications relatives aux tableaux et la suppression par l'intermédiaire de "restrictions secondaires". Dans le cas de la suppression de cases, les règles ou critères utilisés sont appliqués à chaque case de données ou total figurant dans un tableau et consistent à remplacer par un symbole le chiffre figurant dans une case, ou dans chaque case d'une colonne ou d'une ligne d'un tableau statistique.

Dans le cas de la suppression par l'intermédiaire des spécifications relatives aux tableaux, les règles peuvent s'appliquer à une variable ou une caractéristique distincte, à une publication, à un ensemble de tableaux ou à un seul tableau statistique. Ce type de suppression peut se traduire par l'exclusion d'une ligne ou d'une colonne des spécifications relatives à un tableau, par le regroupement de lignes ou de colonnes, par l'élimination d'une variable dans les spécifications relatives à un tableau ou par l'interdiction de publier ou de diffuser un ensemble

particulier de données. Cette suppression peut s'appliquer uniquement à certains niveaux ou certaines régions géographiques, ou à certains tableaux croisés.

Les critères de diffusion prenant la forme de restrictions secondaires relatives à l'utilisation des données peuvent empêcher l'utilisateur de diffuser de nouveau les données visées, en lui interdisant, par exemple, de citer des estimations peu fiables dans un rapport.

Enfin, la suppression peut s'appliquer à un ou plusieurs des supports de diffusion utilisés pour un programme statistique.

2.2 Qualité et utilité

Juran et Gryna (1980) définissent la qualité comme "l'aptitude à être utilisé". Dans le même ordre d'idées, nous considérons comme "aptes à être utilisées" les données suffisamment précises qui sont conformes aux exigences des utilisateurs. Toutefois, dans cet article, la qualité est assimilée à la précision. Les données qui répondent aux besoins des utilisateurs sont simplement dites "utiles" ou ayant une utilité. Ce sont des données recueillies à l'aide de concepts convenant aux exigences des utilisateurs et qui permettent d'offrir à ces derniers les tableaux croisés détaillés s'appliquant aux régions géographiques dont ils ont besoin. (Les données doivent aussi être actuelles et être recueillies moyennant un coût raisonnable, mais l'étude de ces deux éléments n'entre pas dans le cadre de la présente discussion.)

Entre autres, Juran et Gryna caractérisent l'aptitude à être utilisé par la qualité de la conception et le degré de conformité aux paramètres de conception. Nous établirons un certain parallèle avec ces deux critères en considérant parfois la variance et le biais comme deux aspects distincts de la qualité.

3. HISTORIQUE ET QUESTIONS GÉNÉRALES

3.1 Historique

Les critères relatifs à la qualité des données diffusées constituent une des facettes éventuelles de la présentation des données statistiques et des normes qualitatives. À Statistique Canada, c'est le Comité des méthodes et des normes qui a le mandat d'élaborer les normes statistiques. Le Comité s'efforce de déterminer s'il est nécessaire que le Bureau dispose de normes écrites. L'étude des critères de diffusion figure présentement sur l'ordre du jour du Comité. Le comité a déjà préparé certaines normes et lignes directrices sur la présentation des tableaux. Entre autres, ces normes et lignes directrices exigent des auteurs qu'ils établissent un juste équilibre entre l'utilité et la qualité des données lorsqu'ils élaborent des tableaux statistiques.

Le Comité consultatif des méthodes statistiques du Bureau, dont les membres sont des statisticiens de l'extérieur, a aussi étudié les critères de diffusion, sur le plan des principes, ainsi que les pratiques actuelles, et a exprimé ses vues à cet égard. Pour donner suite à cette étude, un groupe de travail doit examiner les pratiques de façon plus approfondie et considérer, de concert avec des questions d'ordre plus général, l'effet qu'aurait la modification de ces pratiques sur les divers programmes statistiques.

Le présent mémoire peut être considéré comme un examen préliminaire de ces questions, mais sans lien direct avec les activités du groupe de travail et des comités.

3.2 Questions générales

L'étude des critères de diffusion peut être considérée comme une facette de l'étude plus large de la qualité et de l'utilité des produits statistiques, laquelle débute par la prise en considération des objectifs et de la conception du programme statistique. Cependant, on peut également soutenir que l'évaluation des avantages des critères de diffusion doit tenir compte du fait que les données susceptibles d'être supprimées ont déjà été recueillies et traitées et que certaines de ces données pourraient être diffusées moyennant des frais additionnels relativement minimes, sinon inexistants. On peut également supposer que tous s'entendront pour reconnaître que les données non aptes à être utilisées ne devraient pas être diffusées. L'obtention d'un consensus sur ce qu'on entend par

une donnée non apte à être utilisée et plus particulièrement sur le degré de qualité à partir duquel les données deviennent non aptes à être diffusées, semble moins probable. De fait, jusqu'à maintenant, une bonne partie des discussions préliminaires sur le sujet des critères de diffusion ont porté sur l'acceptabilité de la suppression de cases. Le support de diffusion ainsi que le type de publications (normalisées ou personnalisées) visés par les critères ont également fait l'objet de discussions.

À l'intérieur du sujet plus vaste de la qualité et de l'utilité, nous nous proposons d'étudier les questions suivantes:

1/ les bénéfices nets découlant ou pouvant découler de l'application par Statistique Canada de critères de diffusion afin de contrôler la qualité des données diffusées et publiées;

2/ La nature ou la forme des critères de diffusion pouvant ou devant être appliqués par le Bureau.

Par souci de simplification, il est possible de reformuler ces points sous forme de questions.

1/ Y a-t-il lieu d'établir des normes de qualité minimales pour tous les produits statistiques? Sur le plan de la variance? Sur le plan du biais? Ces normes permettront-elles d'accroître l'utilité des produits statistiques, en termes absolus ou compte tenu des coûts? Incombe-t-il au Bureau d'implanter de telles normes? Le bureau peut-il tirer profit de l'implantation de ces normes?
2/ Quels critères de qualité devrions-nous utiliser? En matière de présentation des tableaux? À l'échelle de la case de données?
3/ À quels produits et à quels types de données ces critères de qualité devraient-ils être appliqués? Aux indicateurs économiques? Uniquement aux publications répertoriées dans le catalogue ou également aux autres véhicules de diffusion? Uniquement aux produits normalisés ou également aux produits personnalisés?

Pour remettre ces questions en contexte, il convient de noter que la suppression, telle qu'elle est actuellement appliquée par Statistique Canada, ne semble pas représenter en soi une entrave sérieuse à l'utilisation des données. Bien qu'il y ait des exceptions, nombre des utilisateurs consultés seraient d'accord pour reconnaître que c'est la qualité et non la suppression qui représente la principale limitation (p. ex., Division de l'évaluation des programmes, 1988). En revanche, ce point de vue est susceptible de varier en fonction de la mesure dans laquelle le Bureau a actuellement recours à la suppression.

4. UTILISATION DES CRITÈRES DE DIFFUSION

4.1 Critères de diffusion de Statistique Canada

Il n'existe à Statistique Canada aucune norme explicite permettant de déterminer si les données sont d'une qualité suffisante pour être diffusées. De même, aucune ligne directrice n'interdit de soumettre la diffusion des données à certaines restrictions explicites relatives à la qualité.

Le Bureau dispose déjà de normes et de lignes directrices pour la présentation des tableaux dans ses publications (Comité des méthodes et des normes, 1990). Ces normes et lignes directrices prévoient la prise en considération de la qualité des données. En voici un extrait:

"Le tableau ou l'ensemble de tableaux doit être présenté de façon à restreindre le nombre de cases dépourvues d'information et à éviter de suggérer implicitement un degré d'exactitude supérieur à la réalité."

"Si, pour une classification commune dans un tableau, toutes les cases de deux ou plusieurs lignes ou colonnes sont ou risquent d'être:

a) en blanc ou à valeur zéro (autrement que par arrondissement aléatoire);
b) supprimées pour des raisons de confidentialité ou
c) de qualité médiocre ou inacceptable;

alors, il convient de regrouper ou de supprimer ces lignes ou colonnes ou de repenser le tableau, sauf s'il faut les conserver par souci d'uniformité entre les cycles d'une enquête."

Une ligne directrice semblable vise les cas où une grande partie des cases d'un tableau sont ou risquent d'être "de qualité médiocre ou inacceptable".

Il revient aux chefs de programme de définir ce qu'ils entendent par des données de piètre qualité et de décider si de telles données doivent être diffusées. Les lignes directrices reconnaissent la nécessité de maximiser l'utilité des données, compte tenu des objectifs de chaque programme statistique, des utilisateurs connus, ainsi que du coût et des contraintes physiques relatives à la présentation. Elles n'empêchent donc pas la diffusion de données de piètre qualité, cette condition étant d'ordinaire réalisée suite à la présentation de données plus fiables. Cependant, elles visent à limiter la quantité relative de données non fiables dans les produits statistiques du Bureau. Ces lignes directrices n'imposent aucune restriction à la diffusion des données dans des produits personnalisés. Enfin, elles ne prévoient ni n'empêchent le recours à la suppression de cases, bien que la proportion de données supprimées doive implicitement être limitée.

Il est probable que la manifestation la plus commune des principes sous-tendant ces lignes directrices soit la variation du niveau de détail des tableaux croisés selon la taille des régions géographiques ou des diverses sous-populations. Toutefois, les décisions relatives au niveau de détail ne sont pas toujours fondées sur des critères explicites de diffusion.

L'examen de plusieurs programmes de Statistique Canada a révélé qu'il était plus fréquent de procéder à l'application systématique de critères de diffusion dans le cadre des enquêtes-échantillon. Il est beaucoup plus rare qu'on applique de tels critères dans le cadre d'un recensement ou d'une enquête utilisant une combinaison de méthodes de recensement et de méthodes d'échantillonnage. Les critères de diffusion actuellement appliqués épousent six formes fondamentales.

1/ Coefficient de variation limite. Comme on peut s'y attendre, les critères de diffusion les plus fréquemment utilisés sont fondés sur un coefficient de variation limite de l'erreur d'échantillonnage estimée ou prévue. Ainsi, on utilise un tel critère dans le cadre de l'enquête sur la population active, de l'enquête sur les finances des consommateurs, de l'enquête sur les dépenses des familles, ainsi que de la plupart des autres enquêtes-ménages réalisées par le Bureau. À cet égard, il semble que les pratiques de l'enquête sur la population active servent de modèle commun ou de précédent pour les autres enquêtes. (De même, la plupart des enquêtes-ménages portent sur un échantillon prélevé dans celui de l'EPA.) L'enquête sur la population active utilise un coefficient de variation de 33 1/3% comme critère de diffusion: on applique la suppression de cases pour tous les éléments d'information qui excèdent cette limite. Cette pratique s'applique à tous les tableaux, tant normalisés que personnalisés. Chaque fois qu'on supprime une estimation, le taux correspondant est également supprimé.

   Les données de l'enquête sur la santé et les limitations d'activités ont également fait l'objet de suppressions fondées sur l'erreur relative d'échantillonnage estimée. Les données de cette enquête postcensitaire de 1986 pour le Canada, les provinces et les régions métropolitaines de recensement ont été éliminées des tableaux lorsque le coefficient de variation était supérieur à 25%. Dans le cas des tableaux dont les données étaient ventilées à l'échelle de régions infraprovinciales autres que les régions métropolitaines de recensement, on a utilisé un critère de 33% pour éviter de supprimer une trop grande quantité de données.

   Dans le cas des statistiques annuelles de l'enquête sur les voyages internationaux, la limite supérieure du coefficient de variation a été fixée à 16.5%. Pour la plupart des publications répertoriées au catalogue, cette limite est appliquée par l'intermédiaire des spécifications relatives aux tableaux.

   Enfin, il arrive, pour certaines enquêtes possédant des critères de publication des données fondés sur le coefficient de variation, que ces critères ne soient appliqués ni aux tableaux personnalisés ni aux données diffusées sur bandes de microdonnées.

2/ Erreur d'arrondissement limite. Il est rare qu'on utilise l'erreur introduite par l'arrondissement comme critère de diffusion. On le fait, entre autres, dans le cadre de l'enquête sur la population active, où les

estimations des totaux sont d'ordinaire arrondies au millier ou à la centaine le(la) plus près. On évalue l'erreur attribuable à cet arrondissement sous forme de proportion de l'erreur d'échantillonnage prévue pour une case de taille donnée et on lui fixe une limite. Les cases et les taux correspondants pour lesquels l'erreur est supérieure à la limite fixée sont ensuite supprimés.

3/ Comparaison avec des données repères. Il existe des programmes statistiques pour lesquels certaines des données recueillies chaque mois auprès d'un échantillon de répondants peuvent être comparées à des données plus fiables, recueillies beaucoup moins fréquemment. Ces données plus fiables sont parfois utilisées comme données repères. En pareil cas, si la valeur mensuelle s'écarte de la valeur repère dans une mesure supérieure à la limite fixée, la valeur inscrite dans la case mensuelle est supprimée. L'enquête mensuelle sur les manufactures utilise un tel critère de suppression de cases fondé sur une comparaison de données avec les données du recensement annuel des manufactures. Les limites de tolérance varient selon la longueur de la période s'étant écoulée entre la tenue de l'enquête mensuelle et celle du recensement utilisé comme recensement repère.

4 Taux d'imputation limite. L'imputation constitue une source éventuelle d'erreurs pour la plupart des programmes statistiques. Il arrive parfois, mais selon toute vraisemblance peu souvent, qu'on établisse un taux d'imputation limite comme critère de diffusion des données: c'est le cas, en particulier, de l'enquête annuelle sur le commerce de détail où les lignes des tableaux provinciaux (Grand groupe de commerces) pour lesquelles 30% ou plus des données ont été imputées sont regroupées.

5/ Restriction relative à la publication secondaire. Pour certaines enquêtes, les données non publiées sont diffusées sans suppression dans le cadre d'une entente aux termes de laquelle toute publication subséquente des données est soumise à certaines restrictions: soit que les données ne respectant pas les critères de diffusion standard ne doivent pas être présentées dans des rapports ou autrement republiées, soit que ces données doivent être assorties d'un indicateur de qualité ou d'une mise en garde. Ainsi, on applique ce genre de règles aux fichiers de microdonnées de l'enquête sur la population active et aux données de l'indice des prix à la consommation à l'échelle infraprovinciale.

6/ Évaluation générale des données et des sources de données. Dans le cas des enquêtes-échantillon, il n'est pas rare qu'on précise les plus petites unités géographiques selon lesquelles les données peuvent être ventilées. Ces limites peuvent être implicites ou clairement établies au stade du plan d'échantillonnage. Les données visées par de telles limites ne sont jamais publiées, peu importe qu'il existe ou non d'autres critères de diffusion s'appliquant aux produits normalisés ou personnalisés.

Il est plus rare qu'on applique des critères de diffusion aux données analytiques. Dans la plupart des cas, les données d'entrée sont fortement agrégées et probablement d'une qualité telle que l'utilisation de critères de diffusion n'aurait qu'un effet négligeable. Une des principales exceptions à cette règle est le système de comptabilité nationale (SCN). Dans certains secteurs produisant des estimations à l'échelle de la branche d'activité, il existe un système d'évaluation de la qualité qu'on utilise parfois pour déterminer quelles données doivent être diffusées. Ce système prévoit une procédure articulée pour évaluer la qualité des divers agrégats et données. La cote de qualité la plus élevé, le "1", est attribuée aux données recueillies dans le cadre d'un recensement ou aux données tirées des dossiers administratifs ne nécessitant qu'un redressement minimal. La plus faible cote de qualité "acceptable" est le "3", tandis que la plus faible cote, "inacceptable", est le "4". Jusqu'à récemment, les données auxquelles on avait attribué une cote de "4" n'étaient pas publiées. Cette règle est encore en vigueur dans certains secteurs du SCN. En revanche, il existe un secteur où la méthode d'évaluation utilisée a fait l'objet d'un remaniement complet en 1990 (Taillon, 1990). Bien que les données actuellement publiées pour ce secteur soient les mêmes que les données publiées en conformité avec l'ancien système d'évaluation et les anciens critères de diffusion, certaines des données actuellement publiées ont une cote de "4".

L'examen des programmes statistiques a également révélé que certains programmes ne sont dotés d'aucun critère de diffusion explicite ou systématiquement appliqué, mais imposent des restrictions à la diffusion des données sur une base empirique. Il n'est pas rare que de telles restrictions soient imposées par l'intermédiaire des spécifications relatives aux tableaux et elles pourraient éventuellement être appliquées dans le cadre de n'importe quel programme statistique.

Le recensement de la population et le recensement de l'agriculture constituent deux exemples importants de tels programmes. Ces programmes ne sont dotés d'aucune norme ou ligne directrice explicite permettant de décider, selon des critères de qualité, si des données peuvent ou devraient être diffusées. Cependant, ils sont dotés de processus officiels de certification permettant d'évaluer la qualité des données. Lorsqu'on décèle des problèmes de qualité, on décide si les données doivent être supprimées ou si elles doivent être assorties d'une mise en garde.

Bien que rarement, ce processus a parfois entraîné une réduction du nombre de catégories selon lesquelles on avait prévu ventiler les données relatives à une variable (p. ex., les données recueillies à l'aide de la question du recensement de 1981 sur le type de construction résidentielle), l'élimination des données relatives à une variable dans les produits normalisés (p. ex., les données recueillies à l'aide de la question du recensement de 1986 sur les autochtones), ou encore la suppression des données sur les caractéristiques démographiques pour certaines régions géographiques, comme ce fut le cas pour les réserves et les établissements indiens partiellement dénombrés au recensement de 1986. (Dans ce dernier cas, la suppression des données a pour une bonne part été rendue nécessaire par l'absence de données sur les caractéristiques démographiques pour certaines régions géographiques.)

Le recensement de la population n'a pas encore étendu le recours à une telle suppression des données aux produits personnalisés. En pareil cas, on discute d'ordinaire de la qualité des données avec les utilisateurs et on fournit à ces derniers de la documentation sur la qualité.

On utilise des pratiques similaires pour le recensement de l'agriculture.

4.2 Comparaison avec les critères de diffusion utilisés par d'autres organismes

Bien sûr, Statistique Canada n'est pas le seul organisme statistique à utiliser des critères de diffusion ou de suppression. L'examen des normes et lignes directrices appliquées par quelques-uns de ces autres organismes nous a permis de faire les constatations suivantes sur le plan des critères de diffusion.

1/ Office of Management and Budget - États-Unis (Office of Management and Budget, 1988)

"... aucune donnée recueillie dans le cadre d'une enquête statistique ne doit être supprimée ou interdite de diffusion à moins que ... la donnée ne respecte pas les normes de qualité relatives à la publication que l'organisme parrain a précisé dans sa demande de ... collecte de données ...". [Traduction] Cette norme (ainsi que certaines restrictions relatives à la confidentialité) s'applique à tous les organismes statistiques fédéraux des État-Unis.

2/ Bureau of the Census, U.S. Department of Commerce (Bureau of the Census, 1975)

- Aucune case de données ne doit être supprimée en raison d'une importante erreur d'échantillonnage ou d'importantes erreurs d'observation, bien qu'il faille faire preuve de jugement au moment de la définition des tableaux devant être publiés.

- Les tableaux non publiés parce que la majeure partie des estimations ne sont pas fiables, en termes d'erreur d'échantillonnage, peuvent être diffusées aux utilisateurs assorties d'une indication appropriée de l'erreur d'échantillonnage.

- L'existence d'un biais important constitue une raison suffisante pour supprimer des données dans un tableau destiné à être publié.

- Lorsque des données non fiables sont diffusées (p. ex. dans des totalisations personnalisées), le récipiendaire doit être avisé par écrit qu'il est tenu d'indiquer les estimations relatives à l'erreur d'échantillonnage dans toute publication où ces données figurent ainsi que d'y faire état de tout biais grave dont il est au fait.

Shapiro et coll. (1985) ont décrit les pratiques mises en oeuvre par le Bureau of the Census:

> "... lorsque les échantillons sont de si petite taille que les estimations de l'erreur d'échantillonnage en deviennent elles-mêmes très peu fiables, les estimations sont souvent supprimées. La valeur de ces "très graves erreurs" peut varier de 15% à 100% ou plus, le critère généralement utilisé étant l'erreur relative entachant les principales statistiques diffusées pour une enquête donnée. Ainsi, si un total pour l'ensemble des É.-U. est entaché d'une erreur relative de 1%, il peut arriver que les données ventilées à l'échelle des régions comportant une erreur supérieure à 15% soient supprimées; en revanche, si l'erreur relative entachant le total pour l'ensemble des É.-U. est de 10%, il est possible que les données ventilées à l'échelle des régions soient supprimées uniquement lorsque l'erreur relative est égale ou supérieure à 75%."
>
> "La suppression des données pour lesquelles le taux de non-réponse est très élevé est une pratique courante dans tous les secteurs économiques, mais on n'y a point recours dans les secteurs démographiques où il est rare que les taux de non-réponse soient très élevés. Comme l'existence de taux de non-réponse plus élevés accroît la possibilité d'introduction de biais substantiels dans les estimations d'enquêtes publiées, il arrive parfois qu'on supprime la publication des données. Toutefois, le taux de non-réponse nécessitant la suppression des données varie d'une division à l'autre et même d'une enquête à l'autre à l'intérieur de la même division. En général, lorsque le taux de non-réponse est supérieur à ... 25% ..." [Traduction]

Le Bureau of the Census et Statistique Canada ont en commun de mettre en oeuvre une grande diversité de pratiques, mais ces pratiques varient d'un organisme à l'autre. Ainsi, il semble que Statistique Canada applique davantage la suppression aux totaux ou chiffres figurant dans les cases, tandis qu'un examen de quelques publications du Bureau of the Census laisse supposer que cet organisme met plutôt l'accent sur la suppression des moyennes, des médianes et des pourcentages, pour des raisons de qualité, sans supprimer les totaux de cases correspondants. Entre autres, on justifie cette mesure en disant qu'elle a pour objet de permettre aux utilisateurs de calculer leurs propres agrégats (Bureau of the Census, 1989). De fait, aucun des deux organismes ne supprime les chiffres figurant dans les cases en vue de dissimuler les données. Toutefois, dans le cas des tableaux publiés par Statistique Canada, il est beaucoup plus difficile pour l'utilisateur d'obtenir ou d'estimer les données supprimées. Lorsque le nombre de cases dont les données ont été supprimées est assez élevé, ces données sont en pratique dissimulées.

Tout comme à Statistique Canada, il semble que les critères de diffusion ne fassent pas l'objet d'une application universelle dans les autres organismes statistiques, du moins pas dans le cas de la suppression de cases. D'un organisme à l'autre, les enquêtes du même type ne sont pas soumises aux mêmes critères de diffusion. Prenons pour exemple les enquêtes sur la population active. Tout comme Statistique Canada, le Department of Statistics de Nouvelle-Zélande et le Bureau of Labor Statistics des États-Unis utilisent la suppression de cases lorsque l'erreur relative d'échantillonnage est élevée. Toutefois, dans le cas de la Nouvelle-Zélande, les pourcentages correspondants (p. ex., taux de chômage) ne sont généralement pas supprimés lorsque le total estimatif l'est (Department of Statistics, 1990). De son côté, le Bureau of Labor supprime surtout les pourcentages, mais il arrive aussi que les totaux estimatifs pour les États et certaines régions géographiques soient supprimés lorsque les critères de calcul beaucoup plus stricts établis pour ces régions ne sont pas respectés (Bureau of Labor, 1988 et 1990). Enfin, le Bureau of Statistics d'Australie n'utilise pas la suppression de cases dans le cadre de son enquête sur la population active, mais signale par un indicateur quelles sont les données pour lesquelles le coefficient de variation est supérieur à 25% (Bureau of Statistics d'Australie, 1990).

On découvrirait que l'attitude adoptée par les divers organismes en matière de suppression de cases est fort différente si on examinait, par exemple, les enquêtes sur les dépenses et le revenu des ménages et les enquêtes sur les dépenses des familles.

Les pratiques mises en oeuvre pas ces autres organismes en matière de présentation des tableaux n'ont pas été évaluées pour les fins du présent article.

4.3 Mesure de la qualité des données et critères de diffusion

Bien sûr, les questions relatives aux critères de diffusion ne sauraient être étudiées séparément des questions relatives à l'évaluation de la qualité et à l'information des utilisateurs de la qualité des données. Pour qu'on puisse envisager d'avoir recours à la suppression, il doit exister une certaine forme d'évaluation de la qualité ou de certification des données. Si tel est le cas, n'est-il pas nécessaire et suffisant d'informer les utilisateurs de la qualité des données? Les réponses implicites données par les divers programmes statistiques à cette question peuvent être très différentes.

En termes de présentation de renseignements sur la qualité des données correspondant aux critères utilisés en matière de suppression, les programmes statistiques de Statistique Canada étudiés peuvent être regroupés en quatre catégories, selon qu'ils:

1/ publient des mesures correspondantes de la qualité des données à l'intérieur des tableaux, pour toutes les cases de données;
2/ publient des mesures correspondantes de la qualité des données à l'extérieur des tableaux, pour toutes les cases de données;
3/ signalent par un indicateur les données peu fiables et indiquent les critères de qualité correspondants pour ces données et les données supprimées, mais sans publier d'indices précis de la qualité des données pour les données non assorties d'un indicateur et non supprimées;
4/ n'attribuent pas d'indicateurs aux données en fonction de leur qualité ni ne publient de mesures de la qualité des données.

En conséquence, les données publiées par les programmes statistiques des catégories 3 et 4 et qui ne sont ni assorties d'un indicateur ni supprimées peuvent implicitement être utilisées sans qu'on ait à se soucier de leur qualité. En d'autres termes, ces données sont aptes à être utilisées pour n'importe quelle fin, tandis que les données assorties d'un indicateur sont aptes à être utilisées à certaines fins et que les données supprimées ne sont pas aptes à être utilisées. Pour certaines enquêtes, ces restrictions sont exposées plus ou moins explicitement dans les publications.

5. EFFET DES CRITÈRES DE DIFFUSION

5.1 Étendue de la suppression de cases

L'importance ou l'importance éventuelle des critères de diffusion représente plus qu'un simple sujet de discussion philosophique. Lorsque ces critères sont fondés sur les principales composantes de l'erreur, leur effet peut donner une indication claire et facilement intelligible de la proportion des données d'un programme statistique qui sont aptes à être utilisées selon le concepteur de l'enquête. Dans certains cas, les critères de diffusion n'ont que peu d'effet visible sur les données publiées dans les publications répertoriées au catalogue. Dans d'autres cas, la majorité des données de certains tableaux sont supprimées.

Le tableau 1 indique l'effet quantifié de la suppression de cases pour certaines publications relatives à 4 enquêtes distinctes. Ces enquêtes sont toutes des enquêtes-échantillon et les critères utilisés sont fondés sur le coefficient de variation. Lorsqu'on examine ces résultats, on peut supposer que les tableaux tels que définis correspondent aux besoins des utilisateurs et sont conformes aux normes du plan d'enquête. Les publications étudiées aux fins de cette évaluation présentent certaines différences sur le plan des stratégies de présentation des données et des indices ou mesures de la qualité des données. Ces différences sont exposées dans la dernière colonne du tableau 1.

La proportion de cases supprimées dans ces publications varie d'un tableau à l'autre, mais surtout pour les tableaux plus courts (nombre de pages) comme ceux de la publication de l'enquête sur la population active. Dans le cas de cette enquête, la proportion de cases supprimées varie de 0%, pour les principaux tableaux de données en vue desquels l'enquête a été conçue (Singh et coll., 1984), à 45% pour les tableaux plus détaillés. Bien sûr, la proportion de cases supprimées dans les tableaux croisés où le niveau de détail est le plus élevé est

beaucoup plus élevée. Dans la publication de l'enquête sur la population active, 16% des 8,635 cases détaillées sont supprimés, le taux de suppression par tableau s'échelonnant entre 0% et 71%.

Tableau 1: Proportion de cases supprimées en raison d'une erreur relative d'échantillonnage élevée - certaines enquêtes et publications

Enquête Publication	Répartition géographique	Cases supprimées Total	%	Pratiques de présentation des données
Enquête sur la population active[1]	Canada et provinces	18,200	9%	- Le niveau de détail des tableaux varie selon la taille de la province - La plupart des tableaux comprennent des indices de la qualité des données (erreur d'échantillonnage)
Enquête sur la santé et les limitations d'activité[2]	Canada, provinces et territoires	39,700	18%	- Le niveau de détail des tableaux ne varie pas par province ou par région infraprovincial - Les cases jugées de qualité médiocre sont assorties d'un indicateur. Aucun autre indice de l'erreur d'échantillonnage pour les données publiées
Enquête sociale générale[3]	Canada et certaines provinces	11,900	36%	- Indices de l'erreur d'échantillonnage pour les données publiées sont les intervalles "de 16.5% à 33%" et "moins de 16.5%"
Enquête sur l'équipement ménager[4]	Canada et Provinces	3,300	13%	- Le niveau de détail des tableaux ne varie pas selon la province - La plupart des tableaux comprennent des indices de la qualité des données (erreur d'échantillonnage)

[1] La population active, n° 71-001 au catalogue, mars 1990. Le pourcentage de cases supprimées tient compte des cases supprimées en raison de l'erreur d'arrondi. Les critères de diffusion sont conformes aux critères exposés dans le texte.

[2] Faits saillants: personnes ayant une incapacité au Canada, n° 82-602 au catalogue, mars 1990 et Données infraprovinciales - infraterritoriales pour la Nouvelle-Écosse, l'Ontario et la Colombie-Britannique, n°s 82-605, 82-608 et 82-612, mars 1989. Les critères de diffusion sont conformes aux critères exposés dans le texte. Contrairement aux autres enquêtes étudiées ici, les cases à valeur zéro ne sont pas supprimées.

[3] Profil de la victimisation au Canada, n° 11-612 # 2 au catalogue, mars 1990. Le coefficient de variation limite utilisé comme critère de diffusion est de 33%.

[4] Équipement ménager 1990, n° 64-202 au catalogue, octobre 1990. Le coefficient de variation limite utilisé comme critère de diffusion est de 25%. On utilise en outre un critère de diffusion fondé sur l'erreur d'arrondi semblable à celui de l'enquête sur la population active. Le pourcentage de cases supprimées tient compte des cases supprimées en raison de l'erreur d'arrondi et des cases à valeur zéro.

L'enquête sur la population active (EPA) est particulièrement utile pour illustrer l'effet des règles de suppression et la proportion des données qui ont été jugées inaptes à être publiées, même pour des enquêtes dont l'échantillon est d'une taille relativement importante. Les principaux utilisateurs consultés à cet égard ont convenu que des tableaux plus détaillés ne seraient pas aptes à être utilisés en raison de la piètre qualité des

données. Bien que certains utilisateurs se soient opposés à ce qu'on supprime des cases dans certains tableaux, même eux ont uniquement suggéré qu'on modifie les spécifications relatives aux tableaux (Division de l'évaluation des programmes, 1988). Par ailleurs, il a été impossible de déterminer l'aptitude des cases supprimées à être utilisées, du point de vue de tous les utilisateurs, ni la valeur de cette suppression.

5.2 Effet éventuel de la suppression de cases

Quel serait l'effet de l'application de la suppression de cases à des enquêtes non dotées de critères de diffusion? Pour le savoir, on a quantifié l'effet éventuel de la suppression de cases sur un choix de tableaux publiés à partir des données du recensement de la population de 1986. À cette fin, on a utilisé des publications de données-échantillon et un coefficient de variation limite d'environ 33% (soit presque le même que pour l'EPA) comme critère de diffusion. Le choix du recensement de la population s'explique du fait que l'échantillon de ce dernier est important (environ 20% des ménages canadiens) et qu'il permet de disposer d'un large éventail de tableaux croisés détaillés et de données régionales. Les taux de suppression ont été calculés avec et sans les cases en blanc ("-"). Ces cases représentent soit une valeur de zéro soit une valeur plus petite que 5 ramenée à zéro par arrondissement aléatoire (procédure visant à assurer la protection du caractère confidentiel des données). Comme on peut s'y attendre, la quantité de données qui seraient supprimées varie énormément d'un tableau et d'une publication à l'autre.

La publication *Activité* (n° 93-111 au cat.) contient des données semblables à celles de l'EPA. Or, 16% des 80,000 cases de données de cette publication seraient supprimées selon le critère fixant la limite du coefficient de variation à 33%. Si on tient compte uniquement des cases n'ayant pas déjà été publiées en blanc ("-"), ce pourcentage serait de 12%. Si on appliquait la limite de 33% à un tableau de données sur les secteurs de recensement pour la région métropolitaine de recensement de Victoria, en Colombie-Britannique, 20% des 15,600 cases de données (17% des cases qui ne sont pas déjà en blanc "-") seraient supprimées (*Profils: Secteurs de recensement, Victoria: Partie 2*, tableau 1, n° 95-170 au cat.). Pour certaines caractéristiques sociales présentant des distributions asymétriques, ce pourcentage serait beaucoup plus élevé. Ainsi, dans le cas du tableau comportant 41,100 cases de données non nulles portant sur la langue maternelle et la langue parlée à la maison (*Rétention et transfert linguistiques*, tableau 2, n° 93-153 au cat.), 64% des cases de données seraient supprimées en vertu de la limite de 33%. Dans ce dernier cas, le pourcentage de cases supprimées serait beaucoup plus faible (36%) si on tenait uniquement compte des cases qui ne sont pas déjà en blanc "-".

Encore une fois, lorsqu'on examine ces résultats, on peut supposer que les tableaux tels que définis correspondent aux besoins des utilisateurs et sont conformes aux normes du plan d'enquête. Aussi, faut-il en conclure que:

1/ nombre d'éléments d'information de piètre qualité répondent aux critères des utilisateurs en matière d'aptitude à l'utilisation; ou
2/ les matrices hiérarchiques rectangulaires types des tableaux constituent parfois une méthode inefficace de diffusion des données aptes à être utilisées.

Dans le premier cas, la suppression aurait un effet défavorable pour les utilisateurs. Dans le second, il serait désavantageux d'utiliser la méthode de diffusion classique pour les tableaux dans lesquels des données ont été supprimées.

6. ARGUMENTS MILITANT POUR ET CONTRE L'UTILISATION DE CRITÈRES DE DIFFUSION

En résumé, on peut dire que les questions générales exposées à la section 3 portent sur:

1/ la valeur et l'à-propos des critères de diffusion du point de vue de l'utilisateur et du Bureau;
2/ la nature ou la forme des critères de diffusion pouvant ou devant être appliqués pour les divers programmes statistiques;

compte tenu du programme statistique visé, des données, des formes d'erreur et du support de diffusion.

On trouve ci-après un résumé de certains des arguments militant en faveur et en défaveur de l'utilisation de critères de diffusion.

POUR	CONTRE
* Démonstration qu'il existe des normes de qualité et qu'elles sont appliquées	* Contrôle de l'information
* Réputation et responsabilité du Bureau	* Les données sont recueillies dans le cadre de programmes statistiques financés par l'État
* Organisme statistique a le savoir-faire nécessaire pour juger de l'aptitude à l'utilisation	* Les critères sont arbitraires
* Utilisateur considère qu'il revient au Bureau d'assurer que les données diffusées sont aptes à être utilisées	* Tous les chiffres ont une certaine valeur
* Risque d'emploi abusif des données confidentielles	* Incompatible avec le plan d'enquête
* Amélioration des spécifications relatives aux tableaux	* Rend l'analyse plus difficile
* Tous les tableaux croisés pouvant être établis ne doivent pas être considérés comme faisant partie du champ d'observation de l'enquête	* Nombre des cases supprimées peuvent être obtenues autrement
* La suppression de cases fondée sur la variance n'entraîne aucune perte d'information	* Les mesures de la qualité des données sont suffisantes
	* Fausse image de la fiabilité des données non supprimées
	* Les utilisateurs ne veulent pas payer pour des données supprimées

Le lecteur aura constaté que chacun de ces arguments ne s'applique pas nécessairement à tous les aspects des questions étudiées.

6.1 Valeur et à-propos des critères de diffusion

Les critères de diffusion représentent une norme de qualité. Selon certains, les producteurs de produits statistiques sont également des utilisateurs et ont donc la compétence voulue pour déterminer ce qu'il vaut la peine de diffuser. Ils disposent aussi de renseignements plus nombreux et plus détaillés sur les limites des données. Par ailleurs, il est nécessaire que l'utilisateur externe puisse tenir pour acquis que le produit respecte certaines normes de qualité de base. Le jugement professionnel nécessaire pour s'en assurer revêt une importance capitale au stade de la conception des programmes statistiques. Les décisions relatives à des éléments comme la taille de l'échantillon utilisé pour un programme statistique sont fondées en partie sur des normes qualitatives. Il devrait s'ensuivre qu'il est approprié et nécessaire que le Bureau applique de telles normes de conception à toutes les étapes, y compris celle de la diffusion. À la lumière de cet argument, l'aptitude à l'utilisation ne doit pas et ne peut pas être assimilée à l'accessibilité.

Cependant, il est également possible d'affirmer que l'aptitude à l'utilisation ne peut être assimilée à la qualité; en tout cas, certainement pas à la qualité exprimée sous forme de coefficient de variation. Il est reconnu que, dans le cas de nombreuses enquêtes, les utilisateurs veulent obtenir un large éventail de données fiables. Bien qu'il soit peut-être possible de produire nombre de ces données au moyen d'un plan d'enquête approprié (Cowan, 1988), on ne connaît ni tous les utilisateurs ni toutes les utilisations éventuels au moment de l'élaboration du plan et on ne peut s'attendre à ce que toutes les demandes de données tiennent compte des contraintes et des limites relatives à la qualité propres au programme statistique. Malgré ces limites, il peut arriver que le programme statistique visé constitue la seule source de laquelle l'utilisateur puisse obtenir les données dont il a besoin. Il est peut-être préférable de disposer de données non fiables plutôt que de ne disposer d'aucune donnée ou d'en être réduit à la devinette.

En outre, les utilisateurs de données qui sont le moindrement familiers avec le processus d'enquête ne s'attendront pas à ce que toutes les données soient de qualité élevée. On peut présumer que c'est à la lumière de leur expérience avec d'autres données du Bureau destinées à une utilisation à peu près comparable que les utilisateurs détermineront ce qu'ils entendent par de "bonnes" données. Pour juger de la qualité des données, l'utilisateur n'a besoin que de cette expérience, des données et de mesures de la qualité des données. Selon cet argument, il est tout à fait inutile de supprimer des données.

6.2 Forme des critères de diffusion

En principe, la méthode de suppression doit permettre de réduire au minimum les pertes de données pour l'utilisateur, sans limiter l'utilisation appropriée des autres données. Tant que les données sont présentées sous la forme de matrices hiérarchiques rectangulaires types sans suppression de cases, il faut diffuser certaines données non fiables ou supprimer certaines données fiables. En pareil cas, on peut considérer que la suppression de cases représente une solution de compromis raisonnable. Cependant, on établit une nette distinction entre l'à-propos des restrictions imposées par l'intermédiaire des spécifications relatives aux tableaux et celui de la suppression des cases. On peut faire valoir que les premières sont nécessaires pour des raisons d'économie et d'efficience, et qu'on ne peut s'y soustraire pour la plupart des formes de produits.

En revanche, on peut considérer la suppression de cases comme une mesure paternaliste qui ne sert aucune fin d'économie ni d'efficience; surtout quand on sait que les données supprimées peuvent souvent être obtenues par déduction. Il peut arriver que seules les données aptes à toutes les utilisations ne soient pas supprimées et que certaines données aptes à une utilisation particulière soient supprimées. En outre, la suppression de cases est incompatible avec l'idée selon laquelle les données et les mesures de qualité sont inséparables. Compte tenu de la définition que l'on donne de l'aptitude à l'utilisation, on pourrait prétendre qu'il est préférable de fournir aux utilisateurs les données et les mesures de la qualité, puis de les laisser déterminer eux-mêmes si les données sont aptes à être utilisées pour une fin particulière. Enfin, certains auteurs affirment que nombre d'utilisateurs ne se préoccupent aucunement de la qualité des données (Keane, 1987). Il est possible de faire valoir qu'il faut empêcher ces utilisateurs de faire un usage abusif des données, justifiant du même coup le recours à la suppression de cases. Si les données peuvent être obtenues par déduction, on peut user de "persuasion" pour éviter que les données supprimées soient utilisées. Toutefois, la suppression de cases n'empêche pas les utilisateurs de faire un usage abusif des données de qualité médiocre ou de meilleure qualité.

6.3 Variance et biais

Le type d'erreur considéré est un facteur déterminant pour les arguments donnés pour ou contre la suppression, surtout dans le cas de la suppression de cases. Si on fait exception des restrictions imposées par l'intermédiaire des spécifications relatives aux tableaux, certains soutiennent qu'aucune donnée ne devrait être supprimée en raison d'une variance élevée. Selon eux, cette mesure équivaut à décider, sans consulter les utilisateurs, quelles données sont si peu fiables qu'elles ne sauraient être utiles à aucun utilisateur ni servir à aucune fin. La plupart des cases non fiables sont de valeur relativement faible. Cette donnée constitue un renseignement en soi, la valeur exacte de la case n'a aucune importance (Bureau of the Census, 1975). Cela laisse supposer que l'utilisateur ne peut porter un jugement éclairé sur la qualité et l'utilité des données tout en utilisant des critères arbitraires.

On peut affirmer qu'on devrait permettre le recours à la suppression dans tous les produits et que la suppression constitue même le seul recours possible pour les cases biaisées ou éventuellement biaisées. Un tel biais peut découler, par exemple, d'une erreur de réponse systématique ou d'un taux de non-réponse élevé pour une question. Il est possible que certaines données biaisées n'aient aucune valeur, même comme indicateur général de niveau. Ainsi, il pourrait se révéler nuisible d'ajouter foi à de telles données en les publiant.

Par ailleurs, on peut faire valoir que l'estimation du biais peut s'avérer difficile et que l'établissement d'un critère de suppression est en outre arbitraire. Il peut arriver que le cas d'un biais de 1,000% pour une case à valeur élevée soit clair, mais que celui d'un biais de 50% pour une case de faible valeur ne le soit pas. Cependant, le vrai problème ne réside pas dans la possibilité de mesurer le biais. S'il est possible d'estimer l'importance du biais, il convient peut-être mieux d'en informer l'utilisateur plutôt que de supprimer les données. Ce sont les biais pour lesquels il est impossible d'estimer des limites supérieure et inférieure raisonnables qui font problème. En pareil cas, il peut se révéler impossible de fournir aux utilisateurs des indices de qualité leur permettant de déterminer si les données sont aptes à être utilisées. La suppression constitue peut-être alors la seule mesure responsable à prendre pour certains supports de diffusion.

6.4 Nature des données

On peut soutenir qu'il existe certains cas où l'usage abusif de données non fiables peut avoir de graves conséquences pour la réputation du Bureau et pour l'utilisateur. En pareil cas, il est plus prudent de supprimer les données. Même pour des données moins délicates, le fait de supprimer des données dans les publications peut permettre d'éviter que l'utilisateur occasionnel ou moins sophistiqué prenne d'importantes décisions en se fondant sur des données très peu fiables.

Certaines données possèdent un statut spécial à titre de ce qu'on peut nommer "statistiques officielles". Figurent au nombre de ces données les données dont il est fait état dans les lois fédérales et celles qui ont acquis un statut spécial du fait de l'utilisation qu'on en a historiquement fait: il en est ainsi des chiffres de population du recensement, des taux de chômage, de l'indice des prix à la consommation et des statistiques sur le commerce de marchandises. Il est possible de soutenir qu'il convient de protéger le statut de ces données en veillant à ce que seules les données valables soient diffusées.

Il existe également une différence entre l'utilisation des données de Statistique Canada pour ce qu'on peut vaguement définir comme des fins de recherche et leur utilisation pour les fins de l'élaboration de politiques. Dans le dernier cas, le décideur ne voit pas nécessairement les données publiées et il ne sait pas quelles sont les limites relatives à la qualité ni les concepts exacts utilisés. Il est possible que la personne qui rassemble l'information documentaire ou établit le rapport sur lequel la décision sera fondée se soit sentie poussée à se préoccuper uniquement d'obtenir un "chiffre" de Statistique Canada, sans se soucier de sa qualité. On peut soutenir qu'il est essentiel d'éviter de transmettre des données non fiables aux décideurs des secteurs public et privé.

Par ailleurs, il est possible de faire valoir que, en général, les données utilisées à des fins politiques importantes ne font pas partie des données susceptibles d'être supprimées. Compte tenu de certaines des exigences des utilisateurs (p. ex., D'Costa et coll., 1989), cette opinion est discutable. De plus, il semble qu'il est maintenant possible qu'on fasse un usage abusif des données. En effet, les critères de diffusion en vigueur permettent la publication de données qui pourraient être considérées comme très peu fiables si on devait les utiliser à des fins politiques. En outre, il n'existe que de rares antécédents d'usage abusif de statistiques "officielles" non fiables. Cependant, en ce domaine tout est relatif et fonction de l'utilisateur, des données et des fins pour lesquelles ces données sont utilisées. Actuellement, les données les moins fiables ne sont pas diffusées pour éviter qu'on en fasse un usage abusif, ou ne sont pas diffusées dans des délais ou sous une forme permettant qu'on en fasse un usage abusif. Pourtant, si l'on se fie sur le nombre de demandes relatives à certaines de ces données, leur utilité possible ne saurait être mise en doute.

6.5 Support de diffusion

Il semble qu'on favorise davantage l'utilisation de critères de diffusion dans le cas des publications répertoriées au catalogue. On favorise beaucoup moins l'utilisation de ces critères pour les autres supports de diffusion et les critères de diffusion alors appliqués sont certainement moins restrictifs. Selon certains, il faut avoir recours aux spécifications relatives aux tableaux pour réduire au minimum le nombre de données de piètre qualité diffusées dans les produits normalisés; on ne devrait utiliser la suppression de cases pour aucun type de support, sauf lorsqu'il y a risque de biais important; enfin, on ne devrait soumettre les tableaux normalisés à aucun critère relatif à la qualité des données diffusées et il faudrait se contenter d'aviser l'utilisateur de la qualité des données. Néanmoins, on considère que l'imposition, pour l'ensemble des supports de diffusion, d'une certaine forme de restrictions relatives à la publication secondaire des données ou à leur utilisation dans des rapports pourrait constituer une exception viable et appropriée.

Si on fait exception de ces dernières restrictions, on peut douter que même les critères les plus libéraux en matière de qualité des données diffusées puissent être appliqués aux véhicules de diffusion du futur immédiat. La diffusion électronique des données gagne en importance. De même, on observe une augmentation de la demande de données régionales et de données recueillies dans le cadre de divers programmes statistiques afin de pouvoir les intégrer. Seules la nécessité de protéger le caractère confidentiel des renseignements personnels et l'accessibilité de la technologie nécessaire pour traiter et analyser les données à peu de frais peuvent imposer des limites à l'accroissement de ces demandes. Or, il est peu probable que l'aspect technologique constitue un

problème, il serait plus plausible qu'il exerce une pression à la hausse sur la demande de données. Les utilisateurs de données régionales diffusées électroniquement ne voudront sûrement pas que l'utilisation de critères de diffusion viennent limiter le nombre de données auxquelles ils ont accès.

7. CONCLUSIONS

Dans les limites de la loi et des règlements de l'administration fédérale, il revient à chaque programme statistique de décider s'il doit utiliser des critères relatifs à la qualité des données diffusées et, le cas échéant, quels doivent être ces critères. Aucun intervenant n'a jugé qu'il y avait lieu, dans le sens le plus général, de modifier cet état de fait. En revanche, nous avons pu relever des arguments contre certaines applications et contre les critères utilisés, ainsi que contre l'absence de critères de diffusion.

Il semble qu'il soit nécessaire d'utiliser des critères de diffusion pour la conception des tableaux devant figurer dans les produits normalisés. De fait, il y a peut-être lieu d'utiliser des critères plus rigoureux dans le cadre de certains programmes statistiques. Aucun argument décisif ne milite en faveur de l'utilisation de la suppression de cases fondée sur la variance. Par ailleurs, il semble acceptable d'utiliser la suppression de cases fondée sur le biais lorsque l'erreur est relativement importante ou susceptible de l'être. Il faut étudier la possibilité d'imposer, pour tous les types de produits, des restrictions relatives à la publication secondaire des données et à leur utilisation dans des rapports, restrictions exigeant au moins qu'on indique quelle est la qualité des données lorsque cette dernière laisse à désirer. Autrement, il semble que rien ne justifie l'imposition d'autres restrictions à la diffusion des données dans les produits personnalisés, à l'exception possible des "statistiques officielles". Toutefois, si les "statistiques officielles" doivent être soumises à des critères relatifs à la qualité des données diffusées, il serait plus approprié que ces données et les normes de conception des produits soient définies dans le cadre d'une politique.

Il se peut que le Bureau se voit dans l'obligation de fournir aux utilisateurs un plus grand nombre de mesures de la qualité afin qu'ils puissent évaluer l'aptitude des données à être utilisées. Il est possible que l'établissement de telles mesures s'avère difficile et crucial dans le cas des données régionales et des données intégrées.

BIBLIOGRAPHIE

Australian Bureau of Statistics (1990). *May 1990: The Labour Force, Australia*, Catalogue No. 6203.0, Australian Bureau of Statistics.

Bureau of the Census (1975). *Standards for Discussion and Presentation of Errors in Survey and Census Data*, Bureau of the Census, U.S. Department of Commerce.

Bureau of Labor Statistics (1988). *Geographic Profile of Employment and Unemployment*: 1987, Bulletin 2305, Bureau of Labor Statistics, U.S. Department of Labor.

Bureau of Labor Statistics (1990). *Employment and Earnings*, Volume 37 Number 7, Bureau of Labor Statistics, U.S. Department of Labor.

Cowan, C.D., et Malec, D.J. (1988). *Sample allocation for A Multistage, Multilevel, Multivariate Survey*, Proceedings of the Fourth Annual Research Conference, Bureau of the Census, U.S. Department of Commerce.

Department of Statistics (1990). *Household Labour Force Survey: The New Zealand Labour Force, December 1989 Quarter*, Cat. No. 05.002, Department of Statistics, Wellington, New Zealand.

D'Costa, R., et Kennedy, L.W. (1989). *Evaluation Report: Demography*, Division de l'Évaluation de programmes, Statistique Canada.

Juran, J.M., et Gryna, F.M. Jr. (1980). *Quality Planning and Analysis*, McGraw-Hill Book Company, New York.

Keane, John G. (1987). *Surveying Survey Development: Some Stage-Setting Perspectives*, Proceedings of the Third Annual Reseach Conference, Bureau of the Census, U.S. Department of Commerce.

Methods and Standards Committee (1990). *Standards and Guidelines on the Presentation of Data in Tables of Statistical Publications*, Statistique Canada.

Office of Management and Budget (1988). *Guidelines for Federal Statistical Activities*, Federal Register, Volume 53, Number 12, Office of the Federal Register, National Archives and Records Administration, Washington, D.C.

Program Evaluation Division (1988). *Statistics Canada Labour Data: Identifying the Issues*, Statistique Canada.

Shapiro, G., Pollock, J., Harley, D., et Beach, M.E. (1985). *Statistical Standards at the Census Bureau; Past, Present, and Future*, Proceedings of the First Annual Reseach Conference, Bureau of the Census, U.S. Department of Commerce.

Singh, M.P., Drew, J.D., et Choudhry, G.H. (1984). Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981, *Techniques d'enquête*, 10, 2, 139-154, Statistique Canada.

Taillon, J. (1990). *Monthly Gross Domestic Product by Industry: Quality Assessment*, Gross Domestic Product by Industry, Catalogue 15-001, Statistique Canada.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

LES STATISTIQUES, L'ÉQUILIBRE ENTRE LA PRÉCISION À ATTEINDRE ET L'UTILISATION VISÉE

B.L. Khuong[1]

RÉSUMÉ

Cette présentation tente de décrire les situations que vit actuellement le BSQ à la suite des compressions continues de ses ressources, et identifie quelques avenues possibles que les statisticiens du BSQ auraient probablement avantage à emprunter afin de pouvoir répondre plus efficacement aux besoins de plus en plus variés et exigeants de sa clientèle.

Selon l'approche habituelle, les statisticiens tentent, avec la collaboration d'autres spécialistes, de prévoir les besoins en information avant que ceux-ci ne soient exprimés; dans ce contexte, ce sont donc généralement les statistiques qui suscitent les besoins en information.

Depuis quelques années toutefois, particulièrement au BSQ, la diminution des ressources entraîne dans son sillage la réduction de nouvelles activités, afin que les ressources disponibles puissent se concentrer sur le raffinement des connaissances dans les domaines déjà couverts. Paradoxalement, la réduction généralisée des ressources génère un plus grand besoin d'informations. Cette fois-ci, les besoins sont créés par l'utilisateur et ce dernier est prêt à payer, en autant que l'information générée réponde à ses besoins spécifiques. Pour faire face à cette nouvelle situation, les statisticiens auraient-ils avantage à changer leur approche? Devraient-ils aborder certaines demandes en orientant leur analyse sur les problèmes eux-mêmes et sur la nature particulière des besoins, plutôt que sur les aspects plus théoriques de la technique à utiliser?

MOTS CLÉS: Produits statistiques; précision; statistiques de la "qualité optimale".

1. PROBLÉMATIQUE

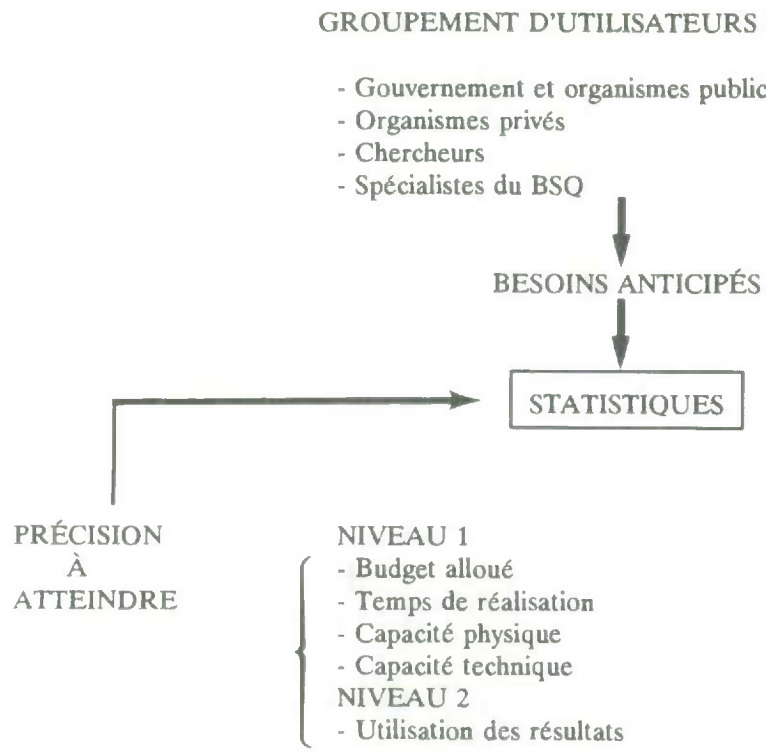
La statistique, comme toute autre discipline scientifique, possède ses propres critères de rigueur qui l'obligent à donner des preuves effectives et chiffrées de sa scientificité et, en particulier, de la précision de ses chiffres. Ces notions de rigueur, de précision, etc. sont imprégnées dans la mentalité des statisticiens durant leur formation universitaire.

Lorsque ces jeunes entament leur carrière au Bureau de la statistique du Québec, leur inquiétude dans la production des statistiques de haute précision est encore plus grande face aux attentes d'une grande partie du public et à la croyance fondée ou non, que les statistiques des bureaux de statistiques sont quasi infaillibles. Ainsi, leur principale préoccupation est concentrée dans le développement des outils ou de techniques qui leur permettront d'améliorer constamment la précision des statistiques produites et ce, indépendamment de l'utilisation que l'on fait de ces dernières.

[1] B.L. Khuong, Directeur de la méthodologie, Bureau de la statistique du Québec, 117 rue St-André, 3e étage, Québec, Québec G1K 3Y3.

La figure 1 schématise le cheminement dit "approche traditionnelle" selon laquelle la préoccupation principale des statisticiens est orientée vers la théorie et la technique pour produire les statistiques les plus précises possibles.

Figure 1: APPROCHE TRADITIONNELLE



Cette obsession à porter attention surtout aux techniques qui permettent d'améliorer la précision n'est pas un mal en soi, car on entend assez souvent des reproches de la part des chercheurs universitaires et des grands utilisateurs du secteur privé adressés aux organismes centraux de statistiques (BSQ, Statistique Canada, BLS aux États-Unis) sur la qualité reprochable de certaines de leurs statistiques.

Dans le contexte actuel de réduction budgétaire généralisé, est-ce que l'on peut toujours utiliser la même approche face aux besoins spécifiques dans une société dont la complexité ne cesse de croître? Selon mes propres constatations, on fait face à deux types de besoins qui requièrent deux approches différentes.

Le premier type concerne les statistiques d'utilisation courante qui répondent à un certain besoin, mais elle ne sont pas assez précises pour une partie des utilisateurs spécialisés. Dans ce contexte, il est tout à fait approprié pour le Bureau de la statistique du Québec en tant qu'organisme central de statistiques, d'adapter l'approche traditionnelle c'est-à-dire celle dont la précision est dictée par la capacité physique, technique, le temps et le budget alloués sans égard à l'utilisation de ces statistiques. Dans ces circonstances, le principal problème à résoudre est celui de l'amélioration continuelle de la qualité.

Ces préoccupations sont déjà imprégnées dans la mentalité des statisticiens en place et de plus, elles sont déjà prévues dans la programmation du BSQ. Il faut de plus admettre qu'aucune nation ne possède des statistiques parfaites, car l'utilisation de ces statistiques d'intérêt général varie d'une personne à l'autre et évolue avec le temps, ce qui entraîne automatiquement un changement dans la précision exigée.

Je suis au BSQ depuis environ deux ans et j'ai eu souvent à faire face également à un deuxième type de besoin, celui qui consiste à fournir des informations spécifiques pour une décision ponctuelle. Dans cette circonstance, les statistiques d'utilisation générale, si précises soient-elles, ne répondent pas tout à fait à ces besoins spécifiques. Or, depuis l'adoption de la nouvelle loi sur le Bureau de la statistique en 1987, ce

dernier est autorisé à répondre également aux besoins spécifiques des utilisateurs en autant que ces derniers acceptent de défrayer les coûts engendrés. Pour accomplir cette nouvelle fonction, le BSQ joue alors le rôle de "consultant" en plus de son rôle habituel d'organisme central de statistiques. C'est dans l'accomplissement de ce rôle additionnel que "l'approche traditionnelle" présente un certain malaise pour répondre avec efficience à la demande.

Ce sont généralement des besoins d'informations spécifiques pour une décision ponctuelle et qui exigent une réponse dans un court laps de temps. En voici deux exemples concrets parmi tant d'autres, l'un dans le domaine des enquêtes, l'autre dans celui de l'analyse des données.

1) Le ministère de l'Énergie et des Ressources désire mesurer et analyser la demande de villégiature sur les terres publiques

Afin de bien estimer la variable décisionnelle, soit la proportion des ménages intéressés à la villégiature sur les terres publiques du Québec, la démarche systématique classique pour fournir au demandeur des données (en tant qu'organisme central de statistique) serait de mener une enquête auprès d'un échantillon de 1 504 000 ménages.

Compte tenu des implications budgétaires de l'implantation d'un tel projet, du coût d'obtention de l'information et du temps nécessaire pour l'obtenir, la décision optimale serait de ne pas effectuer l'enquête.

2) Le ministère de la Santé et des Services sociaux doit établir un critère qui lui permettrait de comparer la performance des hôpitaux

Ce critère serait basé sur le nombre moyen de nuitées passées à l'hôpital pour différentes opérations. Mais il arrive de temps en temps qu'une valeur extrême gonfle exagérément la moyenne de nuitées d'un hôpital. Afin d'éliminer cette valeur extrême, le MSSS a proposé d'utiliser l'hypothèse selon laquelle le nombre de nuitées suit une distribution normale. Et les valeurs qui dépassent alors sont considérées comme extrêmes et éliminées du calcul. Lorsque le problème est présenté aux statisticiens du BSQ, ces derniers émettent des doutes sur la validité de l'utilisation de la loi normale. Intuitivement, ils pensent qu'une distribution de POISSON serait plus appropriée, puisque le nombre de nuitées ressemble à un processus de service d'un phénomène de file d'attente. Après avoir passé plusieurs tests, aucune distribution (Poisson, Beta et X^2) ne correspond parfaitement à la distribution des observations selon le seuil fixé. Donc on ne peut rien faire, bien que ces distributions se rapprochent plus de la réalité que la distribution normale à moins d'investir en temps et argent plus que la valeur d'utilisation de ces résultats.

2. NIVEAU DE PRÉCISION À ATTEINDRE

Dans l'un ou l'autre des deux cas mentionnés ci-dessus, les preneurs de décisions ont besoin d'informations ou d'outils pour les aider à prendre une décision. Mais le fait de fixer trop haut le niveau de précision à respecter, les amènerait probablement à prendre une décision sans ces informations spécifiques et ce à cause du coût élevé et du temps requis pour obtenir une précision qui dépasse l'utilisation visée. Est-ce que l'absence d'informations est préférable à une information moins précise? Y a-t-il des alternatives? Cette dernière question m'amène à en poser une autre sur la qualité des données. Est-ce que le niveau de qualité des données doit être uniforme et statique, quelle que soit la situation? Selon moi, ce niveau ne doit être ni uniforme ni statique.

Il existe un niveau de qualité optimal pour répondre à une exigence donnée, celle-ci étant dictée par l'utilisation attendue. Or, cette utilisation attendue est spécifique et connue avec certitude dans certains cas, et ne l'est pas dans d'autres cas. Peut-on alors séparer les produits statistiques en fonction de ces 2 classes d'utilisation attendue, pour ensuite déterminer le seuil de qualité qui serait la précision à atteindre.

A- Produits destinés à des utilisateurs non spécifiques

Ce sont des produits statistiques qui s'adressent à un très grand nombre d'utilisateurs et qui font ou feront partie de la programmation du Bureau. Pour ce groupe de produits, il est primordial que le niveau de qualité soit uniforme, car ils répondent à une gamme très variée d'utilisations. Mais ce niveau ne doit pas être statique, car il doit évoluer en fonction de la meilleure précision que l'on peut atteindre. Les statisticiens doivent concentrer leurs efforts sur la théorie ou la technique qui leur permet d'améliorer la précision en tenant compte des quatre contraintes énoncées dans la figure.

B- Produits destinés à des utilisations spécifiques et ponctuelles

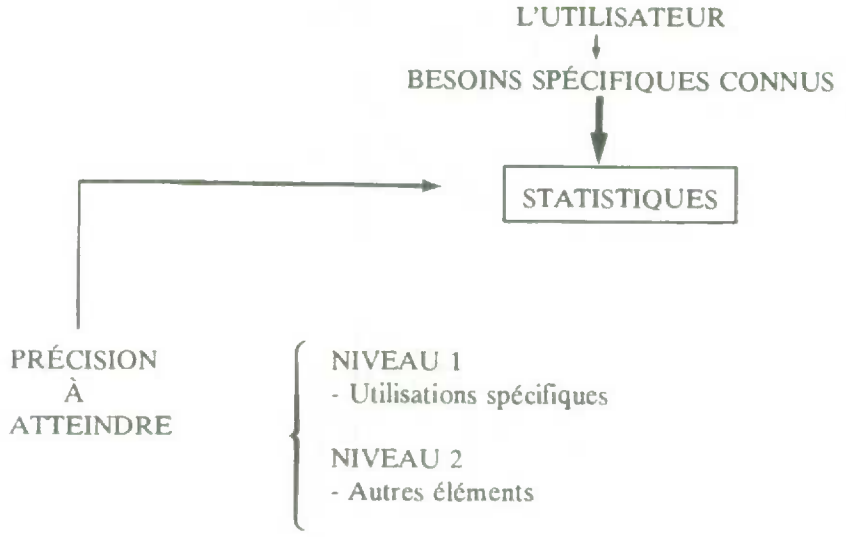
Pour ce deuxième groupe de produits, les statisticiens doivent se pencher davantage sur l'analyse du problème de l'utilisateur, que sur la technique permettant d'obtenir des données avec un niveau de précision identique à celui des produits du premier groupe. Ici, le niveau de précision à atteindre ne doit pas être uniforme et il dépend en particulier de l'utilisation spécifique attendue. Dans ce cas, les statisticiens "font de la statistique", plutôt que "produire des statistiques". Je n'ai pas besoin d'aborder la nuance entre la statistique et les statistiques car MM. Fellegi et Wilk l'ont décrite de façon très explicite dans leur article intitulé "Is Statistics Singular or Plural?"

Lorsque les statisticiens s'impliquent dans l'analyse du problème au-delà de la seule production de statistiques fiables, ils sont en mesure de déterminer de façon "optimale" le niveau de précision à atteindre, en tenant compte de l'utilisation visée. Ils participent activement à l'évaluation de l'impact du niveau de précision à atteindre sur la fiabilité et l'applicabilité des décisions prises à partir des statistiques fournies. Ainsi, ils font plus qu'émettre des mises en garde; ils voient eux-mêmes à ce que ces mises en garde soient prises en compte dans l'utilisation des résultats.

Dans le cas des produits du groupe B, les besoins d'informations pour une utilisation bien déterminée et ponctuelle sont présents. Si, pour des raisons de rigueur au niveau de la qualité à atteindre, les statisticiens se permettent de n'aborder les problèmes qu'à partir des considérations techniques et théoriques, les preneurs de décisions vont se pourvoir d'informations ailleurs. Or, je trouve que fournir des informations moins fiables accompagnées d'une évaluation de l'imperfection est préférable à laisser au preneur de décisions la tâche d'utiliser de l'information sans aucune évaluation de son imperfection.

La figure 2 ci-dessous schématise l'approche dite "ALTERNATIVE" selon laquelle la préoccupation principale des statisticiens est orientée vers la solution du problème ou la nature particulière de l'utilisation de ces statistiques.

Figure 2: L'APPROCHE "ALTERNATIVE"



3. CONCLUSION

En théorie, la notion de qualité des données ne pose apparemment pas de difficulté. En pratique, elle nous conduit à des questions aussi intéressantes et difficiles, que celle de la conciliation de l'exigence scientifique avec la nécessité de produire des données pertinentes à un problème spécifique, dans un court laps de temps et à un coût approprié. La détermination du niveau de qualité à atteindre est une évaluation de l'équilibre entre la précision à atteindre et l'utilisation visée. Dans toutes les situations, il existe deux erreurs qu'il faut éviter dans la conception de la démarche à suivre :

- recherche un niveau de qualité des données plus précis que l'utilisation prévue;

- viser un niveau de qualité pas assez précis pour l'utilisation prévue.

Ainsi, pour produire des statistiques de "QUALITÉ OPTIMALE", les statisticiens doivent au moins tenir compte des 3 éléments suivants:

- éviter d'être seulement "PRODUCTEURS DES STATISTIQUES";

- ne pas répondre à la demande mais plutôt aux BESOINS de l'utilisateur;

- se rappeler que l'utilisation visée est le facteur DÉTERMINANT de la précision à atteindre.

BIBLIOGRAPHIE

Bierman, H., Bonini, C., et Hausman, W. (1977). *Quantitative Analysis for Business Decision* (4e Éd.), Illinois, Richard D. IRWIN.

Deming, W.E. (1990). *Sample Design in Business Research*, New York, John Wiley & Sons.

Fellegi, I.P., et WILK, M.B. (1988). Is Statistics Singular or Plural?, *The Canadian Journal of Statistics*, 16.

SESSION 10

Assurance de la qualité

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

APPLICATION À STATISTIQUE CANADA DES TECHNIQUES D'AMÉLIORATION DE LA QUALITÉ MISES AU POINT DANS L'INDUSTRIE

D.N. Williams[1]

RÉSUMÉ

Les stratégies de développement et les techniques d'amélioration de la qualité mises au point dans l'industrie peuvent être appliquées sans difficulté dans des milieux de travail comme Statistique Canada, où elles peuvent s'avérer très efficaces. Dans cette communication, nous passons en revue et examinons les étapes qu'il faut suivre pour faire cette application, notamment la définition, la mesure et l'amélioration de la qualité. Nous insistons beaucoup sur les différences reconnues entre l'industrie et Statistique Canada en ce qui a trait aux méthodes de travail, par exemple: l'opposition services/production, court cycle de vie/long cycle de vie, forte intensité de main-d'oeuvre/forte intensité de capital (matériel de production), professionnels/cols bleus.

MOTS CLÉS: Satisfaction du client; comprendre le processus; contrôle du processus.

1. INTRODUCTION

Aujourd'hui, dans l'industrie, l'amélioration de la qualité est une priorité essentielle. Des chefs de file venus de tous les coins du monde viendront vous parler de la très grande importance qu'il faut lui accorder. Rendez-vous dans pratiquement n'importe quel service d'une organisation quelconque et demandez-leur si l'amélioration de la qualité progresse...

et vous verrez qu'on vous répondra: "Cela ne nous concerne pas... c'est plutôt l'affaire des responsables de la production".

Un des plus gros obstacles à l'implantation de mesures d'amélioration de la qualité dans une organisation donnée est justement la croyance, largement répandue, que "cela ne nous concerne pas, nous sommes différents, nous nous sommes déjà améliorés, c'est plutôt l'affaire des responsables de la production". Je l'ai entendu dire au moins une fois, de la part des gestionnaires comme des employés, dans chacune des organisations avec lesquelles j'ai travaillé, depuis le service de Recherche et développement, en passant par le Marketing et les Services, et jusqu'aux Opérations de soutien.

Cependant, les organisations qui souhaitaient vraiment améliorer la qualité ont découvert que cela ne concerne pas seulement ceux de l'étage de la production. C'est là que tout commence, mais nous nous sommes rendus compte que tous les aspects de la mise au point, de la production et de la vente d'un produit ou d'un service étaient étroitement reliés, au point que toutes les personnes intéressées doivent unir leurs efforts. L'amélioration de la qualité ne s'applique pas à un secteur plutôt qu'à un autre. Il est inutile que les responsables de la production essaient de réduire le plus grand nombre de défauts possible si le modèle est mal conçu au départ, si personne n'achète le produit ou encore si le produit est expédié n'importe comment.

Il n'existe pas de recettes toutes faites pour l'amélioration de la qualité. Compte tenu de la très grande diversité des contextes de travail, il a fallu adopter des stratégies et des techniques différentes dans chaque cas en vue d'améliorer la qualité. Toutefois, il y a un certain nombre de règles très générales qui donnent toujours de bons

---

[1] David N. Williams and Associates, 418 Edgewood Avenue, Ottawa, Ontario, K1Z 5K5.

résultats. J'aimerais discuter de ces règles avec vous aujourd'hui et que nous essayions de voir comment elles s'appliquent à Statistique Canada.

Il est possible que certaines d'entre elles vous soient familières ou que vous vous rendiez compte que vous appliquez déjà une ou plusieurs de ces règles dans vos domaines de travail. C'est bien, ces règles ne sont pas uniques, et elles ne sont pas non plus le résultat d'un travail de création considérable. Ce qui fait qu'elles sont différentes dans le cas de l'industrie, c'est qu'elles ont été clairement définies et que les entreprises, du moins celles qui veulent vraiment améliorer la qualité, essaient de les mettre en place dans l'ensemble de l'organisation, en les considérant comme un tout. Pas seulement une ou deux ici et là, mais toutes, en bloc. Utilisées de cette façon, elles sont beaucoup plus efficaces.

Règle n° 1: Qualité signifie satisfaction du client

La qualité est définie en fonction de la satisfaction du client; à cette fin, il faut par exemple tenter de comprendre les désirs et les besoins des clients, cerner de façon adéquate leurs attentes et ensuite livrer un produit conforme à ces attentes. Evidemment, à Statistique Canada, la qualité doit aussi tenir compte d'un certain nombre d'exigences précises visant à assurer l'intégrité des données qui sont fournies. Ces exigences internes influent grandement sur la délimitation des attentes des clients.

Quelques exemples de l'importance de bien déterminer les attentes des clients:

- La Société Honda vend huit modèles de voitures de base qui ne comportent que des différences mineures. Si quelqu'un se rendait chez Honda et disait: "Moi ce que j'aimerais vraiment avoir c'est un véhicule qui a du nerf doté d'un moteur V-8, qui loge huit personnes et qui passe de zéro à 90 kilomètres à l'heure en six secondes". Les représentants de Honda se feraient un plaisir de diriger cette personne vers le concessionnaire GM le plus proche. Les gens ne s'attendent pas à trouver une telle voiture chez Honda. Les gens s'attendent à ne pas avoir de problèmes avec leur voiture Honda. Si la voiture Honda a un problème, Honda a un client insatisfait. Lorsqu'on fixe des attentes et qu'on ne les atteint pas, on risque de mécontenter les clients et peut-être de les perdre à tout jamais.

- Il y a quelques années, Mitel avait beaucoup de difficultés à déterminer de façon adéquate les attentes de ses clients. L'entreprise avait tendance à promettre des choses qui n'existaient pas, c'est-à-dire des fonctions dont leurs systèmes de commutation n'étaient pas dotés. Les clients étaient très mécontents de la chose.

Qualité ne signifie par nécessairement dire au client ce qu'il veut entendre ou essayer de lui livrer à peu près tout ce à quoi il peut penser vouloir. Qualité signifie déterminer des attentes et ensuite livrer un produit qui satisfait ou dépasse ces attentes.

Essayez de réfléchir aux attentes de vos clients dans vos domaines de travail respectifs. Leurs attentes sont-elles réalistes, par exemple des données précises présentées dans quelques publications dans des délais raisonnables, ou au contraire s'attendent-ils à voir publier dès qu'ils en ont besoin (c'est-à-dire très rapidement) toutes les données qui existent dans le monde. Demandez-vous comment de telles attentes ont été fixées au départ.

Il n'est pas facile de déterminer des attentes. L'industrie dépense pour cela des millions de dollars par année. Il est encore plus difficile de changer des attentes et encore davantage de les restreindre. Lee Iacoca de Chrysler excellait à ce jeu et les gens le suivaient dans cette voie. D'autres ont moins de succès.

Si vous ne vous occupez pas activement de fixer les attentes de vos clients, ils s'en chargeront pour vous. Pendant longtemps, les gens ont eux-même décidé ce qu'ils attendaient du gouvernement. Les restrictions budgétaires obligent maintenant le gouvernement à envisager sérieusement de faire moins. Il est devenu essentiel de redéfinir les attentes. Il s'agit d'une mesure à la fois difficile et impopulaire.

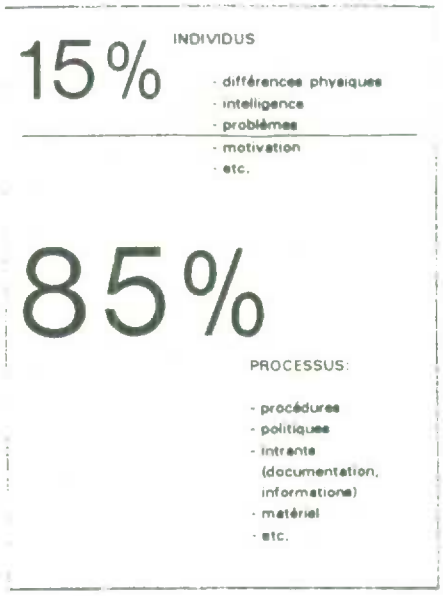
Il reste que si on ne répond pas aux attentes fixées pour les clients, ces derniers seront mécontents; ils se fâcheront et ne reviendront probablement plus jamais. Dans le cas du gouvernement, ils n'ont pas le choix de

revenir ou non, de sorte qu'ils seront seulement fâchés... et ils diront des choses très désagréables au sujet de votre service.

Règle nº 2: Mettre l'accent sur le processus

Alors comment peut-on améliorer la qualité? Ceci nous amène à la deuxième règle: l'unique façon de réaliser des gains de qualité durables consiste à améliorer les méthodes de travail elles-mêmes, de mettre l'accent sur le processus et non sur les gens.

Quand je fais de la formation de groupe, j'aime poser deux questions très difficiles aux participants: tout d'abord je leur demande d'imaginer une semaine de travail parfaite et ensuite de comparer avec une semaine de travail réelle. Ils doivent me faire part des différences observées. La semaine de travail réelle est remplie d'embêtements de toutes sortes, de problèmes urgents à régler, de priorités contradictoires, de clients mécontents, de délais manqués, et j'en passe. Les gens consacrent en moyenne entre 45% et 65% de leur temps de travail à s'occuper de ce genre de choses, ce que nous pourrions appeler du gaspillage. Non pas que les gens veulent gaspiller du temps. Le gaspillage est défini comme de vaines dépenses de temps ou de matériel. Les gens ne sont pas responsables du gaspillage, ils ne peuvent éviter ces problèmes, c'est le processus qui est en cause.



En règle générale, le processus est à l'origine de 85% de ces problèmes et les individus d'un autre 15%. Par processus, j'entends:

- les méthodes,
- la formation,
- les politiques,
- le matériel,
- les installations,
- l'échange de documents et d'informations.

Par individu, j'entends les différences individuelles, notamment les problèmes que chacun apporte de la maison, les différences sur le plan physique, santé, intelligence, *etc.*

Par le passé, les entreprises ont surtout misé leurs efforts d'amélioration de la qualité sur les individus, par des programmes de motivation, des menaces, de l'argent, *etc.* Des campagnes ont été lancées une fois par année avec des slogans cajoleurs à l'intention des employés du genre "Faites-le bien tout de suite" ou "Visez zéro défauts". Le principal résultat a été l'augmentation du moral des troupes une fois les affiches enlevées. D'autres entreprises ont essayé la méthode "pot-de-vin": "Si vous travaillez plus fort, mieux, etc., nous vous donnerons

PLUS D'ARGENT". Le principal résultat a été que les employés ont essayé d'en faire plus pendant un certain temps, mais essayer d'en faire plus ne peut durer toujours. On ne peut espérer acheter de tels regains d'attention que pour un court espace de temps. On s'est vite rendu compte que si le processus n'était pas changé, amélioré, ces mesures ne donnent pas de gains cumulatifs ou de gains à long terme.

Cela ne veut pas dire pour autant que les gens ne sont pas importants. Au contraire, sans eux il est impossible de faire marcher les choses ou d'améliorer le processus, ce qui nous amène à la règle suivante.

Règle n° 3: Les gens veulent faire du bon travail

Cette règle est une conséquence directe de la précédente. En éliminant les blocs qui obstruent le passage, nous permettons aux gens de faire du bon travail. Les entreprises ont dépensé des millions de dollars et des heures et des heures à essayer de trouver des façons d'accroître la motivation et la productivité en mettent l'accent uniquement sur les employés. Ce que nous essayons de démontrer ici, c'est que les entreprises ont peut-être besoin d'éliminer du processus les facteurs de démotivation, lesquels poussent les gens à négliger leur travail.

Il y a bien sûr des gens qui s'en fichent, qui font du mauvais travail. Mais peut-on les blâmer quand on sait que depuis des années le processus ne cesse de se mettre en travers de leur chemin et que personne n'a fait grand chose pour remédier au problème. Si nous devenons convaincus que les gens veulent faire du bon travail, nous pouvons consacrer notre esprit tout entier à la recherche de moyens pour améliorer le processus.

Un autre aspect de la question est que les employés sont les spécialistes dans le processus avec lequel ils travaillent. Pour être certain d'améliorer de façon efficace la qualité, le processus par exemple, vous devez faire participer les employés. Les chefs d'entreprises commencent à se rendre compte qu'ils doivent non seulement croire que leurs employés souhaitent faire du bon travail, mais aussi les former afin qu'ils puissent faire partie intégrante de tout projet d'amélioration, afin qu'ils puissent travailler activement à l'élimination des obstacles qui leur barrent le chemin. Tirez profit de votre meilleure ressource: vos employés.

Règle n° 4: Comprendre le mieux possible votre processus

Bon, nous avons compris que nous devons mettre l'accent sur le processus. Mais que doit-on faire pour cela?

Je me souviens d'une conversation que j'ai eue il y a un certain temps à trois heures du matin avec un gestionnaire mécontent fraîchement arrivé du Japon pour vérifier une exploitation. Ce dernier m'a dit: "Je ne sais pas pourquoi vous les Américains ne pouvez pas comprendre par vous-mêmes cette question de l'amélioration de la qualité. Vous essayez les cercles de qualité, sans succès, vous essayez les robots, encore une fois sans succès, quand donc allez-vous comprendre que l'amélioration de la qualité passe par la meilleure compréhension possible de votre processus. Si vous comprenez parfaitement bien votre processus, vous pouvez vous attendre à des améliorations."

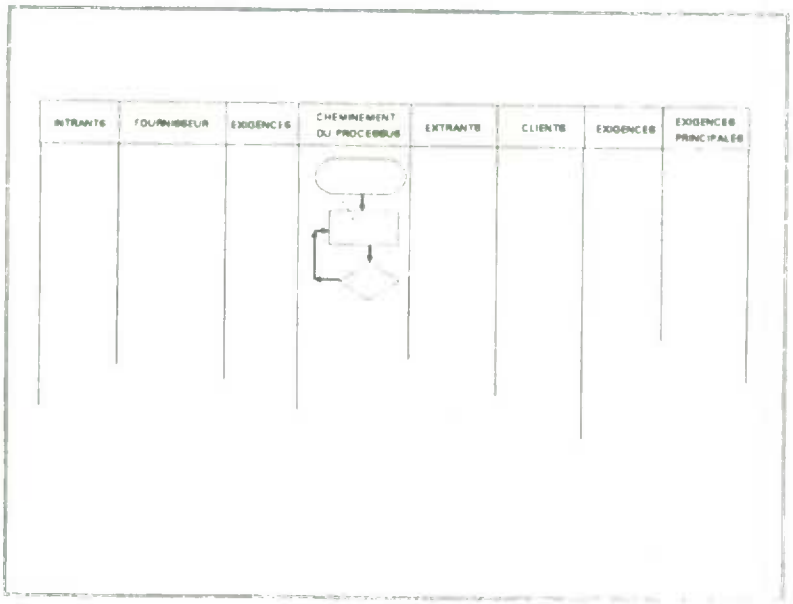
J'ai pensé pendant un certain temps à ce qu'il m'avait dit et ma réaction a été: "Ce n'est pas ça du tout, c'est beaucoup plus complexe que cela." Je comptais à ce moment-là deux ou trois années d'expérience dans mon domaine de travail actuel et j'étais convaincu que ce n'était pas si simple. Mais quand je me suis mis à réfléchir aux diverses choses que j'enseignais aux gens, les outils et les techniques, je me suis rendu compte qu'elles avaient toutes pour but de les renseigner davantage sur le processus.

Tout travail se fait dans le cadre d'un processus quelconque. Sans processus, c'est l'anarchie. Aussi, à ceux qui disent: "Cela ne nous concerne pas", je pose la question suivante: "Etes-vous certain qu'aucun de vos processus de travail ne pourrait être mieux compris?" Par comprendre, je veux dire comprendre tous les aspects du processus, non seulement son cheminement, mais aussi les rôles des employés, des clients, des fournisseurs et des gestionnaires.

Par clients et fournisseurs, j'entends ceux de l'intérieur comme de l'extérieur. Le processus qui consiste à concevoir et à mener même une enquête relativement simple peut exiger la collaboration de nombreuses

divisions différentes, chacune dépendant des intrants et des extrants des autres pour faire son travail. Sans une compréhension de la part de toutes les personnes concernées, fournisseurs et clients, des rôles de chacun et du fonctionnement du processus, comment peut-on s'attendre à ce que les employés produisent un travail de grande qualité dans les délais voulus?

Les deux premiers outils nécessaires à la compréhension d'un processus sont la définition du processus et son contrôle. L'établissement de diagrammes est particulièrement utile à la définition du processus, mais il faut faire plus que simplement décrire les principales étapes du processus, il faut aussi donner de l'information sur les intrants et les extrants et les exigences liées à chacun d'eux ainsi que sur les exigences générales associées à l'ensemble du processus. Grâce à une telle information, vous pouvez commencer à mieux saisir les interactions du processus et le degré de dépendance de chaque élément.



Il est difficile d'obtenir le consensus sur tout cela; les processus sont complexes et de nombreux services sont en cause. Et une fois que vous avez réussi à brosser un tableau du processus, il vous reste à répondre à une question plus difficile encore: "Ce processus fonctionne-t-il vraiment comme il le devrait? Répond-il à toutes les exigences et à toutes les attentes fixées? Si la réponse est "non", vous savez sur quoi faire porter en priorité vos efforts d'amélioration.

Le contrôle est très important à deux égards:

- si vous ne contrôlez pas la qualité du processus, vous ne pouvez pas l'améliorer. En d'autres termes, vous ne saurez pas qu'il y a quelque chose à améliorer et si des changements sont apportés, vous ne saurez pas s'il y a eu amélioration;

- établissement et mise à jour des priorités. Dans l'industrie, on mesure la productivité et les données financières et on rédige des rapports précis à leur sujet. Des objectifs sont ensuite fixés en fonction de ces mesures. On ne peut pas penser vraiment à améliorer la qualité sans d'abord la mesurer et rédiger des rapports détaillés sur la question.

Règle nº 5: Vous ne pouvez pas améliorer un processus sur lequel vous n'exercez aucun contrôle

Il s'agit d'une règle tirée du contrôle statistique des processus. Elle signifie que si un processus n'est ni prévisible, ni uniforme, il ne peut pas être beaucoup amélioré. A mon avis, c'est en l'améliorant qu'on rend un processus uniforme et prévisible, mais certaines personnes ne partagent peut-être pas ce point de vue.

La première étape dans tout projet d'amélioration est l'atteinte d'un degré raisonnable d'uniformité et de prévisibilité concernant le fonctionnement du processus, l'ordonnancement et la livraison des produits. Ce n'est pas là une tâche facile, des groupes de recherche et développement avec lesquels j'ai travaillé ont mis beaucoup de temps pour y arriver. Il importe de souligner que le respect constant des délais de production n'est qu'un aspect de la question, l'uniformité du processus de production est sans doute une tâche beaucoup plus importante.

J'aimerais vous citer un bon exemple, selon une expérience que j'ai vécue récemment avec un groupe de recherche et développement. Le processus semblait simple, le service de R&D était censé recevoir du service de marketing les spécifications du produit et ensuite devait le fabriquer en tenant compte de ces spécifications. Après avoir soigneusement étudié le processus, nous nous sommes rendus compte que l'information dont le service de R&D avait besoin était communiquée de façon sporadique, que le produit n'avaient pas toutes ses parties, que le format était chaque fois différent et que les données étaient loin d'être précises. Cet état de choses a eu pour résultats beaucoup de travail à refaire, des plans bons à jeter à la poubelle et des clients mécontents. Une des premières choses que nous avons faites a été de travailler avec le service de marketing à l'établissement de normes mutuellement acceptables pour ce produit. L'entreprise en question peut maintenant commencer à essayer de trouver des façons de répondre à ces normes.

Des exemples à Statistique Canada vous viennent-ils à l'esprit? Pensez à quel point il est difficile d'obtenir des spécifications uniformes sur les besoins des utilisateurs aux fins de conception d'un questionnaire ou comment il doit être frustrant pour des programmeurs/analystes informatiques d'écrire des programmes pour des analystes en statistique qui n'ont pas communiqué leurs besoins de façon claire.

Règle nº 6: La seule façon dont on peut améliorer de façon durable la qualité consiste à éliminer les causes de la mauvaise qualité

La compréhension du processus doit comprendre la compréhension des problèmes, jusqu'aux causes les plus profondes qui sont à l'origine des problèmes dans le processus. L'amélioration du processus signifie l'élimination permanente de ces causes de problèmes, l'une après l'autre.

Par le passé et pour un grand nombre encore, la seule façon dont les entreprises essayaient de vérifier et d'améliorer la qualité était par l'inspection du produit fini. Si une erreur était décelée, le produit était rejeté, retravaillé et quelqu'un pouvait se retrouver sur la liste noire de l'entreprise. Il s'agit d'un procédé des plus coûteux: il faut payer des gens pour produire, des gens pour trouver les erreurs, des gens pour corriger l'erreur et rehausser la qualité du produit.

Dans le secteur des services, c'est un luxe qu'on ne peut même pas s'offrir car on ne peut corriger un client insatisfait.

Règle nº 7: Plus de 65% des solutions se perdent

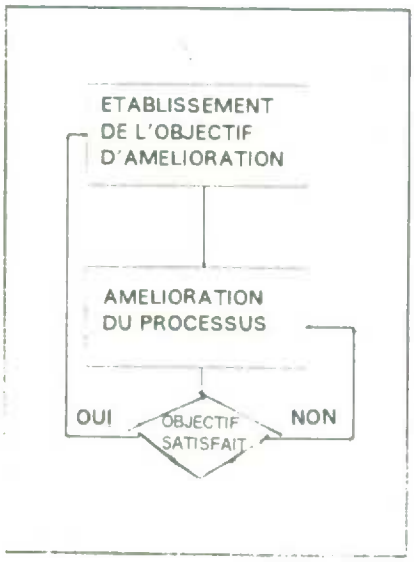
Qu'elles soient bonnes ou non, les solutions peuvent disparaître. Les raisons sont nombreuses, trop même pour que je tente de les énumérer aujourd'hui. Cette règle devait cependant être énoncée pour souligner qu'il ne suffit pas de comprendre les causes d'un problème et d'y trouver une solution, tout reste encore à faire.

Pour que les solutions puissent être mises en oeuvre de façon efficace, il faut faire en sorte qu'elles deviennent permanentes.

Protégez adéquatement les projets d'améliorations; ils peuvent disparaître sans même qu'on s'en rende compte.

Règle nº 8: Etre prêt à recommencer

Pour être efficace, l'amélioration de la qualité doit être continue. Elle doit être inscrite dans les fonctions de chaque employé, faire partie de leurs objectifs. On ne peut s'attendre à ce que quelqu'un réussisse tout du premier coup, mais on peut s'attendre à ce que ces employés travaillent sans cesse à l'amélioration des processus avec lesquels ils travaillent.



Lorsqu'on a atteint un objectif d'amélioration, il faut s'en fixer un autre, cela ne veut pas dire qu'on en a fini avec l'amélioration de la qualité pour l'année ou jusqu'à ce qu'on remanie l'enquête dans cinq ans. Cela signifie qu'on doit sans cesse analyser les processus avec lesquels on travaille ainsi que les mesures de contrôle de ces processus et essayer de les améliorer.

Règle nº 9: Si les gestionnaires ne dirigent pas le mouvement, on ne peut s'attendre à ce que les autres suivent

Il est essentiel que la direction fasse preuve de leadership et s'approprie le projet d'amélioration. Il s'agit d'une responsabilité qui ne peut être déléguée.

Ce sont là les règles qu'il faut essayer d'appliquer à l'ensemble des divisions ou services d'une organisation.

Comme je l'ai déjà souligné, il existe tout un ensemble d'outils et de techniques utiles à la compréhension et à l'amélioration de processus dont je n'ai pu parler en détail aujourd'hui. Un grand nombre de ceux-ci sont de nature statistique, comme les histogrammes, les graphiques de Pareto et les cartes de contrôle. On confond souvent l'amélioration de la qualité avec ces instruments et il est vrai qu'ils ne sont pas tous efficaces dans tous les domaines. Mais il ne faut pas pour autant négliger de faire en sorte de découvrir ceux qui pourraient être utiles.

J'aimerais, pour conclure, non pas vous citer une autre règle, mais tenter de répondre à une question souvent posée au sujet de l'amélioration de la qualité: "Que puis-je en tirer?" Les réponses ne manquent pas dans le cas de l'industrie, comme "la survie de l'entreprise" ou encore "votre emploi".

Ces réponses ne s'appliquent cependant pas toujours dans l'administration publique. Que pouvez-vous en tirer? Je serais porté à dire un plus grand contrôle sur les processus avec lesquels vous travaillez, une plus grande fierté de pouvoir mieux faire votre travail, un moins grand stress face aux embêtements et aux problèmes urgents déjà mentionnés, la capacité de pouvoir réaliser une quantité égale ou supérieure de produits et services, non pas en travaillant plus fort, mais grâce au fait de ne plus avoir à lutter contre le processus, à chaque étape, et par-dessus tout, la très grande fierté de pouvoir livrer un produit de meilleure qualité, votre propre produit ou service, ainsi que l'information essentielle que Statistique Canada fournit.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

PLANS DE CONTRÔLE BASÉS SUR LES SOMMES CUMULATIVES PONDÉRÉES ET LEURS APPLICATIONS

E. Yashchin[1]

RÉSUMÉ

Nous étudions une catégorie de cartes de contrôle pondérées qui peuvent être considérées comme une généralisation de la technique fondamentale de production de cartes à sommes cumulatives. La puissance statistique de ce type de plan est reliée à l'utilisation efficace que font ces derniers des valeurs basées sur les rapports de vraisemblance. La technique des sommes cumulatives pondérées (Cusum) de type 1 se révèle particulièrement utile quand on traite des cartes pour lesquelles la taille de l'échantillon varie. Les sommes cumulatives pondérées de type 2 sont utilisées pour améliorer le rendement statistique d'une carte à sommes cumulatives relativement aux variations de la moyenne du processus. Dans le présent document nous décrivons l'aspect graphique et l'aspect relatif à la prise de décision de la technique ainsi que certaines applications reliées à la surveillance de la qualité des données. Nous traitons aussi de certains types plus généraux de plans pondérés.

MOTS CLÉS: Longueur moyenne d'une série; cartes de contrôle; moyenne mobile à pondération exponentielle; rapport de vraisemblance; contrôle de la qualité.

1. INTRODUCTION

Soit X_1, X_2, ... une suite d'observations reliées à un certain processus, qui peut représenter, par exemple, des mesures, des moyennes d'échantillon, des écarts-types d'échantillon ou des proportions empiriques des erreurs trouvées dans des ensembles de données successifs. Dans la pratique, de telles suites sont habituellement surveillées à l'aide de cartes de contrôle. La partie décision d'une carte de contrôle représente un ensemble de critères servant à juger, à un moment quelconque, si le processus qui produit les observations se trouve dans des limites acceptables. Un tel ensemble est appelé un *plan de contrôle*. Les critères du rendement d'un plan de contrôle sont habituellement reliés au comportement de certaines caractéristiques choisies de la longueur de ses séries (LS), comme la longueur moyenne d'une série (LMS) ou un quantile de la longueur d'une série. Dans la majorité des cas, le plan doit être conçu afin d'assurer à la fois une bonne sensibilité (c.-à-d. une longueur de série (LS) courte) par rapport aux tendances indésirables des observations qui arrivent ainsi qu'un niveau raisonnable de protection contre les fausses alarmes.

Toute technique qui utilise des cartes de contrôle doit posséder des fonctions évoluées de prise de décision pour qu'on la considère appropriée. Cependant, il arrive souvent que ces fonctions, seules, ne soient pas suffisantes. Les techniques qui offrent aussi une représentation graphique significative et commode, qui permet non seulement de détecter la présence d'un changement dans le comportement des données surveillées, mais aussi d'établir le point d'origine de ce changement et d'évaluer l'état courant du processus contrôlé, ont une valeur particulière.

Les cartes utilisées dans les applications industrielles comprennent la carte de Shewhart classique, la carte à moyenne mobile (simple ou pondérée, voir Lai (1974), Nelson (1983)), la carte à moyenne mobile à pondération

[1] E. Yashchin, IBM Corporation, Thomas J. Watson Research Center, Department of Mathematical Sciences, C. P. 218, Yorktown Heights, NY 10598.

exponentielle (MMPE, voir Roberts (1959, 1966), Bather (1963), Hunter (1987), Lucas et Saccucci (1990)), la carte à somme cumulée et la carte à somme cumulée - Shewhart (Page (1954), Barnard (1959), Lucas (1982), Yashchin (1985)), la carte de Girshik - Rubin (Girshik et Rubin (1952), Roberts (1966)), les Quangles (North (1980)), les Polyplots (Blazek, Novic et Scott (1987)) et quelques autres. On peut trouver un résumé de méthodes qui permettent habituellement d'obtenir des plans de contrôle avec une bonne résolution dans Yashchin (1987). Une des méthodes les plus populaires est celle de Page qui est obtenue en supposant que la densité visée des observations est $f_0(x)$ et qu'à un moment inconnu elle pourra passer à une "mauvaise" densité $f_1(x)$. Cette méthode fait appel au rapport de vraisemblance et exige le déclenchement d'un signal qui indique que la situation est hors contrôle au moment T si pour un $l \geq 1$, les l dernières observations $(X_{T-l+1}, X_{T-l+2}, \ldots, X_T)$ sont "significatives" au sens du test du rapport des probabilités séquentielles (TRPS) c.-à-d. que si pour un niveau de signal donné h

$$\sum_{i=T-l+1}^{T} \phi(X_i) > h, \tag{1.1}$$

où $\phi(X_i) = \log [f_1(X_i) / f_0(X_i)]$ représente une *valeur* attribuable à la i^e observation. Dans de nombreuses situations, plus particulièrement celles qui utilisent des familles de distributions exponentielles, la fonction de caractérisation est linéaire en X_i. Par exemple, dans le cas où $f_0(x)$ et $f_1(x)$ sont toutes deux normales avec un écart-type commun σ et moyennes μ_0 et $\mu_1 > \mu_0$, la valeur est

$$\log \frac{f_1(X_i)}{f_0(X_i)} = \frac{1}{\Delta}(X_i - k); \quad k = (\mu_0 + \mu_1)/2, \quad \Delta = \frac{\sigma^2}{\mu_1 - \mu_0}. \tag{1.2}$$

Il est évident, que dans des situations non-linéaires, on peut simplement utiliser la transformation $Z = \phi(X)$. Nous pouvons donc travailler dans le cadre linéaire sans perte de généralité. La méthode (1.1) est associée à une représentation graphique à somme cumulée qui nous permet d'effectuer une "autopsie" qui fournira vraisemblablement une idée additionnelle de la nature et de l'importance du changement ainsi que de son origine. Il a été démontré que la procédure de Page possède certaines propriétés d'optimalité (voir, par ex., Lorden (1971), Moustakides (1986), Banzal et Papantoni-Kazakos (1986), Ritov (1990)) et il existe de nombreuses raisons de croire que cette méthode jouera vraisemblablement un rôle dominant dans le domaine du contrôle des processus.

Dans le présent document nous examinons une carte qui suggère d'associer des poids à notre suite d'observations et de généraliser ainsi la technique des sommes cumulatives de la même façon que la technique des moyennes mobiles pondérées généralise la technique des moyennes mobiles simples. Dans les plans pondérés de type 1 (voir, p. ex., Davies et Goldsmith (1972), Bissell (1973), Yashchin (1989)) nous supposons soit que les poids sont donnés à l'avance, soit qu'ils sont disponibles avec les observations. Dans les plans pondérés de type 2, les poids sont essentiellement les paramètres d'un plan de contrôle qui accroissent les possibilités qu'a ce dernier de détecter les variations de la valeur des paramètres contrôlés (voir Yashchin (1989)). Nous discuterons aussi de plans qui comportent l'utilisation simultanée des deux types de pondération.

Dans la section 2, nous présentons les formats graphiques associés à la technique des sommes cumulatives pondérées. Dans la section 3, nous considérons un exemple relié à la surveillance de l'intégrité des données. Dans les sections 4 et 5, nous discutons de la technique pondérée de type 2 et de ses rapports avec le plan à moyenne mobile à pondération exponentielle (MMPE) généralisé. Finalement, dans la section 6, nous considérons un exemple comportant le contrôle simultané d'un grand nombre de paramètres à l'aide de la carte de Tunnel qui fait usage des deux types de poids.

2. FORMATS DES DONNÉES, DE PAGE ET DES SOMMES CUMULATIVES (CUSUM) UTILISÉS POUR REPRÉSENTER DES DONNÉES PONDÉRÉES EN SÉRIE

Supposons que le paramètre du processus contrôlé est la moyenne des observations X_1, X_2, ... , et que la suite correspondante de poids non négatifs est v_1, v_2, ... Un exemple typique d'une telle situation correspondrait au cas où X_i représente la proportion d'articles défectueux trouvés dans le i^e échantillon et où v_i est la taille de l'échantillon correspondant.

Supposons que nous sommes principalement intéressés par la possibilité que le processus pourrait se déplacer vers le haut pour atteindre un niveau inacceptable. Pour détecter la présence de telles situations, on pourrait utiliser un plan supérieur de Page pondéré défini en fonction de trois paramètres: $h^+ \geq 0$ (niveau du signal), k^+ (valeur de référence) et $0 \leq s_0^+ \leq h^+$ (point de départ). Pour appliquer ce plan, nous prenons comme point de départ s_0^+ et calculons la suite de sommes cumulatives

$$s_i^+ = \max \{s_{i-1}^+ + v_i (X_i - k^+), 0\}, \quad i = 1, 2, \ldots . \tag{2.1}$$

Si T est le premier indice i pour lequel $s_i^+ > h^+$, un signal de situation hors contrôle est déclenché au temps T.

Si un critère de signal additionnel est introduit, pour lequel un signal doit être déclenché au temps i si une seule observation x_i dépasse $c^+(v_i)$, la procédure sera appelée un plan supérieur de Page complété par une règle de Shewhart. Dans la majorité des applications, la fonction $c^+(v_i)$ est décroissante en v_i; dans le présent document, nous ne nous attaquerons pas, en détail, aux questions portant sur le choix de cette fonction.

Il n'est pas difficile de constater que, lorsque tous les poids sont identiques, le plan (2.1) devient un plan Cusum - Shewhart habituel, tel que décrit, par exemple, dans Yashchin (1985). De plus, compte tenu de l'expression (1.2), le plan (2.1) découle directement de l'expression (1.1) dans le cas normal avec des tailles d'échantillon variables. Comme d'habitude, la valeur de référence k^+ est choisie de façon à être proche du point milieu entre les niveaux acceptable et inacceptable du processus, alors que le niveau du signal h^+ caractérise le niveau d'accumulation de données permis dans le plan de contrôle, c.-à-d. qu'il représente l'instrument principal pour atteindre l'équilibre désiré entre le niveau de protection contre les fausses alarmes et les exigences en matière de sensibilité.

De la même façon, on peut définir un plan *inférieur* de Page pondéré $\{s_i^-\}$ afin de détecter des changements *à la baisse* dans le niveau du processus. Comme dans le cas des plans de Page habituels, on peut atteindre ce résultat en appliquant un plan *supérieur* de Page pondéré avec paramètres $h^- \geq 0$, k^-, $0 \leq s_0^- \leq h^-$ (et, peut-être, $c^-(v_i)$) à la suite d'observations "réfléchies", $-X_1$, $-X_2$, ... , c.-à-d. que

$$s_i^- = \max \{s_{i-1}^- + v_i (-X_i - k^-), 0\}, \quad i = 1, 2, \ldots , \tag{2.2}$$

avec déclenchement d'un signal si $s_i^- > h^-$. Dans ce cas, la valeur de référence réfléchie, $(-k^-)$ est généralement choisie près du point milieu entre le niveau acceptable et inacceptable (inférieur) du processus. Finalement, on peut effectuer un contrôle bilatéral si l'on combine un plan supérieur et un plan inférieur afin d'obtenir un plan bilatéral de Page pondéré.

Comme dans le cas des plans Cusum - Shewhart habituels, les plans pondérés peuvent être appliqués dans un de deux formats graphiques. Pour représenter ces données dans un format de masque en V (Cusum), on peut choisir une constante de centrage commode t_0 (qui peut correspondre, par exemple, au niveau cible du processus ou à la moyenne de toutes les observations) puis tracer les points (0, 0) et

$$\{\sum_{j=1}^{i} v_j , \sum_{j=1}^{i} v_j (X_j - t_0)\}, \quad i = 1, 2, \ldots$$

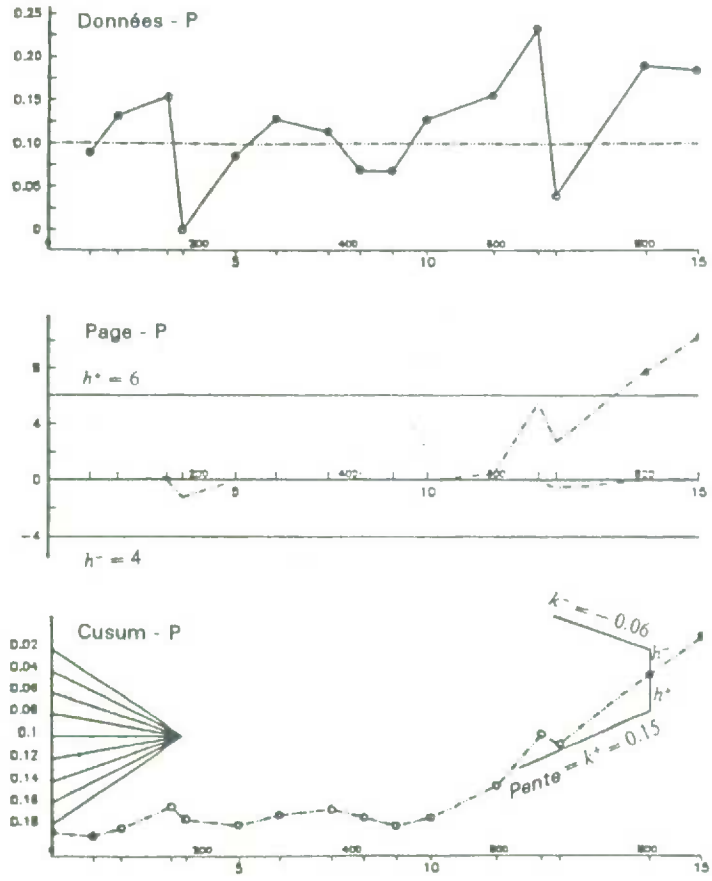
comme on le voit dans la partie inférieure de la fig. 2.1. Puisque, dans la majorité des applications, l'échelle utilisée pour l'axe des y n'a pas d'importance, nous la remplaçons par l'échelle des *pentes* correspondant au rapporteur et calculées selon la convention que la pente correspondant à la ligne horizontale du graphique est égale à t_0; ainsi, s'il arrive que t_0 est zéro, notre pente "généralisée" devient simplement la pente au sens habituel du terme. Dans notre exemple, la constante de centrage choisie est $t_0 = 0.1$, par conséquent, le chemin "plat" d'une trajectoire Cusum correspond au niveau des observations autour de 0.1, comme cela est indiqué sur l'échelle verticale. Si le niveau d'une suite $\{X_i\}$ est 0.02 au-dessus du niveau cible, la trajectoire Cusum pondérée montrera une pente correspondant au rayon du rapporteur qui porte l'inscription 0.12 et ainsi de suite. Il est évident qu'on peut facilement reconstruire l'échelle "implicite" réelle de l'axe vertical à l'aide de la longueur connue du rayon horizontal du rapporteur ainsi que de ses pentes. De plus, si l'on relie deux points de notre trajectoire et que l'on établit un rapport entre la pente (généralisée) du segment résultant et le rapporteur, on peut obtenir la valeur d'une moyenne *pondérée* des observations correspondant à ce segment, de façon analogue à la méthode utilisée pour les chemins Cusum classiques. Dans l'exemple où les observations correspondent à des proportions échantillonnées successives d'articles défectueux et les poids aux tailles d'échantillon correspondantes, cette pente est égale à la proportion d'articles défectueux dans le segment.

Comme dans le cas d'un plan Cusum classique, on peut mettre le plan de contrôle Cusum pondéré en application au moyen d'un masque en V, défini en fonction du niveau du signal et de la valeur de référence des plans supérieurs et (ou) inférieurs, comme on le voit sur la fig. 2.1. Il faut remarquer que les valeurs de référence correspondent aux pentes (généralisées) des bras du masque, exprimées en fonction du rapporteur, c.-à-d. que la valeur de la pente du rayon horizontal correspond à celle de la constante de centrage t_0. Par exemple, le masque illustré à la fig. 2.1 correspond au plan de contrôle bilatéral ($h^+ = 6$, $k^+ = 0.15$, $h^- = 4$, $k^- = -0.06$) et un signal de situation hors contrôle n'est pas déclenché tant que la trajectoire Cusum reste entre les bras du masque. Comme d'habitude, on obtient un contrôle unilatéral en n'appliquant que le demi-masque approprié. Finalement, on peut appliquer une règle de Shewhart complémentaire en faisant la "parabolisation" locale du masque au moyen des rayons des pentes généralisées $c^+(v_i)$ et $-c^-(v_i)$ qui sortent de l'origine du masque.

Dans le format de Page (fig. 2.1, au milieu) la suite $\{s_i^+\}$ (et/ou $\{s_i^-\}$) est tracée par rapport à i et évaluée par rapport à la ligne horizontale correspondant au niveau du signal h^+ (h^-). Ce format est relativement compact pour afficher des données en série, puisque seules les données pertinentes au processus de contrôle sont conservées; par contre, il est étroitement lié à un plan de contrôle particulier ce qui limite sa valeur comme outil graphique. Dans la représentation de Page, il n'est pas nécessaire d'utiliser l'axe horizontal non homogène par rapport aux numéros séquentiels des observations. Cependant, il est encore avantageux de le faire, car cela facilite l'interprétation des pentes et l'estimation du niveau courant du processus et rend ainsi la carte plus significative. De même, nous établissons l'échelle de l'axe des x conformément à la somme cumulative des poids afin de produire le tracé des données.

Nous appelons les formats mentionnés ci-dessus les formats Cusum - P, Page - P et des données - P pour afficher des données en série. S'il examine les tracés de la fig. 2.1, le lecteur remarquera que, comme dans le cas des procédures Cusum classiques, l'utilisation du format Cusum pondéré permet d'évaluer le niveau du processus et de détecter les points de changement beaucoup plus facilement que si l'on utilisait le format Shewhart (des données). L'une des raisons qui expliquent cette situation est reliée au fait que les points basés sur une taille d'échantillon relativement petite (p. ex., les points n° 4 et n° 13 de la fig. 2.1) introduisent un niveau de confusion important dans le tracé des données et peuvent, dans certains cas, empêcher de remarquer un changement dans le niveau. En même temps, ces points ne jouent pas de rôle important dans les tracés Cusum - P où on leur attribue des poids relativement faibles.

Figure 2.1: Formats des données, de Page, et Cusum pour afficher des données en série pondérées. L'échelle horizontale supérieure représente la somme cumulative des poids. L'échelle inférieure correspond aux numéros des observations.



3. PLAN CUSUM PONDÉRÉ DE TYPE 1: UN EXEMPLE

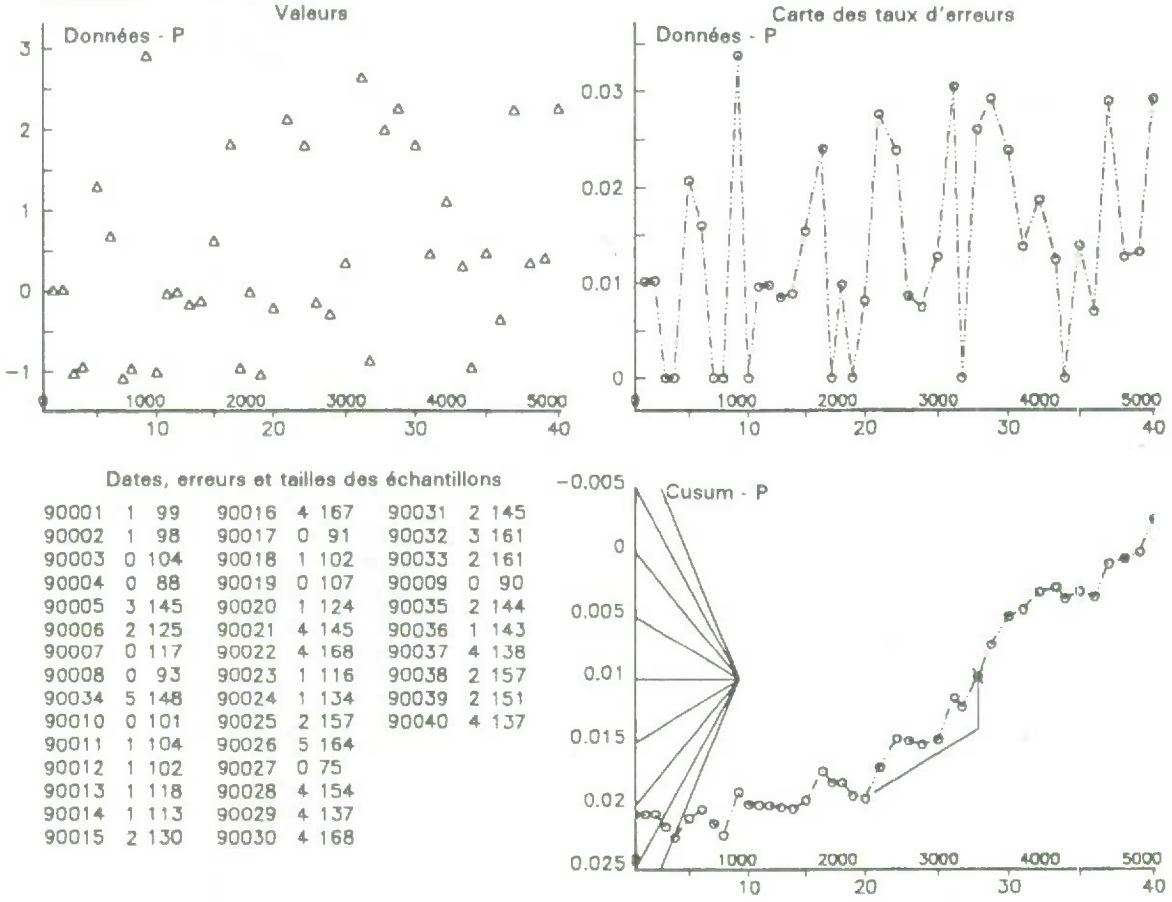
Dans la présente section nous montrons comment on peut appliquer la technique Cusum pondérée pour surveiller le niveau d'intégrité des données dans une base de données. Supposons que la suite de questionnaires imprimés se rapportant aux tendances d'achat de consommateurs "représentatifs" arrive au bureau central d'un organisme qui traite des données, où les données figurant sur ces questionnaires sont introduites dans une base de données informatique par un groupe de commis. Les gestionnaires de la base de données sont intéressés à surveiller la proportion de données erronées introduites par chaque commis. Pour ce faire, le service du contrôle de la qualité extrait au hasard, à la fin de chaque jour, 5% des enregistrements traités. Les enregistrements choisis sont comparés aux questionnaires correspondants et le nombre d'erreurs détectées est relevé. On obtient donc un ensemble de données (voir la fig. 3.1) qui renferme la date (1re colonne), le nombre correspondant d'erreurs détectées (2e colonne) et le nombre d'enregistrements inspectés (3e colonne). Un tel ensemble de données est tenu à jour pour chaque commis. Les tailles des échantillons dans chaque enregistrement sont différentes à cause de l'échantillonnage aléatoire et des différences de productivité parmi les commis.

Le système automatisé de contrôle du processus statistique employé par l'organisme est activé à la fin de chaque jour. Parmi d'autres rapports, le système produit les tracés du type montré à la fig. 3.1 pour chaque commis. La situation montrée dans la fig. 3.1 correspond à la fin du 40e jour de 1990. Elle reflète le rendement d'un commis donné que nous appellerons Alex. La proportion nominale (historique) de valeurs erronées est p_o = 0.01 et la taille typique d'un échantillon est d'environ n = 100. Implicitement, les taux de valeurs erronées sont tracés dans les formats des données - P et Cusum - P, comme on le voit dans la partie droite de la fig. 3.1. De plus, l'ensemble des valeurs z_i calculées au moyen de la formule standard,

$$z_i = \frac{\hat{p}_i - p_0}{\sqrt{p_0(1-p_0)/n_i}} = \frac{\hat{p}_i - 0.01}{0.99/\sqrt{n_i}} \qquad (3.1)$$

est tracé, dans le coin supérieur gauche, dans le format des données - P.

Figure 3.1: Le rapport informatisé utilisé pour surveiller les taux d'erreurs dans l'opération de saisie des données.

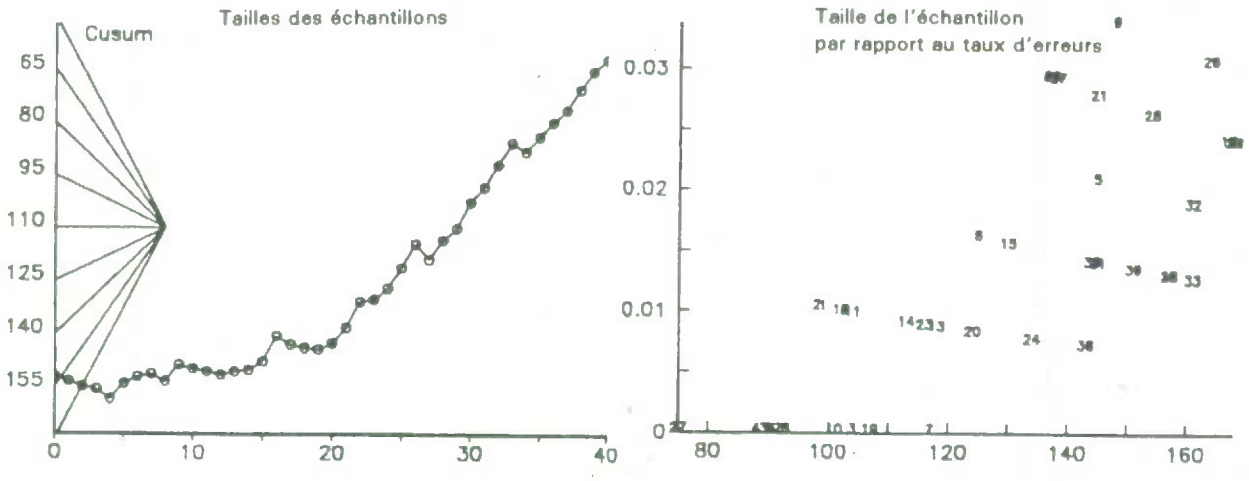


Les taux d'erreurs supérieurs à $p_i = 0.02$ sont considérés inacceptables et devraient donc être détectés aussitôt que possible. Pour détecter la présence de situations hors contrôle, l'organisme utilise un plan supérieur de contrôle pondéré avec comme niveau du signal $h = 4.2$ et comme valeur de référence $k = 0.015$. Pour la taille typique d'un échantillon, $n = 100$, le taux de fausses alarmes de ce plan est d'environ 1 par 190 jours. Par contre, la situation correspondant à $p_i = 0.02$ sera détectée, en moyenne, après 8.5 jours. Compte tenu du fait que le rapport de vraisemblance est à l'origine du plan Cusum pondéré de type 1, on peut s'attendre à ce que le rendement de ce plan par rapport aux tailles d'échantillon variables soit à peu près identique (voir, p. ex., Yashchin (1989, tableau 1)).

Peu après le début d'une alarme (le 28e jour de l'année) le gestionnaire de la base de données s'est entretenu avec Alex de l'importance de l'intégrité des données. À ce moment, il semblait que le taux d'erreurs d'Alex, qui correspondait à ce que l'on s'attendait la majorité du temps, s'était détérioré considérablement autour du 20e jour de l'année, sans raison apparente. Au point indiqué dans la fig. 3.1, la carte Cusum montre clairement que le taux d'erreurs d'Alex est encore autour de 0.02, bien que cette carte n'ait provoqué le déclenchement d'aucun nouveau signal de situation hors contrôle. Il faut remarquer que les deux cartes des données - P correspondant aux taux d'erreurs et aux valeurs ne montrent pas bien la présence d'un changement ainsi que le point où ce dernier commence.

En étudiant les tracés, le gestionnaire de la base de données a réalisé que les symboles sur l'échelle inférieure correspondant à l'axe des x du tracé ne sont pas très uniformes, ce qui laisse supposer que la charge de travail augmente dans la période qui suit le 20e jour de l'année. Effectivement, pendant cette période, il lui manquait toujours au moins deux employés, surtout parce que c'est la saison de l'année pendant laquelle les personnes s'absentent pour des raisons de santé. En l'absence de ces employés, leur part du travail a été répartie parmi ceux qui restaient, ce qui a requis du travail supplémentaire, dans certains cas. Le gestionnaire a donc soupçonné que le problème de qualité du travail d'Alex pourrait être relié au fait que ce dernier ne sait pas comment s'y prendre pour traiter des charges de travail très variables. Le gestionnaire a donc demandé au service du contrôle de la qualité de tracer les tailles d'échantillon pertinentes sur un graphique et de produire un diagramme de dispersion des tailles de l'échantillon par rapport aux taux d'erreurs. Les tracés résultants sont présentés dans la fig. 3.2. Ils montrent clairement l'augmentation de la charge de travail d'Alex qui a commencé il y a 20 jours. Le diagramme de dispersion laisse aussi supposer que le taux d'erreurs d'Alex tend à être beaucoup plus élevé les jours où sa charge de travail est élevée. Cette conclusion est appuyée par plusieurs tests statistiques standards qui ne seront pas mentionnés dans le présent article.

Figure 3.2: Le tracé Cusum des tailles des échantillons quotidiens pour un commis particulier (à gauche) qui montre l'accroissement de la charge de travail. Le fait de tracer ces tailles d'échantillon par rapport aux taux d'erreurs quotidiens (à droite) révèle le lien entre la charge de travail et le niveau d'erreurs dans la saisie des données. Les symboles de traçage représentent les numéros séquentiels des points.



4. PLANS PONDÉRÉS DE TYPE 2

Dans les plans pondérés de type 2, les poids ne sont pas associés aux données, ils représentent plutôt des paramètres du plan de contrôle. Dans ce cas, la pondération est utilisée pour accentuer l'importance relative des observations les plus récentes. Supposons qu'on nous donne un ensemble de poids positifs $1 = w_0 \geq w_1 \geq \cdots \geq 0$. Alors, un prolongement naturel du plan (1.1) demande le déclenchement d'un signal au moment T si, pour un $l \geq 1$, les l dernières observations $(X_{T-l+1}, X_{T-l+2}, \ldots, X_T)$ satisfont l'inégalité

$$\sum_{i=T-l+1}^{T} w_{T-i}\phi(X_i) > h \qquad (4.1)$$

Sans perte de généralité, nous supposerons que le paramètre contrôlé est la moyenne du processus et que la fonction de caractérisation est linéaire en X_i. Alors, le plan de contrôle exige le déclenchement d'un signal au moment T si, pour un $l \geq 1$

$$\sum_{i=T-l+1}^{T} w_{T-i} v_i (X_i - k) > h, \tag{4.2}$$

où, comme d'habitude, k se trouve à peu près à mi-chemin entre le "bon" et le "mauvais" niveau du processus. Les poids de type 1, $\{v_i\}$ apparaissent habituellement de façon naturelle comme partie de la fonction de caractérisation $\phi(X_i)$, par ex., ils représentent souvent les tailles d'échantillon associées à X_i. Comme on le démontre dans Yashchin (1989), ce type de plan nous permet d'améliorer considérablement la sensibilité par rapport aux variations de la moyenne du processus sans sacrifier beaucoup la sensibilité relative aux changements de niveaux. L'article qui vient d'être mentionné renferme une discussion portant sur des applications du plan Cusum pondéré de type 2, y compris le cas important du contrôle de processus à plusieurs variables.

La représentation graphique de l'expression (4.2) correspond exactement à la fig. 2.1, sauf que la longueur des intervalles sur l'axe des x tendra à augmenter. En d'autres mots, au moment T nous pouvons tracer seulement les points (0, 0) et

$$\{\sum_{j=1}^{i} w_{T-j} v_j \, , \sum_{j=1}^{i} w_{T-j} v_j \, (X_j - t_0)\}, \quad i = 1, 2, \ldots, T \tag{4.3}$$

et appliquer le masque V (h, k) au dernier point.

Dans le présent article, nous traiterons du problème lié à la représentation du plan (4.2) dans le format de Page. Le fait que le plan Cusum (1.1) puisse être représenté sous la forme (2.1) est extrêmement important dans certaines applications, surtout parce qu'il n'est pas nécessaire d'analyser, au point i, l'ensemble complet des données qui précèdent ce point pour déterminer si un signal de situation hors contrôle doit être déclenché. Les seules quantités nécessaires pour atteindre la conclusion sont la valeur précédente du plan; s_{i-1} , et le point le plus récent, X_i . Cette propriété permet d'utiliser efficacement la méthode Cusum pour surveiller, en temps réel, des volumes de données considérables et intenses. De plus, le caractère markovien du plan de Page permet de concevoir des cartes Cusum et d'en analyser les propriétés statistiques d'une façon relativement simple. Une question qui se pose naturellement est de savoir s'il existe une représentation semblable pour le plan Cusum pondéré de type 2.

Afin de répondre à cette question, définissons les constantes γ_1, γ_2, ... en fonction de notre suite de poids $\{w_i\}$ au moyen du processus suivant:

$$\gamma_0 = 1, \; \gamma_i = w_i - w_{i-1}\gamma_1 - w_{i-2}\gamma_2 - \cdots - w_1\gamma_{i-1} \, , \quad i = 1, 2, \ldots ; \tag{4.4}$$

les premiers membres de cette suite sont $\gamma_1 = w_1$, $\gamma_2 = w_2 - w_1^2$, $\gamma_3 = w_3 - 2w_1w_2 + w_1^3$, etc. De plus, définissons le plan supérieur de Page pondéré de type 2 de la façon suivante:

Plan de Page pondéré de type 2: Prenons au début $s_0 = 0$ et calculons la valeur de la suite

$$s_i = \gamma_1 s_{i-1} + \gamma_2 s_{i-2} + \cdots + \gamma_i s_0 + v_i (X_i - k), \quad i = 1, 2, \ldots ; \tag{4.5}$$

jusqu'à ce que ce processus soit dépasse h et déclenche un signal, soit devienne inférieur ou égal à 0. Dans ce dernier cas, redonnons à s_i la valeur 0 et relançons le processus (4.5) à partir de zéro, en ne tenant compte d'aucune des données obtenues précédemment.

Nous pouvons maintenant prouver le théorème suivant:

Théorème 4.1. Le plan supérieur de Page pondéré de type 2, tel que défini ci-dessus, et le plan (4.2) sont équivalents si et seulement si $\gamma_i \geq 0$ pour tout i.

Preuve: Nous prouverons le théorème pour le cas $v_i \equiv 1$, $i = 1, 2, \ldots$. La modification de la preuve pour l'appliquer au cas général est très simple. Supposons que le plan de Page a provoqué le déclenchement d'un signal au moment T et que la dernière fois où le plan a été relancé à partir de 0 était il y a r observations, c.-à-d. $s_{T-r} \equiv 0$. Alors il n'est pas difficile de constater que l'on peut obtenir les r dernières observations du plan de Page (4.5) en résolvant le système:

$$\begin{bmatrix} s_T \\ s_{T-1} \\ s_{T-2} \\ . \\ . \\ . \\ s_{T-r+1} \end{bmatrix} = \begin{bmatrix} 1 & w_1 & w_2 & w_3 & w_4 & \cdots & w_{r-1} \\ 0 & 1 & w_1 & w_2 & w_3 & \cdots & w_{r-2} \\ 0 & 0 & 1 & w_1 & w_2 & \cdots & w_{r-3} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 0 & . & . & . & 0 & 1 \end{bmatrix} \times \begin{bmatrix} X_T - k \\ X_{T-1} - k \\ X_{T-2} - k \\ . \\ . \\ . \\ X_{T-r+1} - k \end{bmatrix} \tag{4.6}$$

c.-à-d. qu'effectivement le plan (4.2) doit aussi provoquer le déclenchement d'un signal au moment T.

Inversement, supposons que le plan (4.2) a provoqué le déclenchement d'un signal au moment T et que les r dernières observations sont "significatives" et considérons les valeurs du plan de Page pondéré (4.5). Si la valeur de ce plan, il y a r observations, était 0, alors la relation (4.6) montre que le plan de Page devrait aussi provoquer le déclenchement d'un signal au moment T. Si tous les coefficients γ_i sont non négatifs, ce plan devrait aussi provoquer le déclenchement d'un signal au moment T si $s_{T-r} > 0$, puisque, dans ces conditions, s_T est une fonction croissante pour toutes les valeurs antérieures du plan. Si, toutefois, un de ces coefficients est négatif, on peut trouver une réalisation des observations pour laquelle le plan de Page pondéré ne provoquera pas le déclenchement d'un signal au moment T. En fait, supposons que le premier coefficient négatif de la suite $\{\gamma_i\}$ est γ_2 et considérons la réalisation ci-après des observations:

$$X_1 - k = h, \quad X_2 - k = \epsilon - \gamma_1 h, \quad X_3 - k = -\gamma_1 \epsilon - \gamma_2 h, \quad X_4 - k = h, \tag{4.7}$$

où ϵ est un nombre positif qui ne dépasse pas h. Pour cette réalisation, le plan de Page est relancé après la troisième observation et un signal de situation hors contrôle *n'*est *pas* déclenché. Par contre, puisque la somme pondérée des deux dernières valeurs est

$$h + \gamma_1(-\gamma_1 \epsilon - \gamma_2 h) > h, \tag{4.8}$$

pourvu que ϵ soit assez petit, la procédure (4.2) *peut* entraîner le déclenchement d'un signal lors de la dernière observation. Il est facile de construire une réalisation semblable à celle donnée en (4.7) et d'utiliser l'argument qui précède pour le cas où l'indice du premier coefficient négatif de la suite $\{\gamma_i\}$ est arbitraire; les détails seront omis.

Il n'est pas difficile de généraliser le théorème 4.1 pour les plans inférieurs ainsi que pour le cas où un plan supérieur pondéré est complété par une avance ou par une limite de Shewhart, comme cela est proposé dans Lucas (1982) et dans Lucas et Crosier (1982). Dans le premier cas, le plan (4.5) est lancé à partir d'un point $0 < s_0 \leq h$ plutôt que de 0. Pour compléter le plan avec une limite de Shewhart $c(v_i)$ (c.-à-d. qu'un signal de situation hors contrôle est déclenché si une seule observation dépasse $c(v_i)$) nous "parabolisons" le masque en V à l'aide d'un rayon de pente $c(v_i)$ qui part de l'origine du masque.

Le cas des poids qui décroissent géométriquement, $\{w_i = \gamma^i, i = 0, 1, 2, \ldots\}$ où γ est habituellement compris entre 0.7 et 1 présente un intérêt particulier. Dans ce cas, le plan (4.2) est appelé un *plan basé sur les sommes cumulatives géométriques (plan Cusum géométrique)* avec paramètre γ et représenté par $CG(\gamma)$. L'expression

(4.4), nous permet d'écrire $\gamma_1 = \gamma$, $\gamma_i = 0$, $i = 2, 3, \ldots$, et le théorème mentionné plus haut se ramène essentiellement à un résultat équivalent pour les sommes cumulatives géométriques dont la preuve est donnée dans Yashchin (1989). Le caractère markovien des sommes cumulatives géométriques assure une analyse et une mise en oeuvre relativement faciles de cette technique. Sous la forme générale, le plan supérieur $CG(\gamma)$ est représenté de la façon suivante:

Prenons comme point de départ $0 \le s_0 \le h$ et calculons la suite

$$s_i = \max(0, \gamma s_{i-1} + v_i(X_i - k)), \quad i = 1, 2, \ldots, \tag{4.9}$$

et déclenchons un signal au temps T si $s_T > h$.

Pour concevoir le plan $CG(\gamma)$ ci-dessus, on pourrait simplement supposer que tous les poids de type 1 sont égaux à la même valeur "typique", $v_i \equiv v$ et alors trouver la LMS du plan CG habituel

$$s_i = \max(0, \gamma s_{i-1} + X_i - k) \tag{4.10}$$

avec niveau du signal h/v et avance s_0/v, par ex., à l'aide des tableaux présentés dans Yashchin (1989).

5. PLAN CUSUM PONDÉRÉ DE TYPE 2 ET LE PLAN MMPE GÉNÉRALISÉ

Définissons le plan à moyenne mobile à pondération exponentielle généralisé (MMPE-G) de la façon suivante:

Calculons la suite $\{\bar{s}_i\}$

$$\bar{s}_i = \frac{v_i X_i + \gamma v_{i-1} X_{i-1} + \cdots + \gamma^{i-1} v_1 X_1}{v_i + \gamma v_{i-1} + \cdots + \gamma^{i-1} v_1} \quad i = 1, 2, \ldots, \tag{5.1}$$

déclenchons un signal au temps T si $|\bar{s}_T - Target| > L\sigma(\bar{s}_i)$, où T_{Cible} est le niveau désiré du processus et où $\sigma(\bar{s}_i)$ est l'écart type de $\bar{s}_i$. La valeur de L est choisie de façon à assurer un faible taux de fausses alarmes.

Le plan (5.1) ne peut être représenté comme une chaîne de Markov et, par conséquent, il est généralement impossible de calculer sa LMS exactement. Cependant, ses imperfections statistiques le rendent, de toute façon, inapproprié comme outil principal de contrôle, surtout à cause de son inertie élevée, comme cela est mentionné dans Yashchin (1987). Nous croyons que l'outil principal devrait être le plan $CG(\gamma)$ bilatéral. Toutefois, le fait de représenter le plan MMPE-G sur la carte, avec ses limites de contrôle, nous permet d'évaluer graphiquement l'état du processus et, par conséquent, le plan MMPE-G pourrait bien être utilisé comme un plan de contrôle secondaire, pourvu que sa LMS cible soit beaucoup plus grande que celle du plan principal. Les valeurs recommandées de L sont données dans le tableau 5.1. Nous avons obtenu ces valeurs en supposant, dans l'équation (5.1), (*a*)$v_i \equiv v$, (*b*) X_i sont des variables aléatoires normales, indépendantes et qui suivent la même distribution et (c) que le nombre de variables sur lesquelles la moyenne calculée à l'aide de la formule (5.1) est basée est infini. Dans ces hypothèses, le plan (5.1) est équivalent au plan MMPE habituel (voir, p. ex., Roberts (1959, 1966)), qui part de $\tilde{s}_0 = Target$ et qui est exécuté dans l'intervalle $Target \pm L\sigma(\tilde{s})$ au moyen de la relation

$$\tilde{s}_i = \gamma \tilde{s}_{i-1} + (1 - \gamma) X_i, \quad i = 1, 2, \ldots \tag{5.2}$$

Il n'est pas difficile de constater que pour obtenir la distribution des LS du plan bilatéral précédent, il suffit d'obtenir une matrice de passage du plan $CG(\gamma)$ supérieur (4.10) par rapport au niveau du signal $h = 2L\sigma(\bar{s})/(1 - \gamma)$ et $k = Target - L\sigma(\bar{s})$, de transformer 0 en état absorbant et de calculer la distribution des LS associée à la matrice de passage résultante pour le point de départ $s_0 = h/2$.

Dans de nombreuses situations pratiques, la variance de X_i est inversement proportionnelle à v_i, c.-à-d.

$$\sigma^2(X_i) = \sigma^2/v_i, \quad i = 1, 2, \ldots \tag{5.3}$$

Dans ce cas $\sigma(\bar{s}_i)$ est donné par

$$\sigma(\bar{s}_i) = \sigma \times r_i, \quad \text{où } r_i = \frac{\sqrt{v_i + \gamma^2 v_{i-1} + \cdots + \gamma^{2(i-1)} v_1}}{v_i + \gamma v_{i-1} + \cdots + \gamma^{i-1} v_1} \tag{5.4}$$

La valeur de σ peut être estimée à partir d'un ensemble stable de n données simples $\{\tilde{X}_i\}$ basé sur les poids $\{\tilde{v}_i\}$, p. ex., à l'aide de la formule

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\tilde{v}_i \tilde{v}_{i+1}}{\tilde{v}_i + \tilde{v}_{i+1}} (\tilde{X}_{i+1} - \tilde{X}_i)^2 \tag{5.5}$$

La formule (5.5) présente l'avantage d'être robuste par rapport aux changements possibles dans le niveau du processus qui sont présents dans les données. Cependant, elle produit des estimations biaisées en présence d'une corrélation sériale. Dans les cas où il y a corrélation sériale, on pourrait utiliser la formule

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(\tilde{X}_i - \bar{X})^2}{(n/\tilde{v}_i) - (1/\bar{v})} \quad \text{où} \quad \bar{v} = \frac{1}{n} \sum_{i=1}^{n} \tilde{v}_i, \quad \bar{X} = \frac{1}{n\bar{v}} \sum_{i=1}^{n} \tilde{X}_i \tilde{v}_i \tag{5.6}$$

Dans le cas où les poids sont égaux, les deux formules (5.5) et (5.6) se simplifient pour donner des formules bien connues. De plus, il n'est pas difficile de constater que les valeurs de $\sigma(\bar{s}_i)$ sont invariantes par rapport à l'échelle de v_i. Il vaut aussi la peine de mentionner que l'argument présenté ci-dessus peut être généralisé facilement afin d'inclure des poids de type 2 plus généraux, c.-à-d. non géométriques.

Tableau 5.1: **Les valeurs de L pour lesquelles le plan MMPE correspondant (5.2) a une LMS précisée. Ces valeurs s'approchent beaucoup de L pour le plan MMPE-G (5.1). On suppose que les $\{X_i\}$ sont normales avec $\mu = T_{CSM}$ et écart-type σ. Les limites de contrôle de (5.2) sont** *Target* $\pm L\sigma(\bar{s})$ **où** $\sigma(\bar{s}) = \sigma\sqrt{(1-\gamma)/(1+\gamma)}$.

γ	LMS = 50	100	200	500	1000	2000	5000	10000	20000	50000	100000
0.7	2.17	2.45	2.71	3.02	3.24	3.44	3.69	3.87	4.04	4.25	4.41
0.75	2.12	2.41	2.68	3.00	3.22	3.42	3.67	3.85	4.03	4.24	4.40
0.8	2.05	2.36	2.64	2.96	3.19	3.40	3.65	3.84	4.01	4.23	4.39
0.85	1.96	2.28	2.57	2.91	3.14	3.36	3.62	3.81	3.98	4.21	4.37
0.9	1.81	2.15	2.45	2.81	3.06	3.29	3.56	3.75	3.93	4.16	4.33

6. PLAN CUSUM PONDÉRÉ DE TYPE 2: UN EXEMPLE

Considérons encore une fois la situation décrite dans la section 3 et supposons que le gestionnaire de la base de données est intéressé à obtenir un rapport hebdomadaire sur l'état courant des caractéristiques de qualité de base correspondant aux quatre régions-échantillons, les régions du Centre, du Sud, de l'Est et de l'Ouest. La base de données sommaire utilisée pour surveiller la qualité relative à la région de l'Ouest figure au tableau 6.1. Nous fournirons maintenant l'interprétation de la dernière ligne de données dans ce tableau, correspondant au

18 juin 1990. Les niveaux "cibles" ou "historiques" des paramètres correspondants sont donnés dans la dernière ligne du tableau.

Le 18 juin, 16,000 questionnaires correspondant à la région de l'Ouest ont été reçus et traités par les commis préposés à la saisie des données (le niveau "cible" pour l'entrée des données est de 20,000 questionnaires par semaine). De tous ces questionnaires, 500, c.-à-d. 0.5 x 1000, ont été choisis au hasard par le service du contrôle de la qualité pour évaluer le niveau des erreurs introduites au cours du processus de saisie des données. Le niveau "cible" d'inspection est 5%, c.-à-d. que normalement le service aurait dû inspecter environ 800 formules. La proportion d'erreurs trouvées dans ces formules est appelée "notre taux d'erreurs" et elle est représentée dans le tableau par la colonne "Erreur". Ainsi, 1.1% des questionnaires inspectés contenaient des erreurs introduites par les préposés à la saisie des données. De plus, 10.4% des 16,000 questionnaires renfermaient des valeurs manquantes, 4.9% contenaient des données erronées et 3.5% ont été rejetés aux essais de cohérence des données, qui ont révélé que les réponses, bien que techniquement valides, renfermaient des contradictions logiques. Le taux de réponse était de 29%. Finalement, tous les 16,000 questionnaires sont sujets à une sélection pour empêcher le biais d'échantillonnage. Les gestionnaires sont particulièrement intéressés à maintenir la proportion de répondants de la région de l'Ouest qui font plus de $50,000 par an à environ 20%, la proportion des personnes qui ont 30 ans ou plus à environ 40% et la proportion des hommes à 50%.

Les données en séries correspondant aux quatre régions qui nous intéressent sont représentées, en format CG, avec $\gamma = 0.8$, dans la fig. 6.1. Puisque le plan CG occupe naturellement une région sectorielle, nous produisons un ensemble de cartes ressemblant à un graphique à secteurs avec une origine commune. Cette forme de représentation est appelée une carte de Tunnel (voir Yashchin (1989)). Considérons, par exemple, le tracé utilisé pour surveiller les valeurs manquantes dans la région de l'Ouest. Pour obtenir ce tracé, nous commençons à l'origine de la carte et nous traçons les points (4.3) avec $w_i = \gamma^i$, $i = 0, 1, 2, \ldots$ et v_i proportionnel aux inscriptions dans la colonne "Entrée" des données. Il est évident que le coefficient de proportionnalité n'a aucun effet sur l'apparence du tracé. Supposons donc, pour simplifier, que v_i est choisi en divisant les valeurs dans la colonne "Entrée" du tableau 6.1 par leurs valeurs "cibles", 20. Ce choix laisse supposer que dans de "bonnes" conditions, la valeur typique de v_i est 1.

Estimons maintenant l'écart-type des valeurs dans la colonne correspondant aux valeurs manquantes. Supposons que la variance de ces valeurs est σ^2/v_i où σ^2 est le coefficient de proportionnalité. On peut donc estimer, par exemple, à l'aide des $n = 20$ premiers points des données dans (5.5), que la valeur de σ est $\hat{\sigma} = 0.128$. Toutefois, cette valeur est légèrement inférieure au niveau "historique" de σ, dont on sait qu'il est environ 0.15. Dans l'hypothèse que $\sigma = 0.15$ et que la moyenne du processus est égale à son niveau cible, 10, la LMS du plan CG (0.8) bilatéral avec niveau du signal $h = 0.5$ et $k^+ = 10.1$, $k^- = -9.9$ correspond approximativement à 1,000, nous utiliserons donc ce plan pour détecter les variations de niveau du processus dont la grandeur est égale ou supérieure à 0.2. Le masque en V correspondant est présenté dans la fig. 6.1. De plus, nous présentons les limites de contrôle du plan MMPE (5.1). La formule (5.4) indique que la constante r_{25} à utiliser au point courant est 0.341. Ainsi, dans l'hypothèse que $\sigma = \hat{\sigma} = 0.128$, nous obtenons que l'écart-type estimé du plan MMPE est 0.128 x 0.341 = 0.044. Comme le tableau 5.1 nous propose la valeur $L = 3.84$, les limites de contrôle 10 ± 0.168 du plan MMPE nous assurent une LMS cible d'environ 10,000.

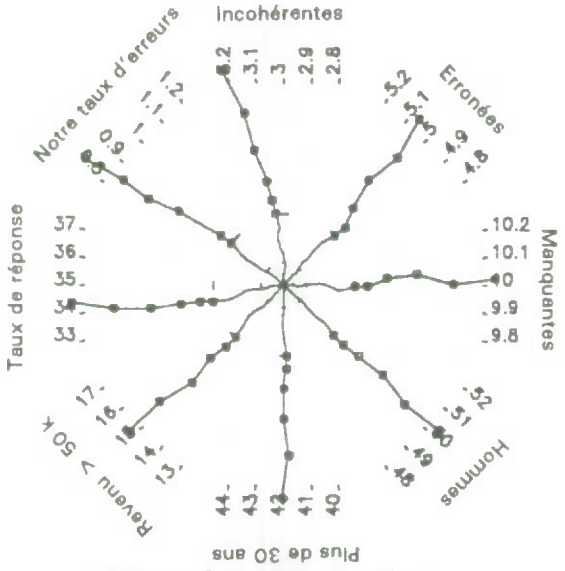
Tableau 6.1: Intégrité des données pour la région de l'Ouest. La dernière ligne du tableau correspond aux valeurs cibles.

Région de l'Ouest										
Date	Entrée	Inspec	Manq	Erron	Incohé	Erreur	Répon	>50k	>30 ans	Hommes
05/25/90	25	1.2	9.9	5.1	2.6	1	30	22	39	50
05/26/90	26	1.2	9.6	5.1	2.7	0.8	31	19	40	51
05/27/90	21	1.0	9.7	5.1	3	1.1	30	19	41	49
05/28/90	22	1.1	9.8	5.3	3.5	0.9	31	19	39	50
05/29/90	22	1.1	9.9	5.1	2.8	0.8	30	20	40	50
05/30/90	16	0.8	10.1	5.2	3	1	29	21	40	51
05/31/90	19	1.0	10.4	4.8	3	1	30	20	40	50
06/01/90	20	1.0	10.2	4.6	3	1.1	30	19	40	50
06/02/90	19	0.9	10.4	5	3.2	0.9	30	21	40	50
06/03/90	15	0.8	10.2	4.9	3.4	0.9	30	20	39	51
06/04/90	15	0.8	10	5.1	2.7	1.1	30	20	40	51
06/05/90	24	1.2	9.7	4.8	2.9	1	30	18	40	50
06/06/90	28	1.3	9.7	4.9	2.8	1.2	31	22	40	49
06/07/90	22	1.0	9.8	4.8	2.7	0.9	30	20	40	50
06/08/90	28	1.3	9.7	4.9	2.8	1.2	31	21	39	51
06/09/90	18	0.9	9.9	4.7	2.8	0.9	30	20	41	49
06/10/90	20	1.0	9.7	4.7	2.8	0.8	29	20	40	49
06/11/90	21	1.0	9.8	5.2	2.9	0.9	29	21	41	49
06/12/90	22	1.1	9.9	5	2.9	1.1	30	21	40	51
06/13/90	15	0.7	9.8	4.9	3	1.2	30	21	38	50
06/14/90	11	0.5	9.6	4.9	3.1	0.9	31	23	39	51
06/15/90	25	0.9	10.3	5.1	3.2	0.9	30	20	40	50
06/16/90	19	0.7	10.2	5.2	3.3	0.9	29	20	40	50
06/17/90	17	0.5	10.3	4.7	3.4	0.9	30	19	42	49
06/18/90	16	0.5	10.4	4.9	3.5	1.1	29	20	40	50
"Cible"	20	5%	10.0	5.0	3.0	1.0	30	20	40	50

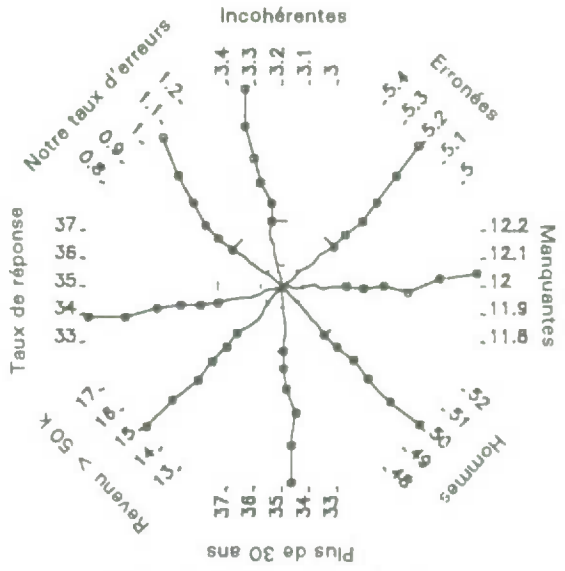
Le tracé correspondant montre clairement qu'il s'est produit, il y a quatre jours, un changement dans le niveau des données manquantes. Le niveau du processus avant le point de changement était d'environ 9.8, valeur qu'on peut déterminer en reliant simplement ce point à l'origine puis en estimant la pente résultante à l'aide du rapporteur. De même, le niveau du processus correspondant aux quatre derniers points était d'environ 10.3. Il importe de se rendre compte que la carte de Tunnel nous permet aussi de "voir" chaque observation en évaluant les pentes de chacun des segments qui relient les points adjacents correspondants.

Nous pouvons surveiller les autres paramètres de la même façon. Dans la figure 6.1, nous ne montrons pas les outils de contrôle associés à ces paramètres. Nous ferons simplement remarquer que les valeurs centrales des rapporteurs correspondent aux valeurs cibles du tableau 6.1 et que les échelles sont choisies de façon que, dans des conditions nominales, aucune trajectoire ne puisse "sortir" du secteur désigné. En général, les valeurs cibles diffèrent d'une région à l'autre. Parmi les autres tendances "intéressantes" qu'on peut tirer de la fig. 6.1, on peut remarquer la variation du niveau des données incohérentes (régions du Centre et de l'Ouest) qui mène à des trajectoires "courbes" pour le plan Cusum - P et la variation à la hausse du taux d'erreurs dans la saisie des données pour la région du Sud. En même temps, le taux pour les erreurs semblables dans la région du Centre s'améliore beaucoup et, actuellement, il est d'environ 0.8% des questionnaires traités. Cela peut laisser supposer, par exemple, l'existence de problèmes relatifs à la répartition de la charge de travail, compte tenu des conclusions auxquelles on en était arrivé pour l'exemple de la section 3. Une autre observation intéressante se rapporte au fait que les niveaux de données manquantes dans les régions de l'Est et de l'Ouest montrent des tendances opposées: par exemple, au point où le niveau de données manquantes dans la région de l'Ouest s'est déplacé vers le haut, le paramètre correspondant dans la région de l'Est s'est déplacé vers le bas. Finalement, la proportion actuelle de répondants qui ont 30 ans ou plus, dans la région de l'Est, est d'environ 41.9, ce qui est significativement plus que le niveau cible pour cette région.

Figure 6.1: **Les cartes de Tunnel utilisées pour surveiller les paramètres portant sur l'intégrité des données. Tous les tracés sont basés sur le plan Cusum géométrique avec γ = 0.8 utilisé avec les poids de type 1. Le masque en V symétrique h = 0.5, k^+ = 10.1, k^- = -9.9 est illustré pour la carte des données manquantes de la région de l'Ouest seulement. La ligne pointillée verticale sur cette carte précise la zone pour le plan MMPE qui, dans les conditions normales, est violée environ une fois par 10,000 jours. Les poids de type 1 pour le taux d'erreurs correspondent aux valeurs dans la colonne "Inspec" du tableau 6.1. Pour les autres paramètres, ces poids correspondent aux valeurs dans la colonne "Entrée".**

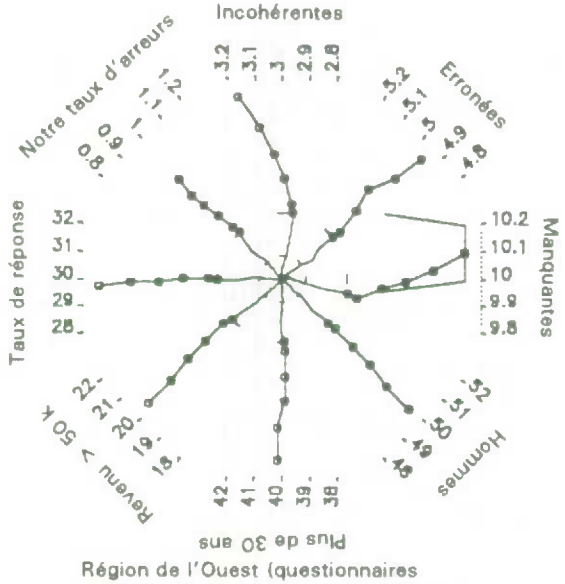


Région du Centre (questionnaires saisis pendant 4 semaines = 65/60K)

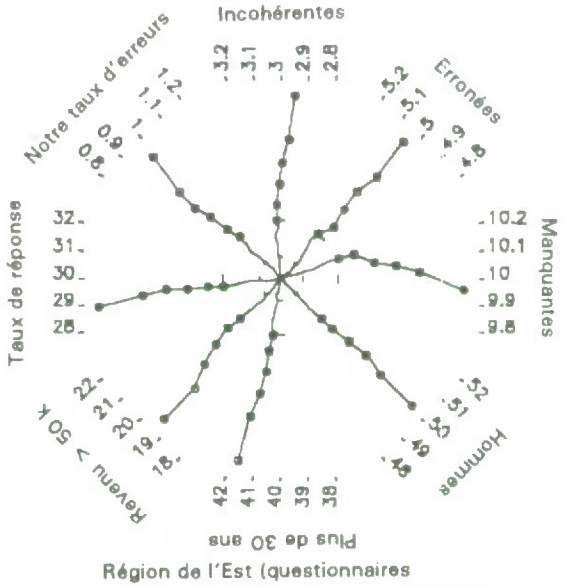


Région du Sud (questionnaires saisis pendant 4 semaines = 58/50K)

18 juin 1990



Région de l'Ouest (questionnaires saisis pendant 4 semaines = 77/80K)



Région de l'Est (questionnaires saisis pendant 4 semaines = 74/80K)

Dans les cartes de Tunnel, l'axe horizontal possède la propriété que, si tous les poids de type 1 sont identiques, $v_i \equiv v$, la longueur totale de l'axe est $v(1 + \gamma + \gamma^2 + \ldots) = v/(1 - \gamma)$. Par conséquent, quand la longueur de la trajectoire est trop courte pour atteindre l'échelle du rapporteur cela montre un sous-échantillonnage possible. Par exemple, la taille des quatre derniers échantillons utilisés pour estimer le taux d'erreurs dans la région de l'Ouest est beaucoup plus petite que celle à laquelle on s'attendrait dans des conditions nominales; donc, la trajectoire résultante correspondant à ce taux d'erreurs est trop courte. En comparant la longueur des segments entre les points adjacents de cette trajectoire aux longueurs correspondantes pour d'autres trajectoires, on peut obtenir une idée approximative de l'importance du sous-échantillonnage. De même, un suréchantillonnage amènera des chemins un peu plus longs, comme c'est le cas pour la région du Centre. Il faut remarquer que, comme cela est indiqué dans l'en-tête de ce tracé, le niveau nominal de saisie pour quatre semaines est de 60 000 questionnaires, alors qu'en réalité on en a reçu 65 000.

L'auteur croit que dans de nombreuses situations pratiques, la technique pondérée de type 2 ainsi que la carte de Tunnel associée offrent la possibilité de montrer l'état d'un grand nombre de paramètres de façon compacte, ce qui permet une interprétation utile des tendances et des données simples particulières les plus récentes. Cette technique est particulièrement profitable conjointement avec la technologie informatique moderne. Par exemple, dans un cadre où l'on utilise un poste de travail, on peut non seulement afficher des cartes du type montré dans la fig. 6.1, mais aussi les "faire" avancer et reculer dans le temps à l'aide de l'animation par ordinateur; on peut aussi appeler et supprimer les masques en V et les limites de contrôle du plan MMPE, explorer des cartes particulières, des données simples, des cartes et des tableaux connexes; autrement dit, utiliser la carte de Tunnel comme point de départ pour du travail exploratoire et de diagnostic.

REMERCIEMENTS

Je désirerais remercier Dr Betty J. Flehinger (IBM Research) pour ses conseils et ses échanges de points de vue utiles sur ce sujet et sur des exemples connexes.

BIBLIOGRAPHIE

Banzal, R.K., et Papantoni-Kazakos (1986). An Algorithm for Detecting a Change in a Stochastic Process, IEEE *Tran. Information Theory*, IT-32, 2, 227-235.

Barnard, G. (1959). Control Charts and Stochastic Processes, *Journal of the Royal Statistical Society* (B), 21, 239-257.

Bather, J. (1963). Control Charts and Minimization of Costs, *Journal of the Royal Statistical Society* (B), 25, 49-80.

Bissell, A.F. (1973). Process Monitoring With Variable Element Sizes, *Applied Statistics*, 22, 226-238

Blazek, L.W., Novic, B. et Scott, D.M. (1987). Displaying Multivariate Data Using Polyplots, *Journal of Quality Technology*, 19, 2, 69-74.

Davies, O.L., et Goldsmith, P.L. (1972). *Statistical Methods in Research and Production*, Oliver & Boyd, Edinburgh.

Girshick, M., et Rubin, H. (1952). A Bayes Approach to a Quality Control Model, *Annuals of Mathematical Statistics*, 23, 114-125.

Hunter, J.S. (1986). The Exponentially Weighted Moving Average, *Journal of Quality Technology* 18, 4, 203-210.

Lai, T.L. (1974). Control Charts Based on Weighted Sums, *The Annuals of Statistics*, 2, 1, 134-147.

Lorden, G. (1971). Procedures for Reacting to a Change in Distribution, *Annuals of Mathematical Statistics*, 42, pp. 1897-1908.

Lucas, J.M. (1982). Combined Shewhart-Cusum Quality Control Schemes, *Journal of Quality Technology*, 14, 2, 51-59.

Lucas, J.M., et Crosier, R.B. (1982). Fast Initial Response for Cusum Quality Control Schemes: Give your Cusum a Head Start, *Technometrics*, 24, 199-205.

Lucas, J.M., et Saccucci, M.S. (1990). Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements, *Technometrics*, 32, 1-29 (with discussion).

Moustakides, G.V. (1986). Optimal Stopping Times for Detecting Changes in Distributions, *The Annuals of Statistics*, 14, 2, 1379-1387.

Nelson, L.S. (1983). The Deceptiveness of Moving Averages, *Journal of Quality Technology* ,15, 2, 99-100.

North, W.R.S. (1980). The Quangle - A Modification of the Cusum Chart, *Applied Statistics*, 31, 155-158.

Page, E. (1954). Continuous Inspection Schemes, *Biometrika*, 41, 100-115.

Ritov, Y. (1990). Decision Theoretic Optimality of the Cusum Procedure, *The Annuals of Statistics*, 18, 3, 1464-1469.

Roberts, S. (1959). Control Chart Tests Based on Geometric Moving Averages, *Technometrics*, 1, 239-250.

Roberts, S. (1966). A Comparison of Some Control Chart Procedures, *Technometrics*, 8, 411-430.

Woodward, R., et Goldsmith, P.L. (1964). *Cumulative Sum Techniques*, ICI Monograph No. 3, Oliver and Boyd, London.

Yashchin, E. (1985b). On a Unified Approach to the Analysis Of Two-sided Cumulative Sum Schemes With Headstarts, *Advances in Applied Probability*, 17, 562-593.

Yashchin, E. (1987). Some Aspects of the Theory of Statistical Control Schemes, *IBM Journ. Res. Devel.*, 31, 199-205.

Yashchin, E. (1989). Weighted Cusum Technique, *Technometrics*, 31, 321-338.

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

TECHNIQUES DE CONTRÔLE ET D'AMÉLIORATION DE LA QUALITÉ DES DONNÉES DES GRANDES BASES DE DONNÉES

R.W. Pautke et T.C. Redman[1]

RÉSUMÉ

À mesure que s'accroît l'importance de l'information et des services pour l'économie, la qualité des données revêt une importance de plus en plus grande. Dans cette communication, nous examinons d'abord certaines indications qui laissent supposer que les niveaux actuels de qualité sont plutôt faibles. Nous décrivons ensuite certaines techniques élaborées et mises en oeuvre aux laboratoires AT&T Bell et à la Division des services du réseau de AT&T afin d'assurer un contrôle statistique de la qualité et d'améliorer cette dernière de façon durable. En particulier, nous y décrivons la technique du *suivi des données*, qui a été élaborée pour permettre une évaluation quantitative des processus d'introduction des données dans une base de données. Nous démontrons comment on peut avoir recours au suivi des données pour contrôler et accroître l'exactitude et la cohérence des données, ainsi que pour accélérer les processus. Tout au long de la communication, nous illustrons nos propos au moyen de ce que nous avons convenu d'appeler le «processus d'accès».

MOTS CLÉS: Suivi des données; processus d'accès; exactitude; cohérence; actualité.

1. INTRODUCTION ET SOMMAIRE

Le présent article porte sur la qualité des données qu'on trouve dans les bases de données, surtout celles utilisées par les grandes sociétés et les organismes gouvernementaux. Au cours des dernières années, ces données sont devenues un élément essentiel pour le fonctionnement de nombre de ces entreprises. Svanks (1988), en particulier, souligne que les données constituent souvent la ressource la plus précieuse d'une société. De même, on semble s'entendre pour reconnaître qu'il peut être extrêmement coûteux pour une société d'utiliser des données de piètre qualité (Ballou, D.P. et coll., 1989) (Lipiens, G.E., 1989). Pourtant, malgré son importance, la qualité des données n'a fait l'objet que d'une attention relativement limitée dans les ouvrages sur la qualité et les services d'information de gestion, où le nombre de références au niveau de la qualité des données est particulièrement peu élevé. En même temps, ces ouvrages indiquent qu'on observe, au moins dans certaines bases de données, des taux d'erreurs assez élevés. Ainsi, Lipiens, Garfinkel et Kunnathur (1982) estiment que, selon les ouvrages antérieurs sur le sujet (Pritzker, L. et coll.) (Terry, M.E., 1963), les taux d'erreurs se situent généralement entre 1 et 10%. Par ailleurs, certains ouvrages plus récents laissent supposer qu'il existe certaines bases de données où ces taux sont beaucoup plus élevés. Par exemple, Johnson et coll. (1981), Laudon (1986), et Morey (1982) font état, pour diverses bases de données, de taux d'erreurs dans l'intervalle 10% à 50%[2]. Toutefois, aussi renversants que puissent être ces taux d'erreurs, ils ne reflètent pas toute l'étendue du problème posé par la qualité des données, puisqu'ils ne portent que sur l'exactitude des données. Ces chiffres ne reflètent ni les incohérences relatives à des éléments d'information supposément identiques entre des bases de données

---

[1] R.W. Pautke, T.C. Redman, AT&T Network Services Division, AT&T Bell Laboratories. Crawfords Corner Road, Holmdel, New Jersey, USA 07733.

[2] Les mesures de la qualité des données spécifiques à AT&T sont couvertes par des droits de propriété et, à ce titre, ne sont pas présentées dans cet article. Tous les exemples cités sont réels dans la mesure où ils illustrent adéquatement les principaux points abordés.

qui se recoupent, ni le caractère incomplet des bases de données (données omises pour des segments entiers de la population visée), ni le caractère désuet des données.

Au cours des dernières années, nous avons été en mesure d'améliorer de façon durable l'exactitude des données à l'aide d'une technique que nous appelons le *suivi des données* (Huh, Y.U. et coll., 1990) (Huh, Y.U. et coll.). Après avoir constaté que le repérage et la correction des erreurs à l'intérieur même des bases de données étaient extrêmement coûteux et fastidieux et ne permettaient souvent d'obtenir que des résultats insatisfaisants, nous avons décidé de porter notre attention sur les *processus* permanents selon lesquels de nouveaux enregistrements sont ajoutés aux bases de données de façon plus ou moins continue, les enregistrements existants sont mis à jour, et ainsi de suite. Il serait vain de s'attendre à améliorer de façon durable la qualité des données tant et aussi longtemps que ces processus n'auront pas été améliorés et ne feront pas l'objet d'un contrôle statistique. Essentiellement, le suivi des données a pour objet de recueillir des renseignements sur le rendement d'un processus afin de pouvoir y apporter les améliorations nécessaires et d'être en mesure de le contrôler.

Dans le présent article, nous considérons le suivi des données dans une perspective plus générale et cherchons de quelles façons nous pouvons étendre son utilisation à l'étude de la cohérence et de l'actualité des données. Pour les fins de cet article, nous définissons l'*exactitude*, la *cohérence* et l'*actualité* comme suit.

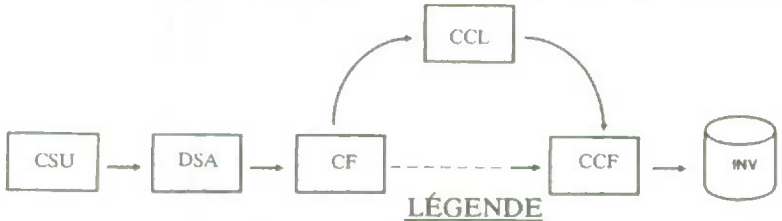
1. *Exactitude*. L'exactitude est une mesure de l'accord avec une source donnée[3].

2. *Cohérence*. Deux ensembles de données sont dits cohérents si la correction du premier implique celle de l'autre. (Nota: Il est possible de donner d'autres définitions de la cohérence des données. Pour plus de renseignements, voir Caby.)

3. *Actualité*. Les données sont dites actuelles si elles sont à jour.

Bien sûr, il existe d'autres aspects de la qualité des données et on peut définir la cohérence de façon beaucoup plus générale; le lecteur désireux de prendre connaissance d'un exposé plus général du sujet est invité à consulter Levitin et Redman.

Figure 1: Le processus d'accès: CSU-DB d'inventaire



LÉGENDE

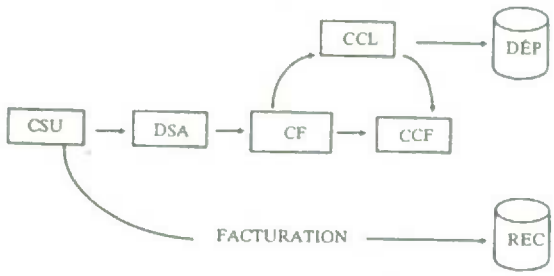
CSU	-	**Commande de service universel**
DSA	-	**Demande de service d'accès**
CF	-	**Commande ferme**
CCL	-	**Activités de la compagnie de téléphone**
CCF	-	**Confirmation de commande ferme**
INV	-	**Base de données d'inventaire**

Le processus d'accès est lancé lorsqu'un client passe une commande de service et qu'une «commande de service universel» est créée. Une «demande de service d'accès» est ensuite établie à partir de la CSU. Lorsque la DSA est établie, elle est traduite dans un format commercial normalisé et une «commande ferme» est transmise à la «compagnie exploitant le centre local». La CCL confirme alors qu'elle va assurer la prestation du service requis par l'intermédiaire d'une «confirmation de commande ferme». Une fois le service fourni et soumis à un essai, l'enregistrement complet est stocké dans la «base de données du système de gestion de la capacité d'accès».

---

[3] Il arrive parfois qu'on confonde les concepts d'exactitude et de *précision*. Selon Shewhart (1939), l'exactitude est une mesure du degré de correction, tandis que la précision «est une mesure du degré de reproductibilité».

Figure 2: Le processus d'accès sur les dépenses et les recettes

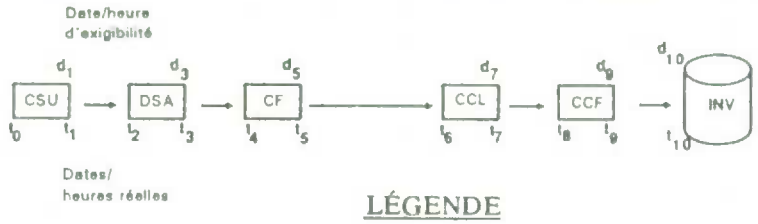


LÉGENDE

CSU	-	**Commande de service universel**
DSA	-	**Demande de service d'accès**
CF	-	**Commande ferme**
CCL	-	**Activités de la compagnie de téléphone**
CCF	-	**Confirmation de commande ferme**
DÉP	-	**Dépenses**
REC	-	**Recettes**

Comme on le voit ci-dessus, le processus d'accès donne également lieu à l'introduction de données dans les bases de données sur les «dépenses» et sur les «recettes».

Figure 3: Le processus d'accès: Considérations relatives à l'actualité



LÉGENDE

CSU	-	**Commande de service universel**
DSA	-	**Demande de service d'accès**
CF	-	**Commande ferme**
CCL	-	**Activités de la compagnie de téléphone**
CCF	-	**Confirmation de commande ferme**
INV	-	**Base de données d'inventaire**

La présente figure illustre les mêmes sous-processus que la figure 1. La longueur des cases et des flèches représente respectivement le temps consacré à l'exécution d'activités à valeur ajoutée et les délais d'attente. Les dates/heures t_0, t_1..., t_{10} indiquent les dates/heures réelles de lancement et d'achèvement des divers sous-processus. Les dates/heures d_1, d_3, d_5, d_7, d_9 et d_{10} sont les dates/heures nominales auxquelles les activités à valeur ajoutée doivent être terminées et d_{10} la date/heure nominale à laquelle l'enregistrement doit être introduit dans la base de données d'inventaire.

Après avoir passé en revue, à la section 2, les idées de base ayant présidé à l'élaboration du concept de suivi des données et avoir décrit une façon d'implanter cette technique, nous verrons, aux sections 3 et 4, comment on peut l'utiliser pour aborder chacun des problèmes de qualité susmentionnés (exactitude, cohérence et actualité). Enfin, nous présenterons les conclusions auxquelles nous en sommes arrivés à la section 5. Tout au long de cet article, nous illustrerons les principaux points étudiés à l'aide du processus d'accès utilisé par AT&T pour commander et recevoir un service d'accès des compagnies de téléphone. Les figures 1 à 3, qui présentent divers

aspects du processus d'accès, servent respectivement à illustrer les caractéristiques clés du processus sur le plan de l'exactitude, de la cohérence et de l'actualité des données.

2. SUIVI DES DONNÉES

Selon une des idées les plus fondamentales exposées dans les ouvrages sur la qualité, il vaut mieux *prévenir* les défectuosités et/ou les erreurs plutôt que de les détecter et les corriger (c.à-d. que la *prévention* est préférable à l'*inspection* et à la *réfection*). Dans cette perspective, on peut comparer une base de données à un lac: pour s'assurer que les eaux du lac sont saines, il convient d'abord de s'assurer que les eaux d'alimentation du lac sont elles-mêmes saines. À cette fin, il faut éliminer les sources de pollution situées en amont et/ou s'assurer que les sources d'alimentation en eau du lac sont propres. Il s'agit d'une approche fondamentalement différente de celle qui consiste à «dépolluer le lac». En effet, si on dépollue les eaux du lac sans éliminer les sources de pollution, il est clair que le lac se polluera de nouveau.

On peut appliquer le même concept à une base de données et aux processus d'acquisition de l'information connexes. Pour améliorer la qualité d'une base de données, il faut d'abord s'assurer que ces processus permettent d'introduire des données de qualité dans la base.

On peut pousser encore plus loin l'analogie avec le lac. Le fait d'introduire uniquement de l'eau saine dans un lac aura pour effet général d'assainir ce dernier. Cet assainissement est facilité par les processus naturels (biologiques, chimiques, etc.) qui contribuent à éliminer l'eau à mesure qu'elle est réintroduite par les affluents. Dans les bases de données, le «roulement» des entités représentées, des valeurs des données, etc., a le même effet.

C'est en nous inspirant de cette analogie et en nous fondant sur les concepts relatifs à la gestion des processus (Process Control Management and Improvement Guidelines), que nous avons adopté l'approche de haut niveau suivante en vue d'améliorer la qualité des données comprises dans une base de données.

1. Soumettre les processus d'introduction des données dans la base à un contrôle statistique.

2. Si les niveaux de qualité observés ne satisfont pas aux exigences, améliorer le processus jusqu'à ce qu'ils y satisfassent. S'assurer que le processus amélioré fait l'objet d'un contrôle statistique.

3. Il est maintenant possible de prédire la qualité des données qui feront partie de la base à un moment quelconque dans le futur. Si cette qualité n'est pas satisfaisante, axer les efforts d'épuration de la base de données sur les segments pour lesquels le roulement est relativement faible.

Il importe de souligner le rôle fondamental que joue le contrôle statistique. Premièrement, il nous permet de prédire les niveaux de qualité futurs (tant de la base de données que des processus) et donc, d'établir une comparaison significative entre ces niveaux et les exigences établies. Deuxièmement, il nous permet de recueillir les données nécessaires pour élaborer les mesures visant à améliorer la qualité et déterminer les zones d'application de ces mesures. Le suivi des données a donc pour objectif principal d'établir un contrôle statistique.

Les étapes clés de la mise en oeuvre du suivi des données sont les suivantes:

1. Prélever un échantillon aléatoire d'enregistrements au début du processus.

2. Suivre les enregistrements sélectionnés tout au long du processus. À cette fin, noter le contenu de chaque enregistrement sélectionné avant et après qu'il ait été soumis à chaque sous-processus. Dans chaque cas, noter également l'heure du début et de la fin du traitement.

3. Déceler les défectuosités et les erreurs engendrées par le processus et les sous-processus qui le composent.

4. Résumer, à des intervalles appropriés, le cheminement des enregistrements sélectionnés à l'intérieur du processus. Élaborer les graphiques et les sommaires nécessaires pour vérifier la conformité aux normes et analyser les modifications ou les erreurs afin de réétablir le contrôle et proposer de nouvelles façons d'améliorer le processus.

Le reste de cet article traite de la mise en oeuvre de ces étapes. La présente section est consacrée à l'examen des étapes 1 et 2, qui ont pour objet le rassemblement des données relatives au suivi.

2.1 Étape 1: Échantillonnage

Nous avons choisi d'utiliser la technique de l'échantillonnage parce qu'elle présente de nombreux avantages par rapport au contrôle complet (Cochran, W.G., 1977). Les caractéristiques du plan d'échantillonnage seront déterminées en fonction de nombreux facteurs, comme la structure de la population au sein de laquelle l'échantillon doit être prélevé, la nature du processus et le degré de précision requis. Ainsi, nombre de processus informatiques assurent un «traitement par lots», en ce sens que les enregistrements y sont traités par groupes. En pareil cas, il convient de prélever un échantillon aléatoire au sein de chaque lot. Si les enregistrements sont traités «un à un», il est peut-être plus approprié de choisir les enregistrements complètement au hasard, selon une probabilité de sélection prédéterminée.

2.2 Étape 2: Suivi

Une fois les enregistrements sélectionnés, il faut leur attribuer un indicateur (ou «drapeau») spécial, afin qu'il soit possible de suivre leur cheminement à l'intérieur du processus. Cet indicateur peut prendre une des deux formes suivantes.

- **Zone unique d'identification**. S'il existe une zone ou un ensemble de zones servant d'identificateur unique pour chaque enregistrement et si cette zone ne peut être modifiée au cours du processus, on peut l'utiliser pour caractériser les enregistrements sélectionnés. Lorsqu'on utilise cette méthode, les valeurs figurant dans cette zone pour les enregistrements sélectionnés sont stockées en mémoire pour consultation ultérieure et il n'est pas nécessaire de modifier la structure des données.

- **Zone d'indication**. On peut ajouter à chaque enregistrement une zone spéciale destinée à signaler si l'enregistrement a été sélectionné ou non. Cette méthode permet de repérer les enregistrements facilement et rapidement tout au long du processus. Cependant, il faut modifier la structure des données pour introduire la zone d'indication et il est important que cette zone soit cachée aux personnes qui utilisent le processus.

Tout au long du processus, on enregistre, au début et à la fin de chaque sous-processus, la date, l'heure et le contenu des enregistrements sélectionnés. L'échantillonnage et le suivi peuvent se faire à la main ou de diverses façons plus ou moins automatisées. Nous allons maintenant décrire brièvement une de ces approches, que nous avons appelée le système *décentralisé de saisie des images* (DSI). On peut utiliser le système DSI:

- lorsqu'il est possible d'avoir recours à des zones d'identification, comme il est décrit plus haut; et,

- lorsque chacun des sous-processus composant le processus est au moins partiellement informatisé[4].

Nous avons choisi d'étudier ce système parce que nombre de processus importants satisfont déjà à ces critères et que nous croyons que ce nombre va aller en augmentant dans le futur.

Essentiellement, la mise en oeuvre du système DSI consiste à intégrer aux systèmes informatisés relatifs à chaque sous-processus les éléments suivants.

---

[4] Le processus d'accès respecte ces critères.

- Un «échantillonneur». Cet élément, qui est intégré au premier sous-processus, détermine si un enregistrement doit être suivi ou non et introduit la valeur appropriée dans la «zone d'indication».

- Des «filtres» qui sont capables de reconnaître les enregistrements sélectionnés et de saisir leur contenu au moment voulu. Ces filtres, qui sont intégrés à chaque sous-processus, saisissent le contenu des enregistrements lorsque le sous-processus est terminé. Les filtres doivent également noter la date et l'heure du lancement et de l'achèvement du sous-processus.

- Un logiciel de «transmission». Ce logiciel, qui transmet les enregistrements saisis dans une mémoire auxiliaire, doit également être intégré à chaque sous-processus.

Dans la mémoire auxiliaire, les enregistrements relatifs à chaque sous-processus sont rassemblés et stockés en vue de leur analyse. On trouve à la figure 4 un exemple d'enregistrement ayant été suivi tout au long du processus d'accès. Nous décrivons aux sections suivantes les analyses qu'il faut réaliser pour qu'il soit possible d'améliorer de façon durable l'exactitude et la cohérence (section 3) ainsi que l'actualité (section 4) des données.

3. EXACTITUDE ET COHÉRENCE

Les renseignements recueillis tout au long du processus servent à quantifier l'exactitude, la cohérence et l'actualité des données. Dans le cas de l'exactitude, l'accent est placé sur les modifications survenues dans les zones lorsque les enregistrements cheminent dans une seule base de données. Dans le cas de la cohérence, l'accent est placé sur les modifications survenues dans les zones lorsque les processus débouchent sur des bases de données distinctes. Conceptuellement, comme nous le verrons dans la présente section, on adopte une approche similaire pour l'analyse des données de suivi relatives à l'exactitude et celle des données de suivi relatives à la cohérence.

Dans le cas de l'actualité, l'accent est placé sur les renseignements relatifs à la date et à l'heure. L'approche adoptée aux fins de l'analyse de ces données est exposée à la section 4.

3.1 Étape 3: Repérage des erreurs

Les zones de données peuvent faire l'objet de nombreuses modifications susceptibles d'entraîner des imprécisions et/ou des divergences. Ces modifications peuvent être de trois types.

- **Modifications de normalisation**. Modifications, comme l'insertion ou la suppression de bornes, d'espaces, etc., apportées afin de respecter les divers formats de présentation utilisés par les systèmes informatiques auxquels fait appel le processus. Si on se reporte à la figure 4, on verra que des modifications de ce type ont été apportées dans les zones CKR et CKL au cours du sous-processus DSA.

- **Modifications relatives à la traduction**. Modifications rendues nécessaires du fait des langages différents utilisés tout au long du processus, comme le langage USO utilisé par AT&T et le langage ISI utilisé dans l'industrie des télécommunications entre les sociétés exploitantes (p. ex., AT&T, MCI, Sprint) et les compagnies exploitant les centres locaux (p. ex., NJ Bell, IL Bell). Si on se reporte à la figure 4, on verra que des modifications de ce type ont été faites dans les zones IX IND et S25 au cours du sous-processus DSA.

- **Modifications opérationnelles parasites**. Modifications qui surviennent lorsqu'on change incorrectement la valeur d'un élément d'information. Ces modifications peuvent être classées comme suit:

  -- correction d'une erreur introduite à une étape antérieure;
  -- introduction d'une erreur à l'étape où les modifications sont apportées.

  Quoi qu'il en soit, chaque modification opérationnelle parasite indique la présence d'une erreur dans le processus. Si on se reporte à la figure 4, on verra que des modifications de ce type ont été apportées dans les zones BAN et LSO SECLOC au cours du sous-processus CCF.

Figure 4: Exemple d'enregistrement "suivi"

ZONE	CSU	DSA	CF	CCF	INV	DÉPENSES	RECETTES
CKR	DHBC 728534 ATI	DHBC 728534 ATI	DHBC 728534 ATI	DHBC 728534 ATI	DHBC 728534 ATI	DHBC 728534 ATI	DHBC 728534 ATI
EC CKT				70 HCCB 098012 249 PT	70 HCCB 098012 249 PT		70 HCCB 098012 249 PT
ACTL			SNORCAR2	SNORCAR2	SNORCAR2		SNORCAR2
SBCTYP			E	E	E		E
LSO CLU			SNTOCA01	SNTOCA01	SNTOCA01		SNTOCA01
MSL		08	08	08	08		08
NCI			04D09 15K -	04D09 15K -	04D09 15K -		04D09 15K -
SEC NCI			04DUB C -	04DUB C -	04DUB C -		04DUB C -
NC			HC -	HC -	HC -		04DUB C -
SEC LOC	DIGITAL EQUIP CORP		DIGITAL EQUIP CORP	DIGITAL EQUIP CORP	DIGITAL EQUIP CORP	DIGITAL EQUIP CORP	DIGITAL EQUIP CORP
PU		100	100	100	100		100
BKND	NB	L	L	L	L	NB	L
SAN			272-791-9088	272-791-9100	272-791-9100	272-791-9088	272-791-9100
LATA					722		
MCN	229510	229510	229510	229510	229510	229510	229510
S25	SRBEX	A	A	A	A	SRBEX	A
ICSC			FT04	FT04	FT04		FT04
LSO SECLOC	408727	408 727	408 727	408 970	408 970	408727	408 970
ACT	N	N	N	N	N	N	N
PON			FW308818313	FW 308818313	FW308818313		FW308818313
CKL	1	9001	00081	0081	0081	1	00081
SVC			SPBC	SPBC	SPBC		SPBC
[illegible]	02/24/89 10:00	02/24/89 10:45	3/08/90 8:00	3/20/89 9:00	3/31/89 8:00	02/24/89 10:00	3/20/89 9:00
[illegible]	02/24/89 10:45	02/24/89 17:15	3/08/90 12:30	3/20/89 12:30	4/01/89	02/24/89	3/20/89
[illegible]		02/24/89 17:00	3/08/90 17:00	0/20/89 17:00	4/01/89 8:00		0/20/89 17:00

Les modifications apportées à l'enregistrement sont indiquées en caractères gras. La modification apportée dans la zone CKR à l'interface du sous-processus CSU et du sous-processus DSA est un exemple de modificaton de normalisation; la modification apportée dans la zone S25 à l'interface du sous-processus CSU et du sous-processus DSA est une modification relative à la traduction; enfin, la modification apportée dans la zone LSO SECLOC à l'interface du sous-processus CF et du sous-processus CCF est une modification opérationnelle parasite.

3.2 Étape 4: Récapitulation des résultats

Les données recueillies peuvent faire l'objet de trois niveaux d'analyse:

- une analyse «métrologique» visant à poser un diagnostic sur l'état général du processus;
- une analyse de «localisation» visant à déterminer quels sous-processus (plus précisément, quelles paires de sous-processus) et quelles combinaisons de zones posent les problèmes les plus graves; enfin,
- une analyse de «contrôle» visant à établir un contrôle statistique et à fixer des objectifs d'amélioration.

Les résultats des analyses plus générales sont utilisés pour déterminer quelles analyses détaillées il convient de réaliser, toutes les analyses ayant pour objet d'aider le propriétaire du processus[5] à déterminer les mesures à prendre. Nous verrons dans les deux prochaines sous-sections quels genres de graphiques se sont révélés utiles à cette fin.

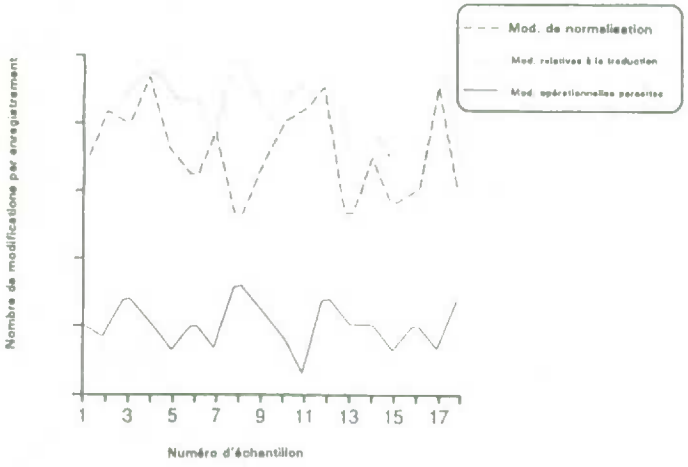
[5] Dans le présent article, on entend par propriétaire du processus la personne ou le groupe de personnes responsable du processus.

3.2.1 Exactitude

La présente sous-section porte sur les aspects du processus d'accès illustrés à la figure 1.

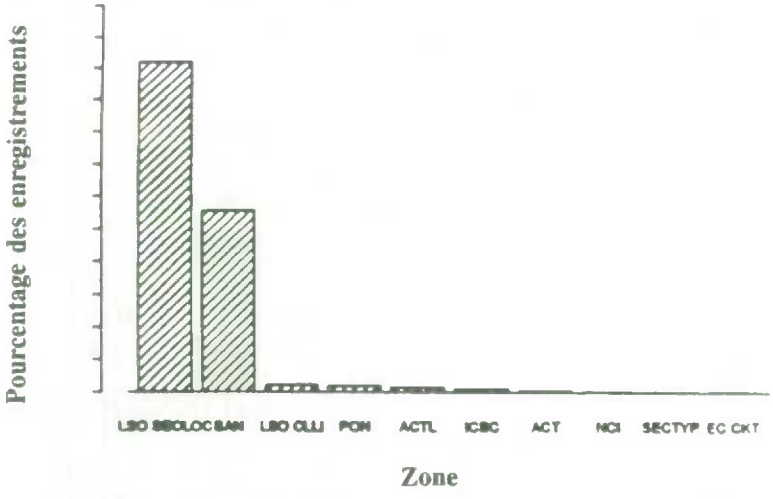
On trouve à la figure 5 un exemple de graphique utile pour l'analyse métrologique. Ce graphique illustre les fluctuations du nombre moyen de modifications de chaque type (de normalisation, relatives à la traduction et opérationnelles parasites) en fonction du temps. Après avoir étudié quelles étaient les principales raisons nécessitant l'apport de modifications de normalisation et de modifications relatives à la traduction, nous avons pu déterminer qu'il s'agissait des variations des exigences relatives au format et au langage machine d'un sous-processus à l'autre. Comme ces modifications sont moins graves[6], nous ne les étudierons pas plus avant. Nous allons plutôt nous concentrer sur les modifications opérationnelles parasites.

Figure 5: Ensemble des modifications apportées au cours du processus d'accès: CSU - DB d'inventaire



Représentation graphique des variations en fonction du temps du nombre moyen de modifications de chaque type par enregistrement pour le segment CSU - BD d'inventaire du processus d'accès.

Figure 6: Modifications opérationnelles parasites: CSU - DB d'inventaire



Pourcentage des enregistrements ayant fait l'objet de modifications opérationnelles parasites au cours du segment CSU - BD d'inventaire du processus d'accès.

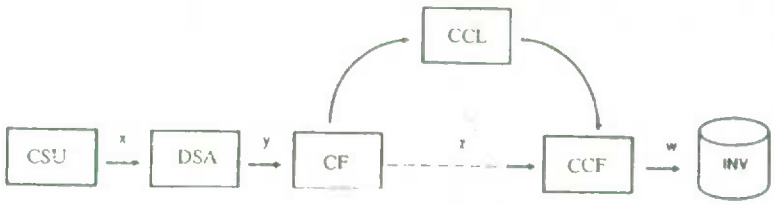
---

[6] Nous n'avons nullement l'intention de laisser croire que les modifications apportées pour des fins de normalisation ou de traduction sont sans importance. Ces modifications se traduisent par un accroissement des délais et des coûts d'élaboration du logiciel et de traitement des données et constituent autant d'occasions d'erreur.

Comme nous l'avons vu, l'analyse de localisation a pour objet de déterminer quels sous-processus ou combinaisons de zones sont les plus susceptibles d'être améliorés. Il est possible d'adopter diverses approches à cette fin. Ainsi, on peut d'abord reporter sur un graphique de Pareto les pourcentages d'enregistrements faisant l'objet de modifications opérationnelles parasites pour chaque zone de données. On trouve à la figure 6 un tel graphique, qui confirme que ce sont bien les zones LSO SECLOC et BAN[7] qui offrent les meilleures possibilités d'amélioration. Nous allons pour l'instant nous confiner à l'examen de la zone BAN et nous étudierons la zone LSO SECLOC à la sous-section suivante portant sur la cohérence.

Deuxièmement, on peut également reporter sur un graphique de Pareto les pourcentages d'enregistrements dans lesquels ces zones ont été modifiées d'un sous-processus à l'autre. Il s'est avéré utile d'indiquer ces chiffres directement sur l'organigramme d'analyse. Le graphique ainsi obtenu (figure 7) est appelé «Sommaire des modifications apportées au cours du processus».

Figure 7: Sommaire des modifications apportées au cours du processus: zone BAN



LÉGENDE

CSU	-	**Commande de service universel**
DSA	-	**Demande de service d'accès**
CF	-	**Commande ferme**
CCL	-	**Activités de la compagnie de téléphone**
CCF	-	**Confirmation de commande ferme**
INV	-	**Base de données d'inventaire**

Pourcentage d'enregistrements dans lesquels la zone BAN a été modifiée à l'interface de deux sous-processus successifs du segment CSU - BD d'inventaire du processus d'accès. Ici, z> > x, y, w (qui sont tous égaux à zéro ou très petits).

Au troisième niveau d'analyse, on utilise des cartes de contrôle p (Statistical Quality Control Handbook) (on trouve à la figure 8 la carte de contrôle p des modifications apportées à la zone BAN à l'interface des sous-processus CF et CCF) sur lesquelles la ligne centrale (LC) continue représente la proportion moyenne d'enregistrements défectueux pour l'ensemble des échantillons, tandis que les lignes discontinues situées de chaque côté de cette dernière représentent respectivement les limites de contrôle supérieure et inférieure (LCS, LCI). Les limites de contrôle servent entre autres à déterminer les points hors contrôle, qui sont indiqués au moyen d'étoiles. Les limites de contrôle discontinues illustrées sur ces cartes s'expliquent par les variations de la taille des échantillons. Le lecteur désireux d'obtenir plus de renseignements sur la sélection d'échantillons de tailles appropriées, le calcul des limites de contrôle, la détermination des points hors contrôle et l'interprétation statistique de ces données est invité à consulter des ouvrages comme *Statistical Quality Control Handbook* ou Wadsworth, Stephens et Godfrey (1986). On peut générer les cartes de contrôle et les analyses statistiques connexes à l'aide des logiciels offerts sur le marché, comme le *SQC Troubleshooter*.

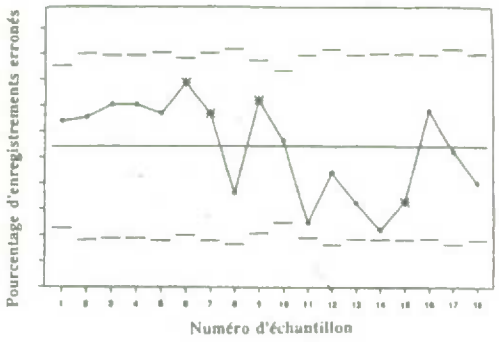
Comme nous l'avons vu plus haut, on trouve à la figure 8 la carte de contrôle p des modifications de la zone BAN effectuées à l'interface des sous-processus CF et CCF. Selon cette carte, le processus affiche un taux moyen élevé d'enregistrements défectueux et on y décèle de nombreux points hors contrôle. Afin de dégager la cause fondamentale de ce phénomène, nous avons généré de nombreux graphiques représentant les variations du nombre moyen de modifications apportées à la zone BAN en fonction de diverses variables. Ainsi, on trouve

---

[7] Les lettres BAN désignent le numéro de facturation (Billing Account Number), tandis que la zone LSO SECLOC indique le bureau de la compagnie de téléphone assurant le service.

à la figure 9 une répartition de ces modifications selon la région, indiquant que le nombre moyen de ces modifications varie considérablement d'une région administrative à l'autre. La découverte de cette variation nous a permis d'améliorer le processus de façon substantielle en faisant part aux autres régions de la méthode utilisée dans la région 6 pour traiter la zone BAN. Or, cette méthode s'est révélée facile à transférer dans d'autres régions et ce transfert s'est traduit par une amélioration sensible de la qualité des données au cours des mois suivants.

Figure 8: Carte de contrôle p: Modifications apportées à la zone BAN dans le cadre du sous-processus CCF

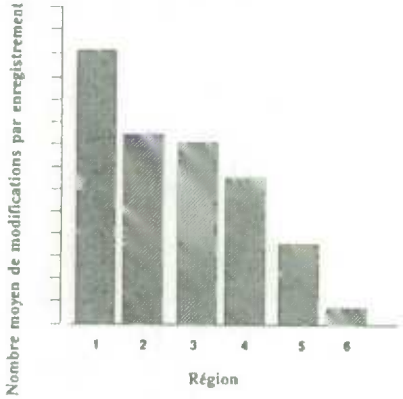


Carte de contrôle p = Modifications de la zone BAN à l'interface du sous-processus «Commande ferme» et du sous-processus «Confirmation de commande ferme».

3.2.2 Cohérence

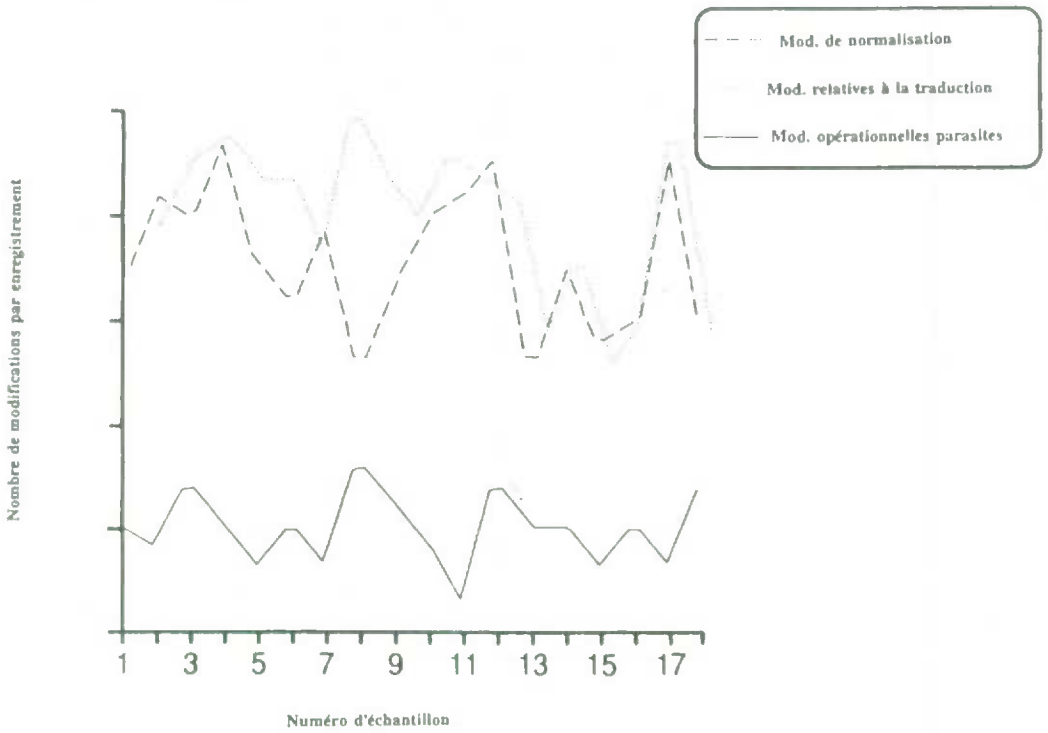
Nous allons maintenant examiner les aspects du processus d'accès illustrés à la figure 2. L'analyse relative à la cohérence s'effectue selon un cheminement très semblable à celui observé pour l'analyse relative à l'exactitude. Premièrement, sur le plan métrologique, on peut représenter en fonction du temps le nombre total d'incohérences observées dans chacune des catégories de modifications (de normalisation, relatives à la traduction et opérationnelles parasites). On trouve un tel graphique à la figure 10. Encore une fois, nous nous confinerons à l'étude des modifications opérationnelles parasites. Dans le cadre de l'analyse de localisation, nous avons généré un graphique de Pareto (figure 11) et un sommaire des modifications apportées au cours du processus (figure 12) qui sont venus confirmer que c'était bien la zone LSO SECLOC qui comportait le plus grand nombre d'incohérences et que la source de ces incohérences se situait à l'interface des sous-processus CF et CCF.

Figure 9: Modifications opérationnelles parasites de la zone BAN dans le cadre du sous-processus CCF selon la région



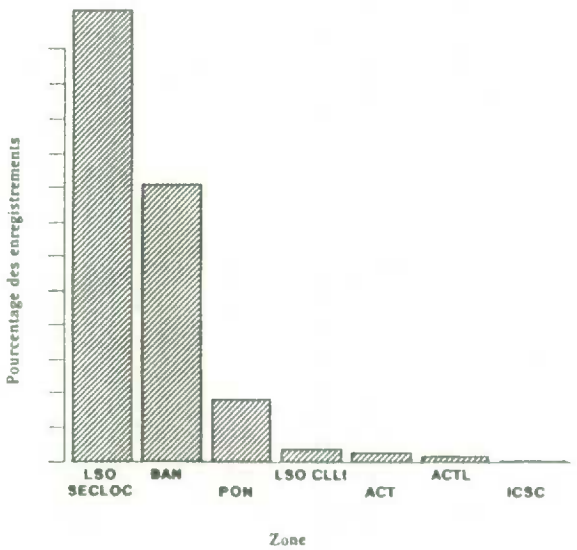
Nombre moyen de modifications opérationnelles parasites de la zone BAN par enregistrement à l'interface des sous-processus «Commande ferme» et «Confirmation de commande ferme» pour diverses régions du pays.

Figure 10: Nombre total de modifications apportées à l'interface des DB sur les dépenses et sur les recettes



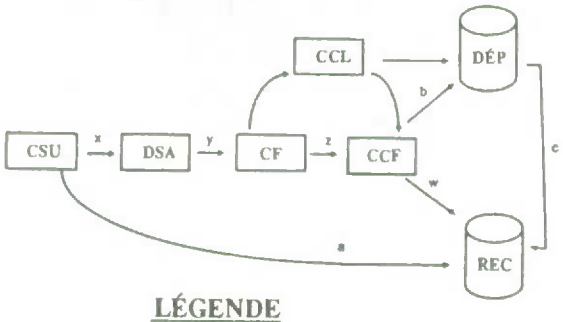
Variation en fonction du temps du nombre moyen de modifications de chaque type apportées par enregistrement au cours du segment CSU - BD sur les dépenses et sur les recettes du processus d'accès.

Figure 11: Modifications opérationnelles parasites apportées 1a l'interface des DB sur les dépenses et sur les recettes



Pourcentage d'enregistrements faisant l'objet de modifications opérationnelles parasites au cours du segment CSU - BD sur les dépenses et sur les recettes du processus d'accès.

Figure 12: Sommaire des modifications apportées au cours du processus: CSU - DB sur les dépenses et sur les recettes



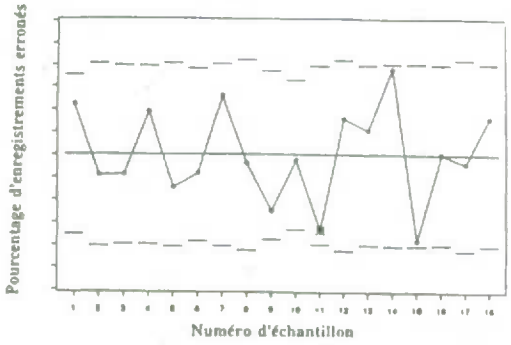
LÉGENDE

CSU	-	**Commande de service universel**
DSA	-	**Demande de service d'accès**
CF	-	**Commande ferme**
CCL	-	**Activités de la compagnie de téléphone**
CCF	-	**Confirmation de commande ferme**
DÉP	-	**Base de données sur les dépenses**
REC	-	**Base de données sur les recettes**

Sommaire des modifications apportées à la zone LSO SECLOC au cours du segment CSU - BD sur les dépenses et sur les recettes du processus d'accès Ici, c > z > > b > a > x, y.

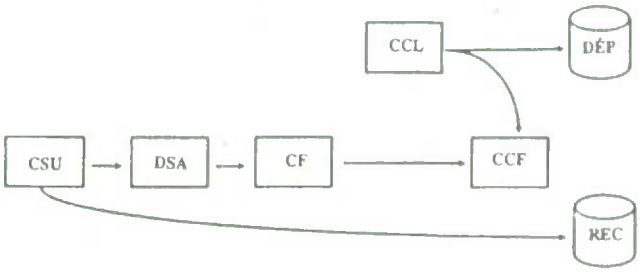
Selon la carte de contrôle p établie pour la zone LSO SECLOC à l'interface des sous-programmes CF et CCF (figure 13), la zone LSO SECLOC semble être «statistiquement sous contrôle», bien que le taux d'erreurs soit élevé. En d'autres termes, il ne semble exister aucun motif spécial de variation. (Bien qu'une étoile indique que l'échantillon n° 11 est hors contrôle, il a été impossible d'en déterminer la cause.) Le comportement apparemment «aléatoire» de ce processus laisse supposer que ce taux d'erreurs est inhérent au processus, tel qu'il est actuellement défini. Le bien-fondé de cette inférence a subséquemment été confirmé: à la suite d'une étude plus approfondie, il s'est avéré que les données de la zone LSO SECLOC ne sont pas transmises du sous-processus CF au sous-processus CCF, mais qu'elles sont plutôt déterminées une deuxième fois dans le cadre du sous-processus CCF. Comme l'illustre la figure 14, il y a donc discontinuité du processus d'accès pour la zone LSO SECLOC à l'interface des sous-processus CF et CCF. En conséquence, la correspondance entre les deux zones LSO SECLOC est aléatoire et cette caractéristique explique le comportement observé sur la figure 13.

Figure 13: Carte de contrôle p: Modifications apportés à la zone LSO SECLOC dans le cadre du sous-processus CCF



Carte de contrôle p = Modifications de la zone LSO SECLOC à l'interface du sous-processus «Commande ferme» et «Confirmation de commande ferme». Bien qu'une étoile indique que le point 11 est hors contrôle (il ne respecte par la règle selon laquelle «deux points parmi trois doivent tomber en dehors de la limite d'avertissement de 2»), il a été impossible d'attribuer cette variation à une cause spéciale. Des analyses ultérieures ont toutefois démontré que le contenu de la zone LSO SECLOC est déterminé indépendamment au cours des sous-processus «Commande ferme» et «Confirmation de commande ferme».

Figure 14: Fonctionnement du processus d'accès pour la zone LSO SECLOC



LÉGENDE

CSU	-	**Commande de service universel**
DSA	-	**Demande de service d'accès**
CF	-	**Commande ferme**
CCL	-	**Activités de la compagnie de téléphone**
CCF	-	**Confirmation de commande ferme**
DÉP	-	**Base de données sur les dépenses**
REC	-	**Base de données sur les recettes**

Fonctionnement réel du segment CSU - BD sur les dépenses et les recettes du processus d'accès pour la zone LSO SECLOC. La figure illustre bien qu'il y a discontinuité à l'interface des sous-processus CCF et CCL.

Afin d'éliminer cette source d'incohérences, le propriétaire du processus se doit de régler ce problème de discontinuité.

4. ACTUALITÉ

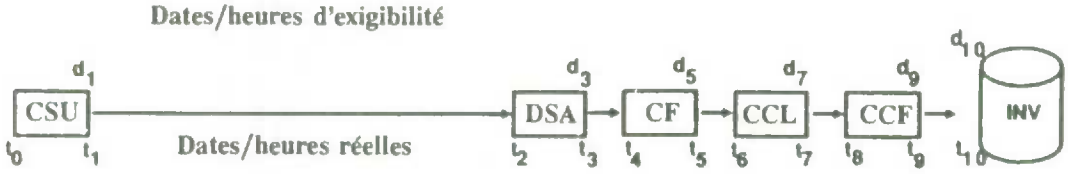
Nous allons maintenant étudier les aspects du processus d'accès illustrés à la figure 3. Dans ce cas-ci, l'accent est placé sur le caractère actuel des données et, en particulier, sur les délais d'exécution et la rapidité du processus.

Une des caractéristiques du premier sous-processus (le sous-processus CSU) permet au client de préciser (sous réserve de certaines limites) le moment auquel il souhaite que le service lui soit offert. À l'aide de ces données, le sous-processus CSU fixe les dates/heures d'exigibilité de chacune des activités à valeur ajoutée. Comme à la figure 3, définissons par d_1 l'heure à laquelle l'exécution du sous-processus CSU doit être terminée, par d_2 l'heure à laquelle l'exécution du sous-processus DSA doit être terminée, et ainsi de suite (c.-à-d. les dates/heures d'exigibilité). Il importe de noter que le client a la possibilité de modifier sa commande jusqu'au moment de la date/heure d'exigibilité. Par ailleurs, le suivi des données peut nous permettre d'obtenir les dates/heures réelles de lancement et d'achèvement des activités. Comme à la figure 3, définissons par t_1 l'heure réelle d'achèvement du sous-processus CSU, par t_2 l'heure réelle de lancement du sous-processus DSA, et ainsi de suite.

Il est particulièrement intéressant de soumettre les données sur les dates/heures réelles d'exécution et les dates/heures d'exigibilité à deux types d'analyses. La première porte sur le respect des dates/heures d'exigibilité. Ainsi, on détermine si les délais prévus pour la prestation du service commandé par un client ont été respectés au moyen de la différence t_{10}-d_{10}. Il faut porter une attention particulière aux valeurs positives de cette différence (c.-à-d., $t_{10} > d_{10}$), car elles indiquent que le client ne peut obtenir le service au moment où il le désire.

La deuxième porte sur la différence entre les heures réelles consécutives. La différence entre l'heure réelle de lancement et l'heure réelle d'achèvement d'un sous-processus donné indique la «durée» de ce sous-processus, tandis que la différence entre l'heure réelle de lancement d'un sous-processus donné et l'heure d'achèvement du sous-processus précédent indique le «délai d'attente».

Figure 15: Fonctionnement idéal du processus d'accès: Actualité



LÉGENDE

CSU	-	Commande de service universel
DSA	-	Demande de service d'accès
CF	-	Commande ferme
CCL	-	Activités de la compagnie de téléphone
CCF	-	Confirmation de commande ferme
INV	-	Base de données d'inventaire

Avec un nouveau programmateur, tous les délais d'attente relatifs au processus sont reportés au début de celui-ci. Les opérations subséquentes sont programmées de façon à être exécutées les unes à la suite des autres et à ce que le service soit offert au moment requis.

Encore une fois, on peut soumettre les données sur le respect des dates/heures d'exigibilité, la durée et les délais d'attente à la procédure d'analyse en trois étapes décrite et illustrée plus haut. Nous nous abstiendrons toutefois d'effectuer de nouveau ces analyses et nous contenterons d'indiquer que, après avoir analysé la «durée» et les «délais d'attente», il est possible de planifier l'exécution du processus différemment. En pareil cas, cette planification a pour objectifs:

1. de réduire le plus possible les délais d'attente entre deux sous-processus, pour plutôt reporter ces délais au début du processus;

2. de fixer l'heure de lancement des activités à valeur ajoutée de sorte que leur exécution soit terminée à la date/heure précisée par le client.

Idéalement, le processus devrait s'exécuter comme il est illustré à la figure 15.

5. CONCLUSION

Le présent article donne une description du suivi des données, nouvelle technique qui permet d'améliorer de façon durable la qualité des données faisant partie de bases de données de grande envergure. Essentiellement, le suivi des données vise à soumettre à un examen attentif les processus dits d'information selon lesquels les données sont introduites dans ces bases. La technique consiste à prélever un échantillon parmi les enregistrements soumis au processus, puis à marquer ces enregistrements de façon à pouvoir les suivre tout au long du processus, c'est-à-dire à pouvoir en noter le contenu à l'achèvement de chacun des sous-processus. L'analyse subséquente des données ainsi recueillies peut permettre de trouver de nouvelles façons d'accroître l'exactitude des données, la cohérence entre les données faisant partie de deux bases distinctes, ainsi que la rapidité d'exécution de l'ensemble du processus (actualité des données). En outre, cette technique permet d'établir un contrôle statistique de la qualité afin de veiller à ce que les améliorations apportées soient durables.

REMERCIEMENTS

Les auteurs tiennent à remercier leurs collègues et collaborateurs des Laboratoires AT&T Bell et de la Division des services du réseau de AT&T pour leur soutien et leurs encouragements. Steve Borbash, Errol Caby, Young Huh, Arnold Lent et Anany Levitin ont apporté nombre de nouvelles idées et contribué à en développer d'autres, Michele Bentley, Darrell Link, Patti McKnight, June Slonin, Lynda Snyder et Beverly Weber ont pour leur part mis en pratique plusieurs des idées exposées dans cet article et les applications qu'ils en ont faites ont permis d'apporter de nombreuses améliorations aux méthodes décrites. Les auteurs tiennent aussi à remercier Monica Falkenthal, Kerin Montgomery et Tim Neet pour leur aide et leur soutien. Enfin, ils souhaitent adresser des remerciements spéciaux à Scott Williamson, sans la perspicacité duquel le présent ouvrage n'aurait jamais vu le jour.

BIBLIOGRAPHIE

Ballou, D.P., et Tayi, G.K. (1989). *Commun. ACM* 32, 3, 320-329.

Caby, E. A Framework for Data Consistency, texte en préparation.

Cochran, W.G. (1977). *Sampling techniques*, Third Edition, New York: John Wiley & Sons.

Huh, Y.U., Keller, F.R., Redman, T.C., et Watkins, A.R. (1990). Data Quality, *Information and Software Technology*. 32, 8, 559-565.

Huh, Y.U., Pautke, R.W., et Redman, T.C. Data Quality Control, soumis au *Journal of Quality technology*.

Johnson, J.R., Leitch, R.A., et Neter, J. (1981). Characteristics of errors in Accounts Receivables and Inventory Audits, *Account. Rev.*, 56, 2, 270-293.

Laudon, K.C. (1986). Data Quality and Due Process in Large Interorganizational Record Systems, *Commun. ACM* 29, 1, 4-18.

Levitin, A.V., et Redman, T.C. The Notion of Data and Its Quality Dimensions, texte en préparation.

Liepens, G.E. (1989). Sound Data are a Sound Investment, *Quality Progress*, 22, 9, 61-64.

Liepens, G.E., Garfinkel, R.S., et Kunnathur, A.S. (1982). Error Localization for Erroneous Data: A Survey, *TIMS/Studies in the Management Sciences*, 19, 205-219.

Morey, R.C. (1982). Estimating and Improving the Quality of Information in a MIS, *Commun. ACM*, 25, 5, 337-342.

Pritzker, L., Ogus, J., et Hansen, M.H. (1965). Computer Editing Methods - Some Applications and Results, *Bulletin of the International Statistical Institute*, 41, 442-465.

Process Control Management and Improvement Guidelines, Issue 1.1, (1988). AT&T.

Shewhart, W.A. (1939). *Statistical Method from the Viewpoint of Quality Control*, Graduate School of the Department of Agriculture, Washington, D.C.

SQC *Troubleshooter*, (1990). AT&T.

Statistical Quality Control Handbook, (1956). AT&T.

Svanks, M.I. (1988). Integrity Analysis, *Information & Software Technology*, 30, 10, 595-605.

Terry, M.E. (1963). The Principles of Statistical Analysis Using Large Electronic Computers, *Bulletin of the International Statistical Institute*, 40, 547-552.

Wadsworth, H.M., Stephens, K.S., et Godfrey, A.B. (1986). *Modern Methods for Quality Control and Improvement*, New York: John Wiley & Sons.

CONFÉRENCIER SPÉCIAL INVITÉ

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

MÉTHODES POUR ESTIMER LA PRÉCISION DES ESTIMATIONS D'ENQUÊTE LORSQU'IL Y A EU IMPUTATION

C.-E. Särndal[1]

RÉSUMÉ

Les chercheurs utilisent une forme ou l'autre d'imputation dans presque toutes les enquêtes de grande envergure. Le présent article expose une méthode d'estimation de la variance lorsqu'on utilise l'imputation simple (plutôt que l'imputation multiple) pour produire un ensemble de données complet. Sauf dans des cas vraiment exceptionnels, l'imputation ne permet jamais de reproduire les valeurs réelles. Pour les fins de cet article, nous considérons donc que l'erreur totale de l'estimation d'enquête est égale à la somme de l'erreur d'échantillonnage et de l'erreur d'imputation. En conséquence, nous calculons une variance globale égale à la somme de la variance d'échantillonnage et de la variance d'imputation. Notre objectif principal est donc d'estimer ces deux composantes à l'aide des données obtenues après imputation, c'est-à-dire des valeurs réellement observées et des valeurs imputées. Nous adoptons à cette fin une méthode axée sur un modèle en ce sens que la forme des estimateurs de la variance est déterminée à la fois par le modèle impliqué par la méthode d'imputation et par la loi de probabilité régissant la sélection de l'échantillon. Nous confirmons les résultats théoriques à l'aide d'une simulation de Monte Carlo.

MOTS CLÉS: Imputation simple; estimation de la variance; modèle d'imputation; inférence à partir d'un modèle.

1. DIVERS TYPES D'IMPUTATION

Le présent article fait état des travaux réalisés parallèlement à l'élaboration du Système généralisé d'estimation (SGE) de Statistique Canada. Comme on le voit dans Lavallée, Leblond et Reinhardt (1990), les estimations de la variance sont automatiquement calculées dans les divers modules d'estimation qui composent le SGE. Il était donc nécessaire d'élaborer des méthodes appropriées d'estimation de la variance pour les cas où l'ensemble de données comporte des valeurs imputées, ce qui se produit dans pratiquement toutes les enquêtes. Avant de poursuivre, je me dois de remercier M. Hidiroglou, P. Lavallée, Y. Leblond, H. Lee et G. Reinhardt pour leur collaboration et pour l'intérêt qu'ils ont manifesté à l'égard des travaux qui ont conduit à la rédaction de cet article.

Les deux principales approches utilisées pour établir des estimations lorsque certaines valeurs sont manquantes sont la pondération et l'imputation. Dans les ouvrages récents, les poids utilisés pour compenser pour la non-réponse sont d'ordinaire considérés comme l'inverse des probabilités de réponse associées à un mécanisme de réponse donné. Comme les probabilités de réponse sont d'ordinaire inconnues, il faut donc les estimer à partir des données dont on dispose. En revanche, l'imputation offre l'avantage de produire une matrice de données complètes. Une telle matrice simplifie le traitement des données, mais elle ne signifie pas nécessairement qu'on peut utiliser directement les «méthodes d'estimation ordinaires». En effet, les valeurs imputées sont obtenues à partir d'un échantillon et, à ce titre, elles possèdent leurs propres propriétés statistiques, comme une moyenne et une variance.

---

[1] C.-E. Särndal, Université de Montréal, C.P. 6128, Succursale A, Montréal, Québec H3C 3J7.

De nos jours, l'imputation est un outil dont l'utilisation est largement répandue. À cet égard, il est intéressant de noter les remarques de Pritzker, Ogus et Hansen (1965) concernant la politique adoptée par le US Bureau of the Census en matière d'imputation: «Fondamentalement, notre philosophie à l'égard du problème de ... l'imputation consiste à faire en sorte d'obtenir par observation directe les données relatives à une très forte proportion des agrégats à tabuler et d'assurer un contrôle de la qualité suffisant pour que pratiquement toute règle d'imputation raisonnable ... nous permette d'obtenir essentiellement les mêmes résultats... Pour ce qui est de l'imputation relative aux données du recensement et des enquêtes-échantillon, nous nous sommes fixé pour objectif de maintenir la proportion de données imputées à 1 ou 2%.»

Idéalement, nous devrions encore nous efforcer de maintenir le taux d'imputation aux environs de 1 ou 2%. Cependant, de nos jours, la majorité des enquêtes réalisées par les organismes statistiques d'envergure présentent des taux d'imputation beaucoup plus élevés. Or, à l'évidence, lorsque 30% des valeurs sont imputées, il est impossible de ne pas tenir compte des effets de l'imputation. Son principal inconvénient tient peut-être au fait qu'elle peut être la source d'une erreur systématique (biais) dans l'estimation ponctuelle. Toutefois, même s'il était possible de trouver une méthode d'imputation n'introduisant aucune erreur systématique appréciable, il nous faudrait tenir compte de l'effet souvent considérable qu'a l'imputation sur la précision (la variance) de l'estimation ponctuelle. Il est donc essentiel que nous disposions de méthodes d'estimation de la variance simples mais valides pour les données d'enquêtes comportant des valeurs imputées, afin d'être en mesure de déclarer avec précision les coefficients de variation des estimations d'enquête.

Au cours des dernières années, les chercheurs ont proposé une grande variété de méthodes d'imputation, que l'on peut distinguer les unes des autres selon différents critères. Ainsi, il est possible de classer ces méthodes en fonction du nombre d'imputations effectuées. Les méthodes *d'imputation simple* consistent à imputer une seule valeur pour chaque valeur manquante. On obtient ainsi une matrice de données complète, dans laquelle les valeurs imputées sont signalées au moyen d'un astérisque. Les estimations sont alors calculées à partir de l'ensemble de données complet. En revanche, l'*imputation multiple* consiste à imputer au moins deux valeurs pour chaque valeur manquante. On obtient ainsi plusieurs ensembles de données complets et les estimations sont calculées à partir de tous ces ensembles de données.

Les méthodes d'imputation se distinguent aussi les unes des autres de par le modèle qui leur est sous-jacent. Certaines méthodes, comme l'imputation par régression, l'imputation par le quotient ou l'imputation par la moyenne, utilisent un modèle *explicite*, tandis que d'autres, comme l'imputation par la méthode «hot deck» et l'imputation dite du plus proche voisin, utilisent un modèle seulement *implicite*. Ces distinctions sont très importantes pour le présent article.

Actuellement, Statistique Canada utilise des méthodes comme l'imputation dite du plus proche voisin, l'imputation par le quotient, l'imputation par la moyenne des valeurs courantes, l'imputation par la valeur historique, l'imputation par la moyenne des valeurs historiques des répondants et l'imputation par une tendance auxiliaire, qui sont toutes des méthodes d'imputation simple. Les valeurs imputées sont calculées par le Système généralisé de contrôle et d'imputation (SGCI), puis elles sont introduites dans le Système généralisé d'estimation (SGE), où les estimations ponctuelles et les estimations de la variance sont calculées par des modules d'estimation distincts. Le présent article est surtout consacré à l'imputation par le quotient qui constitue un exemple de méthode utilisant un modèle explicite.

2. RÉFLEXIONS SUR L'IMPUTATION MULTIPLE

C'est D.B. Rubin qui, vers 1977, a été le premier à proposer l'utilisation de l'imputation multiple. Ses idées sur le sujet sont exposées dans un certain nombre d'articles, en particulier dans Herzog et Rubin (1983) et Rubin (1986), ainsi que dans un livre, Rubin (1987). Tout comme l'imputation simple, l'imputation multiple présente à la fois des avantages et des inconvénients.

Selon Rubin (1986), un des inconvénients de l'imputation simple tient au fait que «... la valeur imputée ne peut à elle seule refléter l'incertitude sur la valeur à imputer. Si une valeur était vraiment appropriée, elle ne serait pas manquante. Ainsi, lorsqu'on assimile les valeurs imputées à des valeurs observées, on sous-estime

systématiquement l'élément d'incertitude même en supposant que l'on connaisse les motifs exacts de la non-réponse.»

L'imputation multiple est attrayante du fait qu'elle indique implicitement que les valeurs imputées présentent une certaine variabilité. Or, c'est précisément cette variabilité (la variabilité à l'intérieur des ensembles de données complets et entre ces ensembles) que mettent à profit les méthodes d'estimation de la variance proposées lorsqu'on utilise l'imputation multiple. Ces méthodes font une utilisation très riche des concepts statistiques de base. (Par ailleurs, il est possible de faire valoir que la sélection d'un échantillon introduit aussi un élément de variabilité. Toutefois, pour la plupart des enquêtes, il est impossible de tirer plus d'un échantillon et l'estimation doit être calculée à partir de cet échantillon unique.)

On peut démontrer à l'aide d'exemples simples que le fait d'assimiler les valeurs imputées aux valeurs observées peut se traduire par une grave sous-estimation de l'incertitude réelle; les chercheurs chargés de prélever les échantillons d'enquête sont depuis longtemps conscients de cette situation. En revanche, on ne peut nier que les utilisateurs assimilent parfois les valeurs imputées aux valeurs observées, ni que cette attitude se traduise par une indication fausse de la précision des estimations. Il est facile, avec les ordinateurs modernes, d'imputer des valeurs en observant une règle quelconque, mais il est plus difficile d'obtenir des estimations valides de la variance.

La citation rapportée plus haut porte à croire que le fait que les valeurs obtenues par imputation simple ne présentent aucune variation nous empêche d'obtenir des estimations raisonnables de la variance et nous amène nécessairement à sous-estimer l'élément d'incertitude. Je ne partage pas cette opinion. De fait, les méthodes que nous allons étudier démontrent qu'il est possible d'obtenir une estimation valide de la variance en utilisant l'imputation simple.

Toute méthode d'estimation de la variance utilisée lorsque certaines des valeurs étudiées sont imputées devrait posséder les propriétés suivantes: a) s'appuyer sur de solides fondements théoriques; b) être résistante aux hypothèses sous-tendant l'imputation; c) être pratique, aisée à mettre en oeuvre et facilement acceptable par les utilisateurs.

Bien que l'imputation multiple possède les propriétés a) et b), il est clair que, dans certains cas du moins, elle ne possède pas la propriété c). Pour les fins de l'élaboration du SGE, nous devons utiliser des procédures faciles à mettre en oeuvre et pouvant être facilement acceptées par l'utilisateur, Or, l'utilisateur d'un ensemble de données (quelqu'un dont le principal domaine de spécialisation n'est pas la statistique) peut facilement comprendre que le statisticien procède à une imputation simple afin de remplacer une valeur manquante par la meilleure valeur possible. Bien qu'il puisse être intéressant, pour certaines fins comme la réalisation d'analyses secondaires, de disposer de plusieurs matrices de données complètes, cette possibilité est souvent écartée en raison des coûts de stockage des ensembles de données multiples.

Il est fort possible que l'imputation multiple se révèle utile dans d'autres contextes et pour des raisons autres que les critères qui président à l'élaboration du SGE. L'imputation multiple constitue une façon de résoudre le problème de sous-estimation de la variance posé par l'imputation des données. Toutefois, elle ne représente pas la seule solution. À cet égard, examinons maintenant, à l'aide d'une méthode élaborée à partir de Särndal (1990), quelles sont les possibilités offertes par les méthodes d'imputation simple.

3. VARIANCE D'IMPUTATION ET VARIANCE D'ÉCHANTILLONNAGE

Chaque règle d'imputation correspond à un modèle (explicite ou implicite) décrivant la relation entre les variables étudiées dans l'enquête. Ainsi, lorsque l'analyste formule une règle d'imputation, il se trouve du même coup à choisir un modèle. Le raisonnement que nous développons dans les pages qui suivent est fondé sur le principe suivant: si on considère la règle d'imputation valable pour le calcul des estimations ponctuelles (aucune erreur systématique), la règle est également valable pour établir les estimations de la variance correspondantes. En d'autres termes, il revient au constructeur du modèle de s'assurer simultanément de réduire le biais au maximum et d'établir des estimations de la variance appropriées. S'il est incapable de relever ce défi, il doit se garder d'imputer.

Soit $U = \{1, ..., k, ..., N\}$ une population finie et y une des variables étudiées dans l'enquête. Notre objectif est d'estimer le total de la variable y dans la population, $t = \Sigma_U y_k$. (Lorsque C est un ensemble quelconque d'unités de population, où $C \subseteq U$, on utilise Σ_C comme notation abrégée de $\Sigma_{k \in C}$; ainsi, $t = \Sigma_U y_k$ signifie $\Sigma_{k \in U} y_k$.) Nous prélevons un échantillon probabiliste s à l'aide d'un plan d'échantillonnage donné. Les probabilités d'inclusion sont connues et nous pourrions obtenir des estimations ordinaires de la variance selon le plan si toutes les unités $k \in s$ étaient observées. Toutefois, certaines données sont manquantes. Supposons que r est le sous-ensemble de s pour lequel les valeurs y_k sont réellement observées et que le sous-ensemble de $s - r$ est le sous-ensemble complémentaire pour lequel les valeurs sont imputées. Les *données après imputation* sont constituées des valeurs notées $y_{\phi k}$, $k \in s$, de telle sorte que

$$y_{\phi k} = \begin{cases} y_k & \text{si } k \in r \\ y_{imp,k} & \text{si } k \in s-r \end{cases}$$

où y_k est une valeur réellement observée et $y_{imp,k}$ désigne la valeur imputée pour l'unité k. Lorsque $r = s$, aucune donnée n'est imputée puisque toutes les données sont réellement observées.

Définissons l'estimateur de t pour le cas où toutes les données sont réellement observées (c.-à-d. où $r = s$) par $\hat{t} = \Sigma_{k \in s} w_k y_k = \Sigma_s w_k y_k$, où w_k est le poids attribué à la valeur observée y_k. Ainsi, pour un échantillon de n unités prélevé parmi N, au moyen d'un échantillon aléatoire simple sans remise (ÉASSR), $w_k = N/n$ pour tous les $k \in s$ lorsqu'on utilise la moyenne de l'échantillon élargi pour estimer t, et $w_k = (\bar{z}_U/\bar{z}_s)(N/n) = (\Sigma_U z_k)/(\Sigma_s z_k)$ pour tous les $k \in s$ lorsqu'on utilise l'estimateur par le quotient avec z comme variable auxiliaire.

Lorsque les données comprennent des valeurs imputées, l'estimateur de t est défini par $\hat{t}_\phi = \Sigma_s w_k y_{\phi k}$. En d'autres termes, nous supposons que les poids w_k sont identiques aux poids utilisés lorsque toutes les données sont réellement observées. Les modules d'estimation du SGE utilisent le même principe, qui traduit l'hypothèse selon laquelle la règle d'imputation choisie n'introduit aucune erreur systématique dans les estimations ou n'y introduit qu'une erreur systématique négligeable.

Comme la valeur imputée ne correspond pas à la valeur réelle y_k (sauf dans des circonstances vraiment exceptionnelles), l'imputation a pour effet d'accroître la variance du total estimé. Il suffit pour démontrer ce fait d'appliquer la règle d'imputation aux unités de l'échantillon réellement observées: elle introduit toujours une certaine erreur. Or, si la règle n'est pas exempte d'erreur pour les unités répondantes, elle n'en est pas plus exempte pour les unités non répondantes. Dans la section 4, nous exprimons la variance de $\hat{t}_\phi$ comme une somme de deux composantes, une variance d'échantillonnage et une variance attribuable à l'imputation,

$$V_{tot} = V_{éch} + V_{imp}.$$

La variance d'imputation V_{imp} est égale à zéro lorsque toutes les données sont des valeurs réellement observées ou lorsque la méthode d'imputation permet de reproduire exactement la valeur réelle y_k pour chaque unité nécessitant le recours à l'imputation. (En pratique, aucune de ces deux situations n'est susceptible de se présenter.) Selon la méthode exposée à la section 4, nous utilisons les données après imputation, $y_{\phi k}$, $k \in s$, pour estimer chacune des deux composantes et obtenir

$$\hat{V}_{tot} = \hat{V}_{éch} + \hat{V}_{imp}.$$

La composante $V_{éch}$ se calcule en deux étapes. Premièrement, nous calculons l'estimation ordinaire de la variance selon le plan à l'aide des données après imputation. (Ainsi, si on utilise un ÉASSR et si $r = s$,

l'estimation ordinaire non biaisée de la variance de $N\,\bar{y}_s$ est $N^2(1/n - 1/N)\,\Sigma_s\,(y_k - \bar{y}_s)^2/(n-1)$. En calculant cette formule pour les données après imputation, on obtient $N^2(1/n - 1/N)\,\Sigma_s(y_{\phi k} - \bar{y}_{\phi s})^2/(n-1)$, où $\bar{y}_{\phi s}$, est la moyenne des n valeurs $y_{\phi k}$.) Deuxièmement, nous ajoutons un terme pour tenir compte du fait que de nombreuses règles d'imputation se traduisent par l'obtention de données présentant une variabilité «moins que naturelle», ce qui pourrait entraîner une sous-estimation de la variance d'échantillonnage à moins que des mesures correctives ne soient prises. Enfin, nous calculons sans difficulté la composante $\hat{V}_{imp}$ à partir des données après imputation. L'utilisateur reconnaîtra sans peine que la variance obtenue à l'aide de la formule ordinaire n'est pas suffisante en soi et qu'il faut y ajouter un élément parce que la règle d'imputation n'est pas parfaite.

Une des propriétés intéressantes de la méthode tient au fait que si aucune imputation n'est requise, c'est-à-dire si $r = s$, $\hat{V}_{imp} = 0$ et $\hat{V}_{éch}$ est égale à l'«estimateur de la variance ordinaire» que nous aurions utilisé si toutes les valeurs avaient été réellement observées.

4. DÉVELOPPEMENTS THÉORIQUES

L'erreur totale de $\hat{t}_\phi$ se décompose comme suit

$$\hat{t}_\phi - t = (\hat{t} - t) + (\hat{t}_\phi - \hat{t}) = \text{erreur d'échantillonnage} + \text{erreur d'imputation.}$$

L'erreur d'imputation est égale à la différence entre l'estimation inconnue qu'on aurait calculée si les données avaient été entièrement constituées d'observations réelles et l'estimation qu'on peut calculer à partir des données après imputation. L'erreur d'imputation est définie par

$$\hat{t}_\phi - \hat{t} = -\Sigma_{s-r}\, w_k\, e_k$$

où

$$e_k = y_k - y_{imp,k}$$

est un *résidu d'imputation* qu'on ne peut observer pour une unité $k \in s - r$. L'amplitude de e_k est fonction de l'efficacité d'ajustement du modèle d'imputation. Si la méthode d'imputation permet d'obtenir des valeurs de remplacement presque parfaites, les résidus seront petits. Il est possible d'emprunter diverses directions pour développer le présent raisonnement. Dans ce cas-ci, nous utilisons une approche *axée sur un modèle* selon laquelle nous prenons en considération trois distributions de probabilités distinctes, dont les espérances mathématiques respectives sont notées E_ξ, E_s, et E_r, où ξ indique «selon le modèle d'imputation», s indique «selon le plan d'échantillonnage», et r indique «selon le mécanisme de réponse, étant donné s». Le modèle est impliqué par la règle d'imputation, il est donc connu; le plan d'échantillonnage correspond à la loi de probabilité qui régit la sélection de l'échantillon, il est donc aussi connu; le mécanisme de réponse est une loi de probabilité d'ordinaire inconnue régissant la réponse, étant donné l'échantillon s. Pour les fins de notre méthode, il n'est pas nécessaire de connaître le mécanisme de réponse et ce dernier peut être un mécanisme selon lequel la probabilité de réponse est systématiquement liée à la variable y, comme c'est le cas pour les mécanismes non aléatoires.

Supposons que $\hat{t}_\phi$ est globalement non biaisé, en ce sens que $E_\xi\, E_s\, E_r\, (\hat{t}_\phi - t) = 0$. La variance globale est alors définie par

$$V_{tot} = E_\xi\, E_s\, E_r\, \{(\hat{t}_\phi - t)^2\}.$$

On peut la décrire comme une «variance prévue» pour le plan d'échantillonnage et le mécanisme de réponse donnés. Nous obtenons

$$
\begin{aligned}
V_{tot} &= V_{\xi sr}\,(\hat{t}_\blacklozenge) \\
&= E_\xi\, E_s\, E_r\, \{(\hat{t}_\blacklozenge - t)^2\} \\
&= E_\xi\, E_s\, E_r \{(\hat{t} - t) + (\hat{t}_\blacklozenge - \hat{t})\}^2 \\
&= E_\xi\, V_s + E_s\, E_r\, V_{\xi c}
\end{aligned}
\tag{4.1}
$$

où $V_s = E_s\{(\hat{t} - t)\}^2$ est la variance de $\hat{t}$ selon le plan, en supposant que $\hat{t}$ est non biaisé selon le plan pour le total t. (Pour un estimateur légèrement biaisé selon le plan, V_s est l'erreur quadratique moyenne de $\hat{t}$ selon le plan.) On notera que $(\hat{t} - t)$ est fonction de s seulement, et non de r. En outre,

$$V_{\xi c} = E_\xi\{(\hat{t}_\blacklozenge - \hat{t})^2 \,|s,r\}$$

est la variance de l'erreur d'imputation selon le modèle, étant donné s et r. L'indice c signifie «conditionnel». Le calcul de l'équation (4.1) repose sur deux hypothèses: 1) il est possible de permuter l'espérance mathématique E_ξ à l'intérieur de $E_s\, E_r$; 2) le terme composite

$$2\, E_\xi\, E_s\, [(\hat{t} - t)\, E_r\, \{(\hat{t}_\blacklozenge - \hat{t})\,|s\}] \tag{4.2}$$

s'annule ou s'approche suffisamment de zéro pour qu'il soit possible de ne pas en tenir compte. Cette condition serait remplie, par exemple, si l'erreur d'imputation prévue était égale à zéro ou négligeable pour le mécanisme de réponse, étant donné l'échantillon probabiliste s obtenu. Même si le terme (4.2) n'est pas exactement égal à zéro pour le mécanisme de réponse réel, il est souvent possible d'utiliser zéro comme valeur approximative du terme (4.2) et d'utiliser la méthode ci-après pour obtenir une estimation de la variance: cette attitude est beaucoup plus valable que celle qui consiste à prétendre naïvement que les données imputées peuvent être considérées comme des données réellement observées.

Si nous posons $V_{éch} = E_\xi\, V_s$ et $V_{imp} = E_s\, E_r\, V_{\xi c}$ dans (4.1), alors

$$V_{tot} = V_{éch} + V_{imp}$$

ou

variance globale = variance d'échantillonnage + variance d'imputation.

Notre objectif est d'estimer la variance globale afin qu'il soit possible de calculer un intervalle de confiance valide pour le total t inconnu. Notre approche consiste à obtenir des estimations distinctes, $\hat{V}_{éch}$ et $\hat{V}_{imp}$, des deux composantes $V_{éch} = E_\xi\, V_s$ et $V_{imp} = E_s\, E_r\, V_{\xi c}$. Les données dont on dispose pour établir ces estimations sont $y_{\blacklozenge k}$, $k \in s$.

Le raisonnement suivi pour obtenir $\hat{V}_{éch}$ et $\hat{V}_{imp}$ est le suivant:

i) Estimation de la variance d'échantillonnage. Soit $\hat{V}_s$ l'estimateur (non biaisé ou presque non biaisé selon le plan) ordinaire de la variance V_s selon le plan. Nous notons $\hat{V}_{\blacklozenge s}$ la quantité obtenue en calculant $\hat{V}_s$ à partir des données après imputation, $y_{\blacklozenge k}$, $k \in s$. Or, pour de nombreuses règles d'imputation, $\hat{V}_{\blacklozenge s}$ sous-estime $V_{éch}$. Nous compensons cette sous-estimation en évaluant l'espérance mathématique

$$E_\xi (\hat{V}_s - \hat{V}_{\phi s}) = V_{dif}$$

et en trouvant un estimateur de V_{dif} non biaisé selon le modèle, noté $\hat{V}_{dif}$. À cette fin, il peut s'avérer nécessaire d'estimer certains paramètres du modèle ξ. En conséquence,

$$E_\xi (\hat{V}_{dif}) = E_\xi (\hat{V}_s - \hat{V}_{\phi s}).$$

Alors

$$\hat{V}_{éch} = \hat{V}_{\phi s} + \hat{V}_{dif}$$

est globalement non biaisé pour la composante $V_{éch} = E_\xi V_s$, comme le démontre le calcul suivant:

$$\begin{aligned} E_\xi E_s E_r (\hat{V}_{éch}) &= E_s E_r \{E_\xi (\hat{V}_{\phi s}) + E_\xi (\hat{V}_{dif})\} \\ &= E_s E_r \{E_\xi (\hat{V}_s)\} = E_\xi E_s (\hat{V}_s) \\ &= E_\xi V_s = V_{éch} \end{aligned}$$

ii) Estimation de la variance d'imputation. Notre approche consiste simplement à trouver un estimateur de $\bar{V}_{\xi c}$, noté $\hat{V}_{\xi c}$, qui est non biaisé selon le modèle, de sorte que $E_\xi (\hat{V}_{\xi c}) = V_{\xi c}$. Encore une fois, il est possible que nous ayons à estimer certains paramètres inconnus du modèle ξ à cette fin. Alors $\hat{V}_{\xi c}$ est un estimateur globalement sans biais de la variance d'imputation V_{imp}, puisque

$$E_s E_r E_\xi (\hat{V}_{\xi c}) = E_s E_r E_{\xi c} = V_{imp}.$$

Enfin, on obtient un estimateur globalement sans biais de V_{tot} à l'aide de

$$\hat{V}_{tot} = \hat{V}_{éch} + \hat{V}_{imp}$$

où $\hat{V}_{éch} = \hat{V}_{\phi s} + \hat{V}_{dif}$ et $\hat{V}_{imp} = \hat{V}_{\xi c}$. Notons que le terme $\hat{V}_{dif}$ a pour rôle de compenser la variation éventuellement «moins que naturelle» des données après imputation. On observe souvent une telle variation lorsque $y_{imp,k}$ est égale à la valeur prédite à l'aide d'un modèle de régression, c'est-à-dire «la valeur appartenant à la droite». En pareil cas, la valeur prédite ne reflète pas la dispersion des valeurs autour de la droite.

Cette méthode d'estimation de la variance a pour avantage de nécessiter uniquement la formulation d'hypothèses minimales sur la relation entre la propension à répondre et la variable étudiée y. Si le modèle d'imputation choisi se révèle approximativement valide, la méthode d'estimation fonctionne même en présence d'une relation systématique, par exemple, si les unités pour lesquelles les valeurs y_k sont élevées sont moins susceptibles de répondre.

EXEMPLE. Prélevons un échantillon s de n unités parmi N à l'aide d'un ÉASSR et dénotons le nombre de répondants par m. La valeur imputée pour les unités nécessitant une imputation est la moyenne pour les répondants. En d'autres termes, $y_{\phi k} = y_k$ si $k \in r$ et $y_{\phi k} = \bar{y}_r$ si $k \in s - r$, et $\hat{t}_\phi = (N/n) \sum_s y_{\phi k} = N \bar{y}_r$. Donc $\hat{V}_{\phi s} = N^2 (1/n - 1/N)\{(m - 1)/(n - 1)\} S_{yr}^2$, $\hat{V}_{dif} = N^2 (1/n - 1/N)\{(n - m)/(n - 1)\} S_{yr}^2$, et $\hat{V}_{imp} = N^2 (1/m - 1/n) S_{yr}^2$, où $S_{yr}^2 = \sum_r (y_k - \bar{y}_r)^2/(m - 1)$. En conséquence, $\hat{V}_{éch} = N^2 (1/n - 1/N) S_{yr}^2$ et $\hat{V}_{tot} = N^2 (1/m - 1/N) S_{yr}^2$. Le tableau suivant indique, pour différents taux d'imputation, la contribution des trois termes à $\hat{V}_{tot}$, en supposant que N est beaucoup plus grand que m et n, et que $(m - 1)/m \approx (n - 1)/n \approx 1$.

Taux d'imputation en %	contribution à $\hat{V}_{tot}$ en pourcentage		
100 (1 - m/n)	$\hat{V}_{\phi s}$	$\hat{V}_{dif}$	$\hat{V}_{imp}$
10	81	9	10
20	64	16	20
30	49	21	30

Le tableau illustre bien le danger qu'il y a à considérer les valeurs imputées comme des données réelles: lorsque le taux d'imputation est de 30%, l'estimation ordinaire de la variance $\hat{V}_{\phi s}$ obtenue pour cet exemple est inférieure à la moitié de la variance totale correctement estimée. Il est utile d'utiliser l'imputation par la moyenne pour les répondants à titre d'exemple, car elle permet d'obtenir des résultats particulièrement simples. Toutefois, en pratique, cette méthode d'imputation n'est d'ordinaire ni justifiée ni efficace. Elle n'est pas justifiée parce que le modèle qui la sous-tend n'est pas assez sophistiqué pour éviter que les estimations ponctuelles soient entachées d'une erreur systématiques; elle est d'ordinaire inefficace parce que les résidus e_k sont élevés.

5. APPLICATION À L'IMPUTATION PAR LE QUOTIENT

La méthode suppose qu'on connaît une valeur auxiliaire positive x_k pour chaque unité $k \in s$. Si $k \in s - r$, nous imputons la valeur $y_{imp,k} = \hat{B}\, x_k$ où $\hat{B} = (\Sigma_r\, y_k)/(\Sigma_r\, x_k)$. Les données après imputation sont

$$y_{\phi k} = \begin{cases} y_k & \text{si } k \in r \\ \hat{B}\, x_k & \text{si } k \in s - r \end{cases}.$$

Le modèle impliqué par l'imputation par le quotient est

$$y_k = \beta\, x_k + \epsilon_k \tag{5.1}$$

où les ϵ_k sont les erreurs non corrélées du modèle de sorte que

$$E_\xi(\epsilon_k) = 0, \quad V_\xi(\epsilon_k) = \sigma^2 x_k. \tag{5.2}$$

Supposons que l'échantillon s est prélevé au moyen d'un ÉASSR et dénotons les tailles respectives de s, r, et $s - r$ par n, m, et $n - m$. S'il n'était pas nécessaire d'imputer des données, l'estimateur de $t = \Sigma_U\, y_k$ serait $\hat{t} = N\, \bar{y}_s$. Si on utilise les données après imputation, nous obtenons

$$\hat{t}_\phi = (N/n)\, \Sigma_s\, y_{\phi k} = N\, \bar{x}_s\, \bar{y}_r / \bar{x}_r. \tag{5.3}$$

(La barre supérieure et l'indice s, r, ou $s - r$ indiquent qu'il s'agit d'une «moyenne simple»; ainsi, $\bar{y}_r = \Sigma_r\, y_k/m$, $\overline{x_{s-r}} = \Sigma_{s-r}\, x_k/(n - m)$, etc.) En utilisant les résultats de la section précédente, nous obtenons $V_{tot} = V_{éch} + V_{imp}$ avec $V_{éch} = E_\xi\{N^2\,(1/n - 1/N)\, S_{yU}^2\}$ et $V_{imp} = E_s\, E_r\{N^2\,(1/m - 1/n)\, C_1\, \sigma^2\}$, où $S_{yU}^2 = \Sigma_U\,(y_k - \bar{y}_U)^2/(N - 1)$ et $C_1 = \bar{x}_s\, \overline{x_{s-r}}/\bar{x}_r$ est une constante connue. Dans cette application, le terme composite (4.1) est égal à zéro. Notre méthode d'estimation de la variance nous permet d'obtenir $\hat{V}_{tot} = \hat{V}_{éch} + \hat{V}_{imp}$, où

$$\hat{V}_{éch} = N^2 \, (1/n - 1/N) \, \{S^2_{y\phi s} + C_0 \, \hat{\sigma}^2\} \tag{5.4}$$

$$\hat{V}_{imp} = N^2 (1/m - 1/n) \, C_1 \, \hat{\sigma}^2 \tag{5.5}$$

où $S^2_{y\phi s} = \Sigma_s \, (y_{\phi k} - \overline{y_{\phi s}})^2 / (n - 1)$ est la variance calculée pour les données après imputation. En outre, nous avons choisi d'estimer σ^2 à l'aide de la formule non biaisée selon le modèle

$$\hat{\sigma}^2 = \frac{1}{\bar{x}_r \, \{1 - \frac{1}{m} \, (cv_{xr})^2\}} \, \frac{\Sigma_r \, (y_k - \hat{B} \, x_k)^2}{m - 1}$$

où $cv_{xr} = S_{xr} / \bar{x}_r$ est le coefficient de variation de x dans l'ensemble des répondants r. On obtient la constante C_0 par

$$C_0 = \frac{1}{\sigma^2} \, E_\xi \, (S^2_{ys} - S^2_{y\phi s})$$

où

$$S^2_{ys} = \frac{1}{n - 1} \, \Sigma_s \, (y_k - \bar{y}_s)^2$$

est la variance de l'échantillon (inconnue) pour des données composées uniquement d'observations réelles. Après évaluation,

$$C_0 = \frac{1}{n - 1} \left\{ \Sigma_{s-r} \, x_k - \frac{\Sigma_{s-r} \, x_k^2}{\Sigma_r \, x_k} + \frac{1}{n} \, \frac{\Sigma_{s-r} \, x_k \, \Sigma_s \, x_k}{\Sigma_r \, x_k} \right\}.$$

Si m n'est pas trop petit, les approximations $\hat{\sigma}^2 \approx (\Sigma_r \, e_k^2) / (\Sigma_r \, x_k)$ avec $e_k = y_k - \hat{B} \, x_k$ et $C_0 \approx (1 - m/n) \, \bar{x}_{s-r}$ sont assez précises pour la plupart des applications.

Nous pouvons exprimer la composante de variance d'imputation comme suit:

$$\hat{V}_{imp} = N^2 \, (1/m - 1/n) \, A \, \bar{x}_s \, \hat{\sigma}^2$$

où $A = \bar{x}_{s-r} \, / \, \bar{x}_r$. La constante A reflète l'effet de sélection attribuable à la non-réponse. Si les unités de grande taille sont moins portées à répondre que les unités de petite taille, il peut arriver que A soit beaucoup plus grand que 1, et que, pour un échantillon s et un nombre m de répondants donnés, la composante $\hat{V}_{imp}$ ait tendance à être élevée par rapport à un cas où toutes les unités sont également susceptibles de répondre. Intuitivement parlant, il semble tout à fait naturel qu'il en soit ainsi.

Il convient de noter deux cas spéciaux. 1) Si tous les $x_k = 1$, la variance totale estimée devient simplement

$$\hat{V}_{tot} = \hat{V}_{éch} + \hat{V}_{imp} = N^2 \, (1/m - 1/N) \, S_{yr}^2$$

où S_{yr}^2 est la variance des m observations réelles y_k. Ce résultat correspond à la variance obtenue en utilisant un plan d'échantillonnage à deux phases où on a recours à un ÉASSR dans chaque phase. 2) S'il n'est pas nécessaire d'imputer des données, c'est-à-dire si $s = r$, $\hat{V}_{imp} = 0$ et

$$\hat{V}_{tot} = \hat{V}_{éch} + N^2 \, (1/n - 1/N) \, S_{ys}^2.$$

En d'autres termes, notre méthode nous permet alors d'obtenir l'estimateur bien connu de la variance pour un ÉASSR.

Nous avons réalisé une étude de Monte-Carlo portant sur 100 000 séries de réponses successives afin de confirmer la validité des résultats obtenus à l'aide de l'imputation par le quotient. À cette fin, vous avons généré une population finie de taille $N = 100$ conformément au modèle défini par les équations (5.1) et (5.2). Nous avons obtenu la série type de réponses r en prélevant au moyen d'un ÉASSR un échantillon s de taille $n = 30$, puis, étant donné s, en générant r à l'aide d'un mécanisme de réponse sous la forme d'épreuves de Bernoulli indépendantes, soit une épreuve pour chaque $k \in s$, la probabilité du résultat «réponse» étant définie par θ_k. Nous avons utilisé trois mécanismes de réponse distincts: selon le premier mécanisme, θ_k s'accroît en fonction de y_k de telle façon que $\theta_k = 1 - \exp(-a_1 y_k)$; selon le deuxième mécanisme, θ_k s'accroît à mesure que y_k diminue de telle façon que $\theta_k = \exp(-a_2 y_k)$; selon le troisième mécanisme, θ_k est constant et égal à 0.7. Nous avons fixé les constantes a_1 et a_2 utilisées dans les deux premiers mécanismes de réponse (que l'on peut décrire comme non aléatoires) de façon à obtenir une probabilité moyenne de réponse égale à 0.7. Aussi, la taille des séries de réponses r obtenues variait-elle autour d'une valeur moyenne de 21 pour chacun des trois mécanismes. Nous avons calculé pour chaque r l'estimation ponctuelle $\hat{t}_\phi$ définie par l'équation (5.3) ainsi que trois estimateurs de la variance, $\hat{V} = \hat{V}(\hat{t}_\phi)$, distincts. Ces estimateurs étaient: 1) l'estimateur de la variance *basé sur un modèle* $\hat{V}_{tot} = \hat{V}_{éch} + \hat{V}_{imp}$, qui est égal à la somme des équations (5.4) et (5.5); 2) l'estimateur de la variance pour un échantillonnage *à deux phases* $N^2 \, (1/n - 1/N) \, S_{yr}^2 + N^2 \, (1/m - 1/n) \, \Sigma_r \, e_k^2 / (m - 1)$, estimateur qui découle de la théorie de l'échantillonnage à deux phases ordinaire, en supposant qu'on prélève au moyen d'un ÉASSR un sous-échantillon de m répondants parmi les n unités constituant l'échantillon initial (Rao, 1990); enfin, 3) l'estimateur de la variance *non corrigé ordinaire* $N^2 \, (1/n - 1/N) \, S_{y\phi s}^2$ obtenu en assimilant les valeurs imputées aux valeurs réelles. Les résultats obtenus sont présentés dans le tableau suivant.

Estimateur $\hat{V}$	Biais relatif de $\hat{V}$ en %		
	Mécanisme n° 1	Mécanisme n° 2	Mécanisme n° 3
Basé sur un modèle	-0.20	-4.64	-3.99
Deux phases	9.95	-12.49	-1.11
Non corrigé ordinaire	-25.73	-37.90	-33.21

Nous avons calculé le biais relatif d'un estimateur $\hat{V}$ comme $\{\text{moyenne}(\hat{V}) - \text{var}(\hat{t}_\phi)\}/\text{var}(\hat{t}_\phi)$, où la moyenne$(\hat{V})$ est égale à la moyenne des 100 000 valeurs de $\hat{V}$ et var$(\hat{t}_\phi)$ est égale à la variance des 100 000 valeurs de $\hat{t}_\phi$. La simulation confirme le fait que la méthode basée sur un modèle permet d'obtenir une estimation presque non biaisée de la variance, quel que soit le mécanisme de réponse. Ceci n'a rien de surprenant puisque notre

population est conforme au modèle sur lequel est basée la simulation. L'estimateur pour l'échantillonnage à deux phases est efficace lorsqu'on utilise le mécanisme de réponse n° 3 («données manquantes au hasard»); autrement, il est biaisé. Enfin, le fait de traiter les valeurs imputées comme des données réelles se traduit par une grave sous-estimation de la variance réelle pour les trois mécanismes de réponse.

6. VALEURS IMPUTÉES AVEC RÉSIDU AJOUTÉ

Nous distinguons deux genres de valeurs imputées: 1) la valeur imputée $y_{imp,k}$ correspond uniquement à une valeur prédite, $y_{pred,k}$, comme lorsqu'on utilise la valeur appartenant à une droite ou à une surface de régression ajustée. Ainsi, selon la méthode d'imputation par le quotient que nous avons utilisée plus haut, $y_{imp,k} = y_{pred,k} = \hat{B}\, x_k$ où $\hat{B} = (\Sigma_r\, y_k)/(\Sigma_r\, x_k)$; 2) la valeur imputée $y_{imp,k}$ se compose d'une valeur prédite et d'un résidu, de sorte que $y_{imp,k} = y_{pred,k} + e_k^*$. Le terme résiduel, dont l'objet est de faire que les valeurs imputées correspondent plus étroitement aux observations réelles, peut être obtenu au moyen d'un échantillonnage des résidus $e_k = y_k - y_{pred,k}$ calculés pour les unités répondantes $k \in r$. On trouvera ci-après un schéma conçu à cette fin. Les ouvrages sur le sujet recommandent parfois d'utiliser ce genre d'imputation afin de préserver les distributions des valeurs imputées [voir, par exemple, Little (1988)]. Le processus d'imputation est alors plus difficile à mettre en oeuvre. Pour les fins du SGE (dont l'objectif principal est d'obtenir une estimation valide de la précision des estimations d'enquête), il n'est pas évident que les avantages obtenus justifient les efforts supplémentaires devant être déployés.

Précisons quand même un schéma d'imputation par une «valeur prédite augmentée d'un résidu» dans le cas où on prend comme point de départ le modèle d'imputation par le quotient. Nous calculons, pour une unité $k \in r$, $e_k = y_k - \hat{B}\, x_k$ où $\hat{B} = (\Sigma_r\, y_k)/(\Sigma_r\, x_k)$, puis $\tilde{e}_k = e_k/\sqrt{x_k}$. Nous obtenons ainsi une série de m «résidus normalisés» $\tilde{e}_k$. Nous calculons ensuite, pour une unité $k \in s - r$, $e_k^0 = \sqrt{x_k}\, \tilde{e}_k$, où $\tilde{e}_k$ est prélevé dans la série obtenue au moyen d'un ÉASSR et x_k appartient à l'unité pour laquelle la valeur doit être imputée. Les unités pour lesquelles la valeur de x est élevée ont alors tendance à obtenir des résidus e_k^0 plus élevés, comme le prévoit le modèle. Nous posons alors $e_k^* = e_k^0 - (\Sigma_{s-r}\, e_k^0)/(n - m)$. Pour $k \in s - r$, nous imputons $y_{imp,k} = \hat{B}\, x_k + e_k^*$, $k \in s - r$; pour $k \in r$, nous disposons des observations réelles, y_k. Comme la somme des e_k^* pour tous les k appartenant à $s - r$ est posée comme étant égale à zéro, on obtient l'estimateur ponctuel par $\hat{t}_\phi = (N/n)\, \Sigma_s\, y_{\phi k} = N\, \bar{x}_s\, \bar{y}_r / \bar{x}_r$, comme à la section 4, mais sa variance est différente. On peut démontrer que $E_\xi\, E_s\, E_r\, E_\phi\, (S^2_{y\phi s} - S^2_{ys}) = 0$, où E_ϕ indique l'espérance sous le processus de sélection aléatoire des résidus normalisés. En d'autres termes, la différence entre la variance calculée à partir des données après imputation, $S^2_{y\phi s}$, et la variance inconnue d'un échantillon entièrement composé d'observations réelles, S^2_{ys}, est à peu près égale à zéro en moyenne. Nous pouvons donc utiliser $\hat{V}_{éch} = N^2\, (1/n - 1/N)\, S^2_{y\phi s}$ comme estimateur approximativement globalement sans biais de la variance d'échantillonnage, sans qu'il soit besoin d'y ajouter un terme $\hat{V}_{dif}$. Cependant, il est toujours nécessaire de calculer un estimateur de la variance d'imputation $V_{imp} = N^2\, (1/m - 1/n)\, C_1\, \sigma^2$ et de l'ajouter à $\hat{V}_{éch}$.

7. CONCLUSION

Les travaux en cours sur les techniques d'estimation de la variance dont il est fait état dans le présent article ont pour objectifs: 1) d'étudier la robustesse des estimateurs de la variance exposés à la section 4, c'est-à-dire leur sensibilité à l'infirmation des hypothèses du modèle d'imputation; 2) d'appliquer ces techniques à d'autres procédures d'imputation axées sur des modèles strictement implicites, en particulier à la méthode dite du plus proche voisin; 3) d'appliquer ces techniques au cas où plusieurs procédures d'imputation distinctes sont utilisées dans la même enquête; 4) d'appliquer ces techniques à d'autres types d'estimateurs que ceux considérés dans

l'exemple donné dans cet article. Nous exposerons dans une publication en cours de préparation les résultats obtenus lorsqu'on utilise l'estimateur de Horwitz-Thompson, $\hat{t} = \Sigma_s\, y_k/\pi_k$, comme prototype. L'estimateur obtenu à partir des données après imputation est alors défini par:

$$\hat{t}_\phi = \Sigma_r\, y_k/\pi_k + (\Sigma_{s-r}\, x_k/\pi_k)'\, \hat{B}$$
$$= \Sigma_s\, y_k/\pi_k - \Sigma_{s-r}\, e_k/\pi_k$$

où $e_k = y_k - x_k'\, \hat{B}$ est le résidu d'imputation pour l'unité k.

BIBLIOGRAPHIE

Herzog, T.N., et Rubin, D.B. (1983). Using Multiple Imputations to Handle Nonresponse in Surveys. Dans W.G. Madow, I. Olkin et D.B. Rubin (éds.), *Incomplete Data in Sample Surveys*. New York: Academic Press, 209-245.

Lavallée, P., Leblond, Y., et Reinhardt, G. (1990). Generalized Estimation systems, statement of requirements. Rapport, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

Little, R.J.A. (1988). Missing-Data Adjustments in Large Surveys (with discussion). *Journal of Business and Economic Statistics*, 6, 287-301.

Pritzker, L., Ogus, J., et Hansen, M.H. (1965). Computer editing methods: some applications and results. *Bulletin of the International Statistical Institute*, 41, 1, 442-466.

Rao, J.N.K. (1990). Variance Estimation Under Imputation for Missing Data. Manuscrit consulté par courtoisie de l'auteur.

Rubin, D.B. (1986). Initiation à l'imputation multiple pour les cas de non-réponse. *Techniques d'enquête* , 12, 41-52.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley & Sons.

Särndal, C.E. (1990). Estimation of Precision in the Generalized Estimation System when Imputation is Used. Rapport, Secteur de l'informatique et de la méthodologie, Statistique Canada.

ALLOCUTION DE CLÔTURE

Recueil du Symposium 90 de Statistique Canada
Mesure et amélioration de la qualité des données
Octobre 1990

ALLOCUTION DE CLÔTURE

G.J. Brackstone

Voilà qui nous amène à la fin du symposium de 1990. Nous avons traité un large éventail de questions, depuis la communication dynamique que John Early nous a présenté le premier matin et qui a donné le ton et déterminé le contexte du symposium jusqu'à l'excellente étude de Carl Särndal concernant l'incidence de l'imputation sur la variance et l'estimation de la variance et la discussion qui a suivi.

Nous avons exploré de nombreux aspects de la qualité des données, certains d'entre eux correspondant à des domaines traditionnels pour les statisticiens d'enquête et d'autres, à des questions nouvelles. Nous nous sommes penchés sur le plan de sondage global dans le cadre duquel nous tentons d'établir un équilibre entre le coût et la qualité pour toutes les opérations d'une enquête, et nous avons vu les moyens d'améliorer l'efficacité et la productivité des opérations d'enquête individuelles. Nous avons abordé l'éternelle question de la couverture du recensement et avons exploré des éléments nouveaux dans le domaine de l'estimation de la variance. Les problèmes particuliers de qualité reliés aux dossiers administratifs ont été traités tout comme les questions relatives à la qualité qui sont propres au maintien des bases de sondage, lesquelles reposent souvent sur des sources administratives. On nous a parlé du défi que présente l'intégration des données de différentes sources et de qualité variable et, bien entendu, on a traité des programmes d'assurance de la qualité.

Où cela nous mène-t-il? Nous pourrions nous sentir déprimés à l'idée des nombreux problèmes, certains très complexes, qui affectent la qualité et que nous devons résoudre, ou stimulés et prêts à explorer ce que nous avons appris au cours des trois derniers jours en vue de l'appliquer à nos problèmes. J'espère que c'est la deuxième hypothèse qui se réalisera. Aux participants de Statistique Canada, il va sans dire que nous faisons bon accueil aux propositions qui résultent de ce symposium et qui présentent des mesures nécessaires pour améliorer la qualité de nos programmes à un coût raisonnable ou qui nous permettent de maintenir notre niveau qualitatif à un moindre coût.

On nous a beaucoup parlé de la qualité totale, de la gestion de la qualité totale et des programmes de qualité totale. Je crois cependant qu'ici, le terme "total" revêt au moins trois sens différents. Premièrement, il y a le fait que le concept de la qualité s'étend au-delà de la simple variance ou même de l'erreur quadratique moyenne. Il englobe l'actualité, la pertinence et même le niveau de service en ce qui a trait à la diffusion des données. Dans ce sens, nous devons considérer la qualité totale. Deuxièmement, il y a l'idée que le concepteur d'une enquête doit tenir compte de toutes les diverses étapes de l'enquête afin de réaliser le meilleur rapport coût-qualité. Dans cette optique, la notion de qualité se rapporte habituellement à la variance ou à l'erreur quadratique moyenne: la qualité est "totale" en ce sens qu'elle englobe la part d'erreurs de toutes les étapes de l'enquête. Et troisièmement, le terme "total" peut avoir trait à la portée d'un programme sur la qualité en ce sens qu'il faut rechercher la qualité pour tous les programmes et produits d'un organisme. Toutes ces idées sont importantes, mais différentes les unes des autres.

Il importe de maintenir une vue large et équilibrée de la qualité. Il est très approprié de chercher à améliorer la qualité à partir de notions telles que "l'aptitude à l'usage" et "les besoins de l'utilisateur". Lors d'une réunion récente avec des utilisateurs, lorsque nous avons demandé à ces derniers de nous préciser les améliorations qu'ils souhaitaient le plus que nous apportions à nos produits, ils ont immédiatement répondu qu'ils aimeraient disposer de ces produits plus rapidement. Ils n'ont pas mentionné l'erreur quadratique moyenne. Nous devons appliquer les quelques ressources dont nous disposons pour l'amélioration de la qualité aux domaines qui présentent un potentiel réel d'amélioration d'aspects qualitatifs pertinents que les utilisateurs remarqueront et apprécieront.

Je vais maintenant ajouter quelques mots au sujet du symposium même. Il y a eu près de 350 inscriptions. Nous avons été particulièrement heureux de souhaiter la bienvenue aux participants d'Australie, de Finlande, de

Macao, de Suède et des Etats-Unis et aux représentants de nombreuses universités, entreprises et administrations publiques canadiennes. J'aimerais remercier Normand Laniel, Robert Lussier, Mary March et Jeff Smith du comité d'organisation qui ont planifié le programme du symposium et ont veillé a son déroulement au cours des derniers jours. Ils ont fait de l'excellent travail. A cette fin, ils ont reçu l'aide précieuse de Rosie Arena-Ryan, Suzanne Bonnell, Ann Brown, Suzanne Johnston, Christine Larabie, Carole Morin et Benita Therriault de Statistique Canada et de Gillian Murray de l'université Carleton. J'aimerais aussi témoigner ma reconnaissance à Don Royce qui a organisé la session sur l'évaluation de la couverture dans les recensements de la population. Enfin, je tiens à remercier tous les orateurs et les présidents de comités qui ont contribué au succès de ce symposium.

Nous avons commencé à planifier le symposium de l'an prochain. Le thème portera sur les questions spatiales dans le domaine de la statistique. Nous voulons concentrer notre attention sur des sujets qui combinent les théories et l'application de la méthodologie d'enquête et de la géographie. Nous nous attendons à voir des questions telles que l'utilisation de méthodes, d'outils et de techniques géographiques dans les plans de sondage et la collecte des données, le traitement de données codées suivant une grille géographique, l'intégration et l'analyse des données spatiales, les questions relatives à l'affichage et à la diffusion des données géographiques, et les travaux de recherche connexes.

Le symposium de 1990 est terminé. Je vous remercie tous de votre participation et de votre appui.