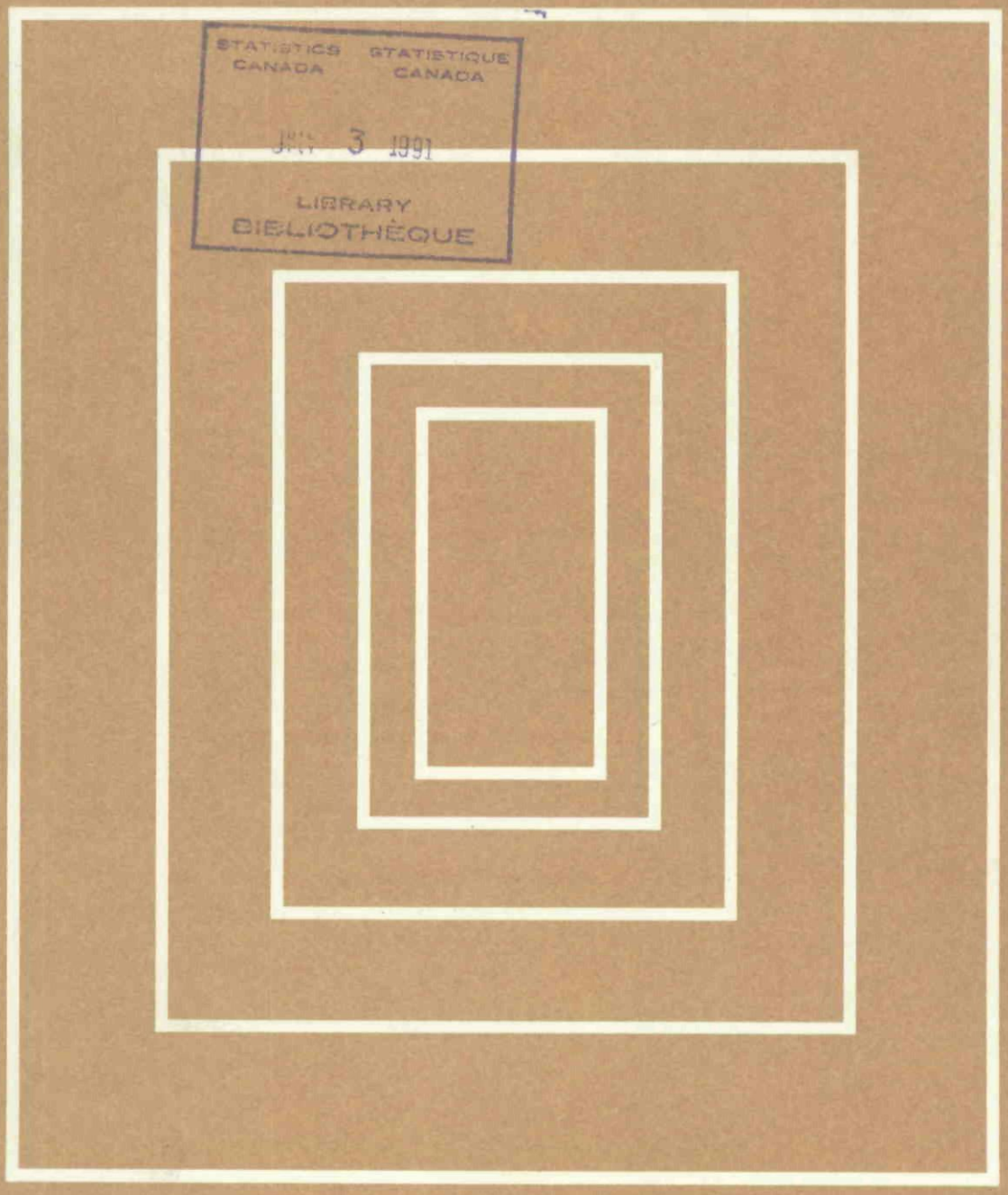
C.3

Catalogue 12-001

# Techniques d'enquête

Une revue de Statistique Canada

Juin 1991 Volume 17 Numéro 1



STATISTICS CANADA STATISTIQUE CANADA
JUL 3 1991
LIBRARY BIBLIOTHÈQUE

Statistique Canada Statistics Canada

Canada

Statistique Canada
Division des méthodes d'enquêtes sociales

# Techniques d'enquête

Une revue de Statistique Canada

Juin 1991 Volume 17 Numéro 1

Publication autorisée par le ministre de
l'Industrie, des Sciences et de la Technologie

© Ministre des Approvisionnements
et Services Canada 1991

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrement sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable du ministre des Approvisionnements et Services Canada.

Juin 1991

Canada : 35 $
États-Unis : 42 $ US
Autres pays : 49 $ US

Catalogue 12-001

ISSN 0714-0045

Ottawa

10100697 88

Ca OOS E

Statistics Canada Library
Bibliothèque Statistique Canada

# TECHNIQUES D'ENQUÊTE

## Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

**Président** G.J. Brackstone

**Membres** B.N. Chinnappa, G.J.C. Hole, C. Patrick, F. Mayda (Directeur de la production), R. Platek, D. Roy, M.P. Singh

COMITÉ DE RÉDACTION

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

**Rédacteurs associés**

B. Afonja, *Nations Unies*
D.R. Bellhouse, *U. of Western Ontario*
D. Binder, *Statistique Canada*
E.B. Dagum, *Statistique Canada*
J.-C. Deville, *INSEE*
D. Drew, *Statistique Canada*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistique Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*
J.N.K. Rao, *Carleton University*
D.B. Rubin, *Harvard University*
I. Sande, *Bell Communications Research, U.S.A.*
C.E. Särndal, *Université de Montréal*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *U.S. Internal Revenue Service*
C.M. Suchindran, *University of North Carolina*
J. Waksberg, *Westat Inc.*
K.M. Wolter, *A.C. Nielsen, U.S.A.*

**Rédacteurs adjoints**

J. Gambino, L. Mach et A. Théberge, *Statistique Canada*

---

POLITIQUE DE RÉDACTION

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

Présentation de textes pour la revue

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

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 35 $ par année au Canada, 42 $ (É.-U.) aux États-Unis, et de 49 $ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

# TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 17, numéro 1, juin 1991

## TABLE DES MATIÈRES

Dans ce numéro ..... 1

**Nouvelles approches pour la collecte et la saisie des données**

G.S. WERKING et R.L. CLAYTON
Amélioration de la qualité des données à l'aide d'un mode de collecte mixte ..... 3

P.A. PHIPPS et A.R. TUPEK
Fiabilité des données introduites au moyen d'un téléphone à clavier ..... 17

J.M. BRICK et J. WAKSBERG
Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire ..... 31

J.G. BETHLEHEM et W.J. KELLER
Le système Blaise ou comment élaborer un système de traitement intégré des données d'enquête ..... 47

J.D. DREW
Recherche et essais pour les méthodes d'enquêtes par téléphone à Statistique Canada ..... 63

---

D.R. BELLHOUSE
Fonctions de vraisemblance marginales et fonctions de vraisemblance conditionnelles approximatives pour l'échantillonnage répété ..... 77

I. SCHIOPU-KRATINA et K.P. SRINATH
Renouvellement de l'échantillon et estimation dans l'Enquête sur l'emploi, la rémunération et les heures de travail ..... 89

P.S. KOTT
Estimation d'un système d'équations linéaires à l'aide de données d'enquête ..... 101

L. BIGGERI et U. TRIVELLATO
L'évaluation des erreurs dans les données de comptabilité nationale: les estimations provisoires et révisées ..... 111

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 1–2
Statistique Canada



# Dans ce numéro

La réalisation d'enquêtes téléphoniques par les méthodes classiques est relativement peu coûteuse et assure un contact direct, bien qu'à distance, entre l'intervieweur et le répondant. Ce sont surtout ces deux caractéristiques – coût peu élevé et contact humain – qui ont fait de l'enquête téléphonique un outil de sondage privilégié. Les nombreuses recherches qui ont été faites à ce sujet durant la dernière décennie sont résumées dans *Telephone Survey Methodology*, colligé par R.M. Groves et coll. (1988), et dans le *Journal of Official Statistics*, vol. 4, n° 4. La section spéciale de ce numéro, consacrée à la collecte et à la saisie des données, fait état d'autres recherches où le téléphone joue encore un rôle important dans un contexte technologique plus nouveau.

Dans le premier article de la section spéciale, Werking et Clayton commentent les recherches qu'a réalisées depuis sept ans le U.S. Bureau of Labor Statistics (BLS) sur les méthodes de collecte de données par téléphone. Ils montrent comment ces recherches ont conduit à la décision de mettre en oeuvre l'interview téléphonique assistée par ordinateur (ITAO) et la saisie de données au moyen d'un téléphone à clavier (TDE – touchtone data entry) dans la Current Employment Statistics Survey en 1991 et 1992. Les auteurs soulignent aussi une nouvelle technique, la reconnaissance de la parole, qui peut remplacer la saisie par boutons-poussoirs.

Phipps et Tupek examinent un des aspects des enquêtes téléphoniques sur lesquels ont porté les recherches du BLS, notamment les erreurs de mesure dans les enquêtes avec saisie au moyen d'un téléphone à clavier. Ils concluent que la plus grande source d'erreurs est l'introduction de chiffres en trop ou de chiffres erronés mais que cela peut être corrigé partiellement grâce au contrôle des données, surtout lorsqu'il existe des données longitudinales. Ils proposent aussi des moyens d'améliorer les systèmes de saisie au moyen d'un téléphone à clavier.

La composition aléatoire (CA) de Mitofsky-Waksberg est une méthode courante de sélection des ménages. Brick et Waksberg examinent une caractéristique opérationnelle non efficace de cette méthode et proposent une version modifiée dont ils étudient les propriétés statistiques. De plus, ils indiquent des critères permettant de déterminer dans quelles circonstances on doit opter pour la méthode originale ou ses variantes.

Bethlehem et Keller présentent le système Blaise, un progiciel innovateur élaboré au Bureau de la statistique des Pays-Bas. Les auteurs montrent comment Blaise sert à intégrer les diverses étapes du traitement des données d'enquête, y compris la collecte et la saisie. Avec l'utilisation croissante des ordinateurs portatifs dans les enquêtes par sondage, des outils comme le système Blaise deviendront indispensables.

Comme la plupart des organismes de statistiques à l'étranger, Statistique Canada étudie la possibilité d'étendre et de perfectionner l'utilisation du téléphone dans les enquêtes par sondage. Dans le dernier article de la section spéciale, Drew résume les recherches et les expériences récentes qui ont été faites en ce qui concerne les enquêtes-ménages. Il expose aussi les conséquences de ces recherches et de ces expériences pour le remaniement de l'Enquête sur la population active.

On peut utiliser un modèle de régression linéaire normal pour l'échantillonnage répété. Bellhouse obtient les fonctions de vraisemblance marginales et conditionnelles pour la matrice des coefficients de corrélation de ce modèle. Il présente des applications de ces fonctions pour l'échantillonnage aléatoire simple et pour des plans plus complexes.

Schiopu-Kratina et Srinath décrivent la méthodologie de l'Enquête sur l'emploi, la rémunération et les heures de travail (EERH) de Statistique Canada. L'EERH est une grande enquête mensuelle menée auprès des entreprises et qui repose sur un échantillon avec renouvellement. La détermination de la taille de l'échantillon à chaque mois est une opération assez complexe vu l'évolution constante de la population. Les auteurs proposent des façons de simplifier le plan de sondage.

Kott montre comment certaines techniques élaborées en vertu de la théorie des sondages fondée sur les plans, notamment l'inclusion des poids d'échantillonnage et l'estimateur de l'erreur quadratique moyenne fondé sur la linéarisation, peuvent servir à l'estimation d'un système d'équations linéaires. Il montre aussi que l'utilisation de poids d'échantillonnage peut être souhaitable en l'absence probable de variables explicatives. De plus, l'estimateur de l'erreur quadratique moyenne fondé sur la linéarisation est quasi-non biaisé pour de nombreuses structures d'erreur.

La nécessité de publier le plus rapidement possible et, en même temps, de pouvoir établir des estimations justes à partir de toutes les données disponibles amène l'élaboration d'un processus de révision des estimations des comptes nationaux. Biggeri et Trivellato examinent ce qui s'est fait récemment au chapitre de l'analyse de la fiabilité des estimations des comptes nationaux révisées ultérieurement. Ils présentent aussi une étude empirique où ils utilisent des données du Canada, de l'Italie et des États-Unis.

Le rédacteur en chef

Au moment de mettre sous presse, nous apprenons le décès de M. M.N. Murthy, Directeur de l'Applied Statistics Centre, Madras, Inde, le 2 avril, 1991. M. Murthy a fait des contributions importantes à la méthodologie d'enquête. Il était l'auteur du livre très reconnu *Sampling Theory and Methods*. Pendant sa carrière, il a travaillé pour l'Indian Statistical Institute et l'Institut Statistique de l'ONU pour l'Asie et le Pacifique. Il était Fellow de l'ASA, la Royal Statistical Society et l'International Statistical Institute. Nous étions comblés de l'avoir comme membre du comité de rédaction. Il manquera à ses collègues et ses étudiants de par le monde. Il laisse dans le deuil sa femme, Vyjayanthi, et sa fille Shashi.

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 3–16
Statistique Canada



# Amélioration de la qualité des données à l'aide d'un mode de collecte mixte

GEORGE S. WERKING et RICHARD L. CLAYTON[1]

RÉSUMÉ

L'établissement d'estimations assujetties à des contraintes de temps pose un problème constant, qui est celui des limitations importantes qui se rattachent à la collecte des données par la poste. Pour surmonter cette difficulté, le Bureau of Labor Statistics des États-Unis, a effectué pendant 7 ans d'importants travaux de recherche sur l'utilisation de l'interview téléphonique assistée par ordinateur (ITAO) et de l'auto-interview assistée par ordinateur (AIAO), dans le cadre de cette dernière méthode les répondants introduisant eux-mêmes leurs données au moyen d'un téléphone à clavier ou par reconaissance de la parole. Le présent article 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; ITAO; collecte des données au moyen d'un téléphone à clavier; 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 des États-Unis 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 il 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, sur les gains et sur les heures par industrie, au niveau détaillé, recueillies au moyen de l'enquête mensuelle réalisée par le Bureau auprès de 350,000 entreprises – la Current Employment Statistics Survey (CESS) – ainsi que de données sur la population active et sur 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 toutes seules comme indicateurs économiques principaux, mais elles servent également à l'élaboration de nombreux autres indicateurs importants de l'activité économique du pays, y compris: le revenu personnel pour le calcul du produit national brut, l'indice des indicateurs économiques avancés, l'indice des indicateurs instantanés d'activité, l'indice de l'activité 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,

[1] G. Werking et R. Clayton, Monthly Industry Employment Statistics Division, Bureau of Labor Statistics, Bureau 2089, 441 G Street, N.W., Washington, D.C., É.-U., 20212.

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" basées seulement sur l'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 initiales. 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. Pour surmonter cette difficulté, le Bureau a mis en oeuvre un programme de recherche sur les techniques de collecte téléphonique automatisée des données dans le but de réduire considérablement l'ampleur et la fréquence des grosses révisions apportées aux estimations provisoires.

Le présent article donne un aperçu des travaux de recherche effectués par le Bureau pendant 7 ans sur les techniques de collecte téléphonique automatisée des données et résume certains des principaux résultats de ces travaux. Vous trouverez dans les sections ci-après 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 téléphone à clavier (IDTC) 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 gouvernement fédéral et les gouvernements des États aux termes duquel il revient au Bureau de préciser le plan de sondage et les procédures opérationnelles de l'enquête, tandis qu'il incombe à chaque État de mettre en oeuvre les activités de collecte des données ainsi que les activités de conciliation. Le Bureau produit et publie, pour l'ensemble du pays, des estimations mensuelles complètes pour toutes les industries au niveau des catégories à 2, 3 ou 4 chiffres, tandis que chaque État produit 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 complets (ou pour l'ensemble de l'univers) pour l'année précédente et on utilise ces chiffres pour procéder à un étalonnage (réalignement) annuel des estimations produites à partie de l'échantillon de la CESS en fonction de ces chiffres obtenus pour l'univers. Cet étalonnage annuel permet d'obtenir des estimations plus précises pour le mois en cours ainsi qu'une estimation annuelle 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 0.2%, et il a été pratiquement nul pendant 4 ans au cours des années quatre-vingt. Alors que l'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 quand on les a 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 la méthode 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'aptitude à l'usage des données ont considérablement augmenté. L'émergence de cette nouvelle façon de voir est dans une large mesure attribuable aux travaux de Deming, de Juran et d'autres auteurs portant sur la gestion de la qualité. Pendant les années quatre-vingt on s'est également beaucoup plus concentré sur l'importance et sur les utilisations 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, cet accroissement de l'utilisation et ce gain de visibilité des statistiques de la CESS se sont accompagnés 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, dans le domaine de la micro-informatique. Cette nouvelle technologie a offert aux organismes statistiques de nombreuses occasions nouvelles d'améliorer le contrôle et la qualité de la collecte des données, notamment: l'interview téléphonique assistée par ordinateur, l'introduction des données au moyen d'un téléphone à clavier, la reconnaissance de la parole, l'auto-interview assistée par ordinateur 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 des programmes permanents 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 alors disponibles de collecte automatisée des données, le Bureau prévoit procéder à une application à grande échelle de ces techniques en 1991.

## 2. PROGRAMME DE RECHERCHE DANS LE CADRE DE 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 méthodes de collecte susceptibles de permettre un accroissement important des taux de réponse en vue d'améliorer 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 pour la 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 ces méthodes de collecte des données à un niveau à peu près équivalent au coût des actuelles méthodes de collecte par la poste?

Le programme de recherche a permis de mettre au point une méthode de collecte mixte par ITAO/IDTC assurant l'obtention du taux de réponse désiré et le respect des contraintes financières imposées. On trouve dans les sections ci-après une brève description de ces méthodes de collecte assistée par micro-ordinateur et des essais réalisés au cours des travaux de recherche ainsi qu'un exposé des taux de réponse obtenus et une analyse des coûts. Les mémoires de recherche dont il est fait état dans la bibliographie renferment de plus amples renseignements sur ces essais. De plus, des résultats récents sur l'erreur de mesure dans le cas de la collecte effectuée à l'aide de téléphones à clavier sont donnés dans un article de Phipps et Tupek, dans la présente publication.

### 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 le point de référence de la CESS est la période de paie englobant le 12$^{e}$ jour du mois, on ne dispose que de 2 1/2 semaines pour recueillir, introduire, contrôler, totaliser, valider et publier les données. Afin de respecter cet échéancier serré, il faut disposer d'une méthode de collecte permettant d'obtenir les données requises dès que l'entreprise en dispose. Les quatre méthodes de collecte des données étudiées sont décrites à tour de rôle ci-après.

**Poste** – Le questionnaire de la CESS est un questionnaire-navette d'une page, expédié par la poste, sur lequel il y a suffisamment d'espace afin que l'employeur puisse inscrire les données pour 12 mois. Après avoir reçu, chaque mois, le questionnaire par la poste le ou vers le 12$^{e}$ 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 d'employés qui ne sont pas des cadres), heures et gains. Une fois le questionnaire rempli, l'employeur le renvoie par la poste à l'organisme d'État, où les données qu'il renferme sont saisies puis contrôlées. Le questionnaire est ensuite classé afin qu'il puisse être réexpédié par la poste à l'employeur pour recueillir les données du rapport du mois suivant. Comme nous l'avons déjà mentionné, ce processus permet actuellement d'obtenir un taux de réponse de 50% au cours des 2 1/2 semaines dont on dispose avant la diffusion des estimations provisoires.

**Interview téléphonique assistée par ordinateur (ITAO)** – Suivant la méthode de l'ITAO, on envoie le questionnaire de la CESS à l'employeur par la poste une seule fois au début de l'année et ce dernier le conserve pour y inscrire les données mensuelles pendant toute l'année. 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, dont la date et l'heure ont déjà été fixées, de l'organisme d'État chargé de réaliser l'ITAO. 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 d'un téléphone à clavier (IDTC)** – Suivant la méthode de l'IDTC, l'employeur suit les mêmes étapes que pour la méthode de l'ITAO, mais, plutôt que d'attendre l'appel de l'organisme d'État chargé de réaliser l'ITAO, il compose lui-même un numéro 800 qui permet de communiquer avec le micro-ordinateur à entrée par téléphone à clavier de cet organisme, puis il se sert du clavier de son téléphone pour introduire les éléments d'information à la suite des messages de guidage appropriés prévus dans l'interview automatisée pour la CESS. Après avoir été introduit, chaque élément d'information est relu pour permettre au répondant de le vérifier.

**Reconnaissance de la parole (RP)** – La déclaration des données par la méthode de reconnaissance de la parole et la collecte des données au moyen d'un téléphone à clavier s'effectuent de la même façon, à la différence près que la première méthode ne nécessite pas l'utilisation d'un téléphone à clavier. Il suffit en effet à l'employeur de lire les données inscrites sur la formule pour que le micro-ordinateur à entrée vocale les traduise et les lui relise pour lui permettre de les vérifier. Le système de RP 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 d'ITAO articulé sur micro-ordinateur destiné à être utilisé dans le cadre d'un essai portant sur deux États, dont la mise en oeuvre a commencé en 1984 (figure 1). C'est le système d'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 sélectionné un échantillon aléatoire de 200 unités dans chaque État, le Bureau a graduellement amélioré les procédures et les systèmes de collecte au cours des 7 années suivantes. Les premiers 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 pour la diffusion des estimations provisoires). Ainsi, le critère choisi pour évaluer l'efficacité des nouvelles méthodes de collecte par ITAO et par IDTC a été leur capacité de faire passer à un taux permanent stable de 80 à 90% le taux de réponse d'échantillons d'unités ayant affiché un taux de réponse de 0 à 20% à la date limite pour la diffusion des estimations provisoires. En 1990, à la conclusion des travaux de recherche sur l'ITAO, le Bureau utilisait cette méthode pour recueillir des données auprès de plus 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, en 1985, il était également devenu évident que la collecte permanente par ITAO serait plus coûteuse que la méthode existante de collecte par la poste. À ce moment, de nouveaux travaux ont été entrepris pour chercher à réduire le coût de la méthode de l'ITAO tout en maintenant le taux de réponse mensuel au niveau élevé que cette méthode permettait d'atteindre. 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 micro-ordinateur 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, une version d'un système de vérification par introduction des données au moyen d'un téléphone à clavier. Après avoir déterminé un système de déclaration au moyen d'un téléphone à clavier articulé sur micro-ordinateur se prêtant à la réalisation d'essais dans le cadre d'une enquête, le Bureau a commencé à effectuer, en 1986, un essai de collecte de données à l'aide de cette technique dans deux États. Il convient de noter que l'IDTC 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 fournir leur réponse dans les délais prévus, tandis que l'IDTC devait être utilisée auprès des entreprises répondant par ITAO 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'IDTC 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'IDTC en arrivent également à leur conclusion, plus de 5,000 unités réparties entre 14 États continuent de déclarer chaque mois leurs données au moyen de cette technique et le Bureau a recueilli les données pour un total de plus de 100,000 questionnaires à l'aide de cette nouvelle méthode de déclaration automatisée.

Dans la foulée des travaux sur l'IDTC, le Bureau effectue actuellement plusieurs essais à petite échelle 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'IDTC, mais que la déclaration par RP a l'avantage important de paraître plus naturelle aux répondants et que ces derniers la préfèrent généralement à l'IDTC. Actuellement, le coût d'acquisition du matériel de RP est environ 15 fois celui du matériel d'IDTC; toutefois, d'ici quelques années, à mesure que ce coût diminuera, la déclaration par RP devrait constituer une solution de rechange rentable pour l'IDTC.

### 2.4 Résultats des travaux de recherche

Au cours des 7 dernières années, le Bureau a été en mesure de déterminer que des données sur la paie sont disponibles dans la plupart des entreprises avant la date limite pour la 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 avec un taux de réponse de 82 à 84% (figure 1). Les taux de réponse ainsi obtenus sont demeurés remarquablement stables au fil des ans, au fur et à mesure que la taille de l'échantillon utilisé pour l'ITAO s'élargissait, pour passer de 400 à 5,000 unités, et que le nombre d'États participants s'accroissait de 2 à 14. Selon les résultats de la recherche, la majorité des entreprises disposent des données à temps pour respecter la date limite pour la diffusion, et l'utilisation de la méthode de l'ITAO permet d'augmenter dans une mesure de 60 à 80% le taux de réponse des entreprises donnant une réponse tardive par la poste, 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 pour la diffusion était la longueur de la période de paie de l'entreprise (figure 3). Il y a généralement une période de paye par semaine, aux deux semaines, deux fois par mois, ou chaque 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 aux deux semaines sont la plupart du temps disponibles dans les délais fixés; toutefois, la majorité des périodes de paie mensuelles se terminent longtemps après la date limite pour la 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 de l'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 pour les autres 40% cette première communication téléphonique sert à rappeler à ces personnes qu'elles doivent communiquer leurs données (figure 1). 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.

| | | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 |
|---|---|---|---|---|---|---|---|---|
| Poste | Taux de réponse | 47% | 47% | 48% | 49% | 49% | 51% | 52% |
| ITAO | Unités | 400 | 400 | 2000 | 3000 | 5000 | 5000 | 5000 |
| | 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 |
| IDTC 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 1.** Sommaire des résultats de la recherche

La durée moyenne d'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 d'ITAO (figure 1) suite à la rationalisation du matériel utilisé et à l'accroissement de l'expérience des intervieweurs. Une autre question très importante prise en considération au cours des essais a été l'incidence de la méthode de l'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 de l'ITAO qu'avec la méthode de collecte par la poste, et qu'avec la première méthode il n'y avait presque pas de répondants d'importance qui sortaient de l'échantillon. Bref, il semble que l'ITAO a reçu un accueil favorable auprès de la majeure partie des répondants et a presque permis d'optimaliser le taux de réponse pouvant être obtenu pour les estimations provisoires.

Par suite de l'accroissement des coûts reliés à l'utilisation de la méthode de l'ITAO, le Bureau a entrepris des travaux de recherche sur la collecte des données au moyen d'un téléphone à clavier. 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 de l'ITAO à afficher des taux de réponse aussi élevés en ayant recours à une méthode de déclaration par IDTC complètement automatisée (figure 1). L'importance de ce résultat s'explique par les économies que l'IDTC permet de réaliser par rapport à l'ITAO. Un des principaux sujets de préoccupation relatifs à la collecte par IDTC était que, contrairement à ce qui se produit pour la collecte par ITAO, où les heures fixées pour les appels s'échelonnent sur toute la journée, il se pourrait que les répondants par IDTC 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 de micro-ordinateurs à entrée par téléphone à clavier pour traiter les appels reçus en période de pointe. Heureusement, ces inquiétudes se sont avérées sans fondement et, bien que les micro-ordinateurs à entrée par téléphone 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 auxquels il faut téléphoner pour leur rappeler qu'ils doivent communiquer leurs données a tendance à s'établir au même niveau (environ 40%), qu'on utilise la méthode de l'IDTC ou la méthode de l'ITAO. On met actuellement à l'essai des méthodes visant à réduire le nombre d'appels servant à rappeler aux répondants qu'ils doivent communiquer leurs données nécessités par la méthode de l'IDTC. Un des principaux avantages que présente cette méthode pour le répondant est q'une interview par IDTC ne prend que la moitié du temps d'une interview réalisée par ITAO, la durée moyenne d'une interview par IDTC s'établissant à seulement 1 minute 45 secondes. En outre, la plupart des entreprises sont déjà pourvues de téléphones 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 de l'IDTC. Bien que cette méthode 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 pour les répondants font que ces derniers réagissent très positivement à cette méthode.

Mentionnons une observation générale relative à l'élaboration d'un programme de recherche sur l'ITAO, soit le fait que la nature du matériel ou du logiciel utilisé au cours des travaux de recherche n'est pas critique, pour autant que ce matériel ou ce logiciel soit assez souple pour être modifié. Il est possible que les résultats finals des essais indiquent que la mise en oeuvre du système d'ITAO comporte 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'amélioration 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 de l'ITAO que celle de l'IDTC 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 de transition afférents à la méthode de l'ITAO et des coûts courants de mise en oeuvre de la méthode de l'IDTC.

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 IDTC (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 de l'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 de l'ITAO) avant que l'on applique pour de bon la méthode de l'IDTC, 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 de l'IDTC.

Sur le plan des catégories de dépenses salariales, l'utilisation de l'IDTC 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 mensuelle d'une importante opération de bureau dans les États. De même, les opérations de saisie et de validation de la saisie par lots ainsi que de contrôle des formules qu'impose la méthode de collecte par la poste ont été complètement éliminées grâce à l'IDTC, dans le cadre de laquelle le répondant introduit à l'aide d'un téléphone à clavier les données relatives à son entreprise, puis valide lui-même chacune des entrées. Autre caractéristique d'importance,

| Catégorie de dépenses | Poste | ITAO | Auto-interview (IDTC et RP) |
|---|---|---|---|
| **DÉPENSES SALARIALES** Envoi postal | ↗ | | |
| Retour par la poste | ↗ | | |
| Saisie 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 16 bits) |

**Figure 2.** Coûts de collecte des données (les flèches indiquent la direction des récentes variations de prix)

l'IDTC possède, sur le plan de la rentabilité et de la qualité, un avantage certain par rapport à la collecte par la poste du fait qu'elle facilite grandement le suivi téléphonique dans les cas de non-réponse. Grâce à l'IDTC, il est possible de produire 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 afin de rappeler à ces personnes qu'elles doivent communiquer leurs données. 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 envoyée par la poste, ou encore si elle en était au stade de l'enregistrement ou à celui de la saisie; en outre, les répondants ayant récemment envoyé leur formule par la poste avaient tendance à s'offenser de recevoir un rappel au sujet d'une activité qu'ils considéraient terminée. Comme la participation à l'enquête était volontaire et qu'on ignorait si le répondant avait rempli son questionnaire ou non, ce n'est qu'aux unités critiques (gros employeurs) qu'on téléphonait pour leur rappeler de communiquer leurs données.

L'utilisation de la méthode de l'IDTC n'a permis de réaliser aucune économie sensible au chapitre de la conciliation, puisque le nombre de rejets au contrôle s'est établi à peu près au même niveau pour la collecte par IDTC que pour la collecte par la poste. Il en a été de même pour les cartes postales de rappel, le nombre des cartes envoyées aux établissements donnant une réponse tardive quand on utilisait 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 moyen d'un téléphone à clavier à la date prévue quand on employait la méthode de collecte par IDTC.

Sur le plan des catégories de dépenses non salariales, les frais d'affranchissement (actuellement 58 cents par unité) sont remplacés par la somme des frais d'appel et de la fraction non amortie du coût de l'appareil utilisé pour l'IDTC (somme s'élevant actuellement à 46 cents par unité). Les frais d'affranchissement augmentent constamment avec une hausse annuelle de l'ordre de 5% environ (figure 2). Cette majoration est causée par un accroissement annuel des frais de main-d'oeuvre (+ 5.7%) et par le prix de l'essence qui augmente aussi de façon générale, les frais de main-d'oeuvre intervenant pour plus de 80% des frais totaux d'affranchissement. Par contraste, pour la méthode de l'IDTC, 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 pour ce qui est de l'obligation nouvelle de téléphoner aux répondants pour leur rappeler de communiquer leurs données, dans les cas de non-réponse complète, on peut démontrer que le passage de la collecte par la poste à la collecte par IDTC 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 l'utilisation de ces deux méthodes de collecte (figure 5), ces économies augmenteront considérablement. On tentera d'utiliser les économies réalisées à l'avenir grâce à l'utilisation de l'IDTC pour compenser les frais relatifs à l'obligation de téléphoner aux répondants afin de leur rappeler de communiquer leurs données dans les cas de non-réponse complète.

Cet examen du rendement et de l'analyse des coûts nous permettent de tirer plusieurs conclusions importantes relativement à la déclaration des données à l'aide de la méthode de l'IDTC. (1) 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. (2) Le fait que les répondants cessent de répondre par la poste et qu'ils passent à la méthode de l'IDTC semble

causer très peu de perturbations dans la production des déclarations mensuelles. Selon les interviews de suivi réalisées par le Bureau auprès des répondants pour lesquels la méthode de collecte des données est maintenant l'IDTC, ces derniers éprouvent très peu de difficulté à s'adapter à cette nouvelle méthode de déclaration. Virtuellement tous les répondants ont rempli leur première déclaration mensuelle par IDTC avec précision et sans aide et de nombreux répondants ont mentionné comment il était facile de produire une déclaration à l'aide de cette méthode. (3) On peut considérer que l'IDTC constitue une méthode de remplacement fiable pour la collecte des données d'enquête. Au cours des quatre dernières années de collecte, on a enregistré ni défaillance majeure du matériel ni interruption du processus de collecte. Au besoin, les défaillances mineures du matériel ont pu facilement être résolues à l'aide d'un micro-ordinateur de secours. En outre, le processus de collecte par IDTC dans les États prévoira une option de renvoi automatique des appels, destinée à réacheminer les appels à un site 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 une différence importante 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'adoption 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 administré par les États 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 plus de 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 coûts 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 importantes supérieures à 100,000 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 IDTC 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 entreprises 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 en définitive 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, les systèmes d'ITAO et d'IDTC ont pu être modifiés 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 de l'ITAO s'était transformée d'une simulation malhabile d'interviews de type enquête-ménage en une technique rapide et efficace à "écrans" et "fenêtres" bien adaptée 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 améliorations dans un large éventail d'États. Cette façon d'aborder les essais a permis de développer, au niveau des États, une confiance solide dans les méthodes et les systèmes proposés. Cette caractéristique s'est révélée 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, dans environ 40% des cas, la différence entre les estimations provisoires et les estimations définitives s'élevait à plus de 50,000 et dans 13% des cas à plus de 100,000 (c.-à-d. qu'il s'agissait de révisions importantes) (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 pour la diffusion, permettraient de maintenir l'ampleur des révisions au-dessous d'un seuil jugé acceptable (modification s'élevant à moins de 50,000 par rapport au chiffre des fluctuations mensuelles relatives à l'emploi). Bien sûr, il aurait été possible de demander à l'ensemble des 175,000 entreprises donnant des réponses tardives de produire leurs déclarations à l'aide des 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 changeant la méthode de collecte utilisée pour le plus petit nombre possible d'entreprises afin de permettre le maintien des révisions au-dessous du seuil de 50,000).

Contrairement aux enquêtes-ménages, les enquêtes auprès des établissements utilisent en général une pondération différentielle pour les diverses unités de collecte, le plan de sondage prévoyant que les unités très importantes doivent faire partie de l'échantillon. Dans 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'il y en a 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. Pour vérifier l'incidence des gros employeurs, les unités comptant 100 employés ou plus 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 nouvelles estimations aux estimations provisoires initiales afin de déterminer l'incidence des unités comptant 100 employés ou plus sur les révisions. Les résultats ont indiqué qu'entre une demie et deux tiers de la révision était attribuable à ces unités. On a ensuite obtenu des résultats similaires en effectuant les mêmes

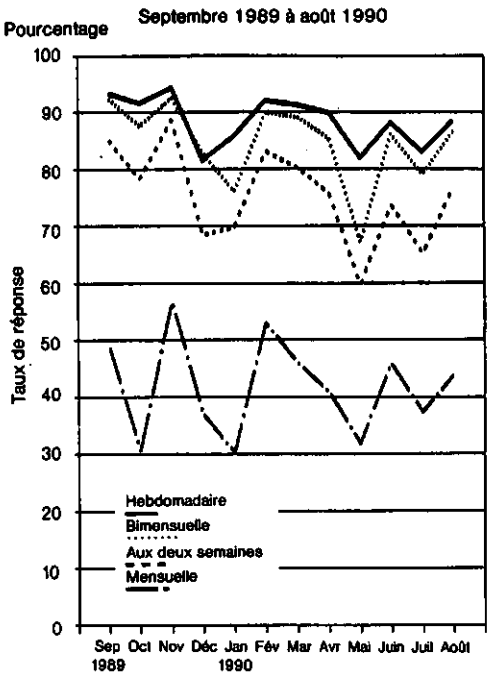



**Figure 3.** CESS, rendement de l'ITAO à la première date limite, selon la durée de la période de paie

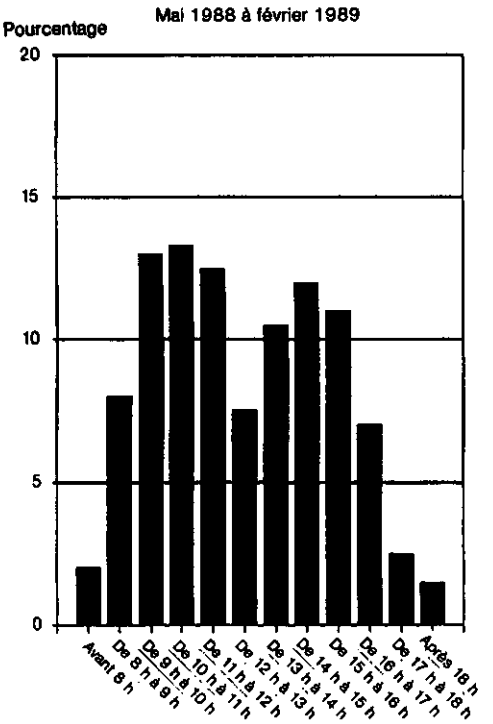



**Figure 4.** Introduction des données au moyen d'un téléphone à clavier

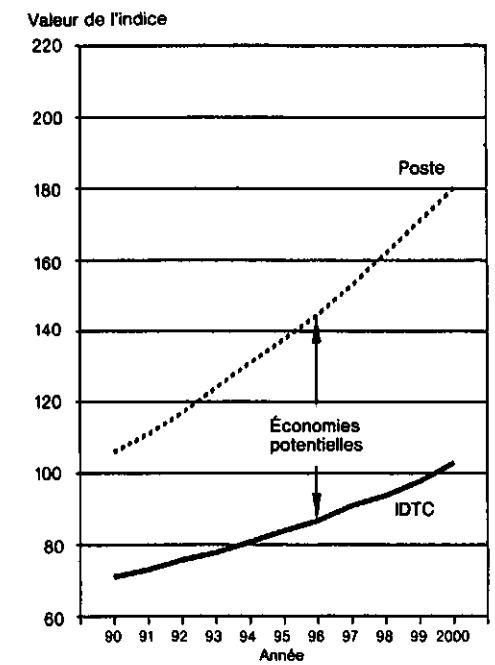



**Figure 5.** Coûts estimatifs par unité, selon le mode: de 1990 à 2000

Nota: Projections non officielles du BLS

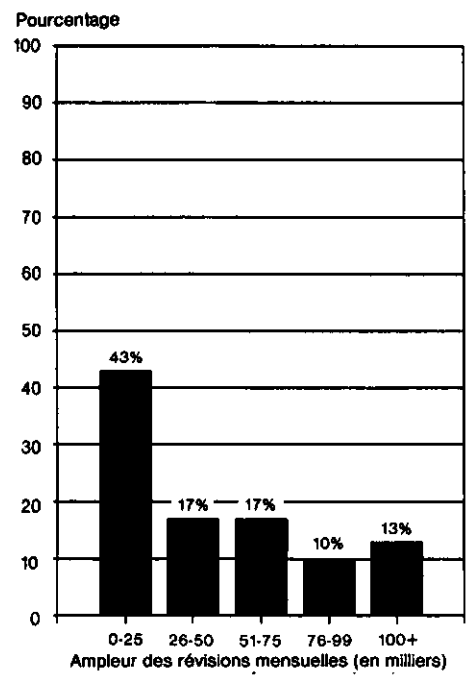



**Figure 6.** Répartition des révisions selon leur ampleur (De 1985 à 1989)

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 comparativement à seulement 60% actuellement. Cette réduction considérable de la taille de l'échantillon visé par le passage à la méthode de l'ITAO/de l'IDTC 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 par ITAO/par IDTC dans 25 États en 1991 et prévoit étendre l'utilisation de ces méthodes à tous les États en 1992. Grâce à la mise en oeuvre de 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 de nos réalisations, 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 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'aptitude à l'usage se sont considérablement accrues. 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, dans le domaine de la micro-informatique. La nouvelle technologie a offert aux organismes statistiques de nombreuses occasions nouvelles d'améliorer la qualité et le contrôle de la collecte des données, par exemple: l'ITAO, l'AIAO, l'IDTC, la RP et le télécopieur. Certaines de ces techniques permettent d'améliorer la qualité et le contrôle tout en réduisant les coûts de fonctionnement courants. 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 diminuant les dépenses.

Lorsque nous examinons l'état de nos programmes statistiques, nous découvrons fréquemment un cadre de mise en oeuvre très rigide. Souvent, la méthode de collecte des données utilisée pour nos enquêtes n'a jamais été modifiée 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 d'une enquête, elle est d'ordinaire bien ancrée dans la structure organisationnelle de l'organisme et il peut être très difficile de la restructurer afin de permettre d'effectuer des modifications d'envergure. C'est à l'intérieur de ce cadre de réalisation des enquêtes 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 suivent l'évolution rapide de la technologie et des méthodes de collecte automatisée des données; enfin, peut-être par dessus tout,
- de continuer à trouver des façons d'incorporer les résultats des travaux de recherche couronnés de succès aux programmes courants.

Ces défis détermineront la future position concurrentielle de nos programmes et de nos organismes en termes de coûts et de qualité.

## BIBLIOGRAPHIE

CLAYTON, R.L., et HARRELL, L., Jr. (1989). Developing a cost model of alternative data collection methods: MAIL, CATI et TDE. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

CLAYTON, R.L., et WINTER, D.L.S. (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.*

JURAN, J.M., GRYNA, F.N., Jr., et BINGHAM, R.S., Jr., éds. (1979). *Quality Control Handbook*, troisième édition. McGraw Hill.

GROVES, R.M.J., et coll., éds. (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.

OFFICE OF MANAGEMENT AND BUDGET (1988). *Quality in Establishment Surveys*. Statistical Policy Working Paper 15.

OFFICE OF MANAGEMENT AND BUDGET (1990). *Computer Assisted Survey Information Collection*. Statistical Policy Working Paper 19.

PONIKOWSKI, C., et MEILY, S. (1988). Use of touchtone recognition technology in establishment survey data collection. Presented at the First Annual Field Technologies Conference, St. Petersburg, Florida.

WERKING, 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*, 349-362.

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 17–30
Statistique Canada



# Fiabilité des données introduites au moyen d'un téléphone à clavier

POLLY A. PHIPPS et ALAN R. TUPEK[1]

RÉSUMÉ

Le Bureau of Labor Statistics a mis en oeuvre, dans le cadre d'une enquête mensuelle auprès des entreprises, une méthode de collecte électronique des données utilisant la reconnaissance des données introduites au moyen d'un téléphone à clavier. Le système d'introduction des données au moyen d'un téléphone à clavier (IDTC) utilise des expressions numérisées pour demander aux participants à l'enquête de répondre aux questions en appuyant sur les boutons d'un poste téléphonique à clavier numérique. Le système IDTC permet de réduire de façon substantielle les coûts d'enquête par suite de l'élimination de nombreuses activités à prédominance de main-d'oeuvre. On en sait toutefois très peu au sujet des erreurs de mesure découlant de l'utilisation de ce mode de collecte. Dans la présente étude, nous évaluons les erreurs inhérentes à l'emploi de la méthode IDTC à partir de trois sources de données: il est ainsi possible d'analyser les erreurs reliées à des aspects précis de l'interface homme-machine. De plus, nous faisons état de certaines caractéristiques de conception du système qui ont des répercussions sur les erreurs dues à ce mode de collecte. Nous terminons en traitant de l'incidence de nos constatations sur d'autres enquêtes.

MOTS CLÉS: Mode de collecte des données; interface homme-machine; auto-interview assistée par ordinateur.

## 1. INTRODUCTION

Le Bureau of Labor Statistics (BLS) des États-Unis diffuse chaque mois des estimations de l'emploi pour l'ensemble des États-Unis. Ces estimations sont établies à partir des données recueillies au moyen d'une enquête réalisée auprès de 350,000 entreprises, la Current Employment Statistics Survey (CESS), qui fournit l'une des premières mesures mensuelles de l'état de santé de l'économie américaine. Toutefois, les estimations provisoires de l'enquête sont établies à partir de données ne provenant que d'environ la moitié des entreprises participant à l'enquête, des estimations révisées étant produites deux mois après la diffusion initiale des données aux médias. Il peut arriver, en raison du faible taux de réponse enregistré au moment de l'établissement des estimations provisoires, que ces dernières doivent faire l'objet de révisions importantes. En 1983, le BLS a entrepris des travaux de recherche sur l'utilisation de techniques de collecte automatisée des données afin d'assurer un meilleur respect des délais de réponse et de réduire l'importance éventuelle des révisions nécessaires.

Traditionnellement, les questionnaires de la CESS étaient recueillis par la poste, par l'intermédiaire des organismes d'État responsables de la sécurité de l'emploi. Des essais réalisés entre 1984 et 1986, au cours desquels on a remplacé la méthode de collecte par la poste par une interview téléphonique assistée par ordinateur (ITAO), ont démontré que la méthode ITAO constituait un moyen efficace de réduire les délais de réponse (Werking, Tupek, Ponikowski et Rosen 1986). Les taux de réponse moyens obtenus pour les estimations provisoires en utilisant la méthode ITAO se sont situés entre 85% et 90%, comparativement à 45% à 50% avec la méthode de collecte des données par la poste. Bien que la méthode ITAO ait permis de réduire

[1] Polly A. Phipps, Office of Employment and Unemployment Statistics, Bureau of Labour Statistics, Room 2821, 441 G Street N.W., Washington, D.C. 20212; Alan R. Tupek, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics, Room 2919, 441 G Street N.W., Washington, D.C. 20212.

les délais de réponse, le coût de l'utilisation exclusive de la méthode ITAO pour la CESS ne peut pas être absorbé à l'intérieur du budget actuel affecté à cette enquête. Des travaux de recherche sur l'IDTC ont été réalisés depuis 1986 afin d'élaborer une autre méthode de collecte des données qui offrirait les mêmes améliorations au chapitre du rendement que la méthode ITAO, mais à un coût moindre (Ponikowski, Copeland et Meily 1989). Dans leur article, Werking et Clayton font une analyse plus poussée de l'utilisation des méthodes ITAO et IDTC pour la CESS.

Le BLS a effectué récemment des essais sur le système IDTC afin d'obtenir des données sur les délais de réponse, sur le coût de la collecte et sur les taux de rejet au contrôle. Ces essais ont montré que la collecte par IDTC utilisant l'ITAO comme méthode d'appoint est aussi efficace que la méthode ITAO et qu'elle permet de respecter des délais de réponse aussi serrés (Werking, Tupek et Clayton 1988). De plus, l'IDTC permet de réduire considérablement les coûts afférents à l'enquête. Selon les estimations actuelles, les coûts unitaires mensuels de la collecte par IDTC sont d'environ 30% inférieurs à ceux de la collecte par la poste, tandis que la collecte par ITAO est 20% plus coûteuse que la collecte par la poste (Clayton et Harrell 1990). On étend maintenant l'utilisation de la méthode de collecte par IDTC à une plus grande partie de la population de la CESS, en commençant par les entreprises comptant le plus grand nombre d'employés. Les travaux de recherche visant à trouver des moyens d'améliorer le système IDTC se poursuivent, même si les répondants ont réservé un accueil très favorable à celui-ci.

Les travaux de recherche auxquels le BLS procède actuellement ont pour objet de cerner les problèmes que les répondants éprouvent à utiliser le système IDTC et de mesurer les erreurs dues à ce mode de collecte des données. Les résultats de ces travaux indiquent les points qu'il faut améliorer pour faciliter la tâche des répondants ainsi que pour réduire les erreurs. La section 2 contient des renseignements généraux sur la méthode IDTC et sur le fonctionnement du système IDTC du BLS. Dans la troisième section, nous traitons des erreurs qui peuvent être reliées à l'utilisation de l'IDTC, qui nécessite une interface homme-machine, comme mode de collecte des données. Dans la section 4, nous décrivons les trois sources de données utilisées aux fins de l'étude. Les problèmes et les erreurs sont analysés à l'aide des résultats d'une enquête de vérification des enregistrements, de données enregistrées mécanographiquement, ainsi que des résultats d'une enquête d'évaluation auprès des répondants. La section 5 fait état des méthodes utilisées, des analyses effectuées et des résultats obtenus à partir de chacune des trois sources de données. Enfin, dans la section 6, nous présentons une évaluation globale des erreurs de mesure dues à ce mode de collecte, certaines suggestions en vue d'améliorer le système ainsi que certaines observations quant à l'utilisation éventuelle de ce système pour d'autres enquêtes.

## 2. INTRODUCTION DES DONNÉES AU MOYEN D'UN TÉLÉPHONE À CLAVIER

La principale raison pour laquelle on envisage d'utiliser le mode de collecte par IDTC est afin de réduire le coût de collecte des données par ITAO, tout en maintenant l'actualité et la qualité des données au niveau actuel. La CESS semblait bien se prêter à la collecte des données par IDTC, puisque seulement cinq ou six éléments d'information numériques y sont recueillis chaque mois. Ces éléments d'information comprennent: le nombre total d'employés, le nombre d'employés de sexe féminin, le nombre d'employés à la production, la rémunération des employés à la production, le nombre d'heures de travail des employés à la production et, dans le cas de certaines activités économiques, le nombre d'heures supplémentaires ou les commissions versées aux employés. Les entreprises sont priées de déclarer les totaux relatifs à chaque élément d'information pour la période de paye dans laquelle tombe le 12$^{e}$ jour du mois. On demande aussi aux répondants employant la méthode IDTC d'indiquer le numéro d'identification de leur entreprise et le mois pour lequel ils fournissent des données.

Grâce au système IDTC du BLS, il est possible:

- de repérer les entreprises répondantes légitimes d'après un appariement avec un fichier de numéros d'entreprise;
- de faire varier la série de questions selon l'activité économique exercée par l'entreprise;
- de repasser toutes les réponses afin que le répondant puisse les confirmer, à l'aide d'une voix simulée sur ordinateur (numérisée) (on demande aux répondants d'inscrire un "1" pour confirmer leur réponse ou un "0" pour la réintroduire);
- d'attendre deux secondes avant que les répondants inscrivent leur réponse, et d'attendre deux secondes entre les chiffres avant de considérer l'introduction des données comme étant terminée (on suppose aussi que l'introduction des données est terminée si toutes les positions de la zone ont été remplies; par exemple, la zone correspondant au "mois" comprend deux chiffres);
- de répéter chaque question jusqu'à trois fois (pour le numéro d'identification, le mois et le nombre total d'employés) ou de demander au répondant de confirmer qu'il n'a pas de réponse pour la question (dans le cas de tous les autres éléments d'information), lorsqu'un répondant ne confirme pas sa réponse ou ne donne pas de réponse dans les deux secondes suivant la lecture de la question;
- d'enregistrer la date, l'heure du début et de la fin de l'appel, et tous les éléments d'information (Werking et coll., 1988).

Les répondants reçoivent par la poste des instructions sur l'IDTC dans lesquelles on leur indique la façon d'utiliser le système et on leur donne des exemples de dialogue entre l'ordinateur et le répondant, comme celui-ci:

| **Ordinateur:** | **Votre réponse:** |
|---|---|
| Inscrivez le nombre total | Dans le cas de 25 employés, appuyez sur le 2 et le 5 |
| Vous avez inscrit 2, 5 | Appuyez sur 1 pour confirmer ou sur le 0 pour réintroduire des données. |

Dans les instructions, on indique aussi aux répondants qu'ils ont le choix d'utiliser le dièse ("#") pour signaler qu'ils ont terminé l'introduction des données pour une question, ce qui permet de réduire la durée de l'interview. En outre, avant de déclarer des données, les répondants peuvent téléphoner pour faire l'essai du système en utilisant un numéro d'identification spécial qui leur est délivré à des fins d'essai.

Un intervieweur communique par téléphone avec les répondants employant la méthode IDTC durant leur premier mois d'utilisation du système pour savoir s'ils ont éprouvé des problèmes et, au besoin, pour leur prodiguer des conseils. Chaque mois, les répondants reçoivent une carte de rappel par la poste et, s'ils n'ont pas répondu d'eux-mêmes avant une date préétablie, un appel de sollicitation leur demandant d'entrer en communication avec le système IDTC le plus tôt possible. En général, les données ne sont pas recueillies au moment de l'appel de sollicitation.

## 3. ERREUR DE MESURE DÉCOULANT DE L'UTILISATION D'UNE INTERFACE HOMME-MACHINE

Il existe peu de précédents en matière d'utilisation d'un système IDTC pour répondre aux questions d'une enquête. Toutefois, la reconnaissance des données introduites au moyen d'un clavier est très répandue dans le cas de services tels que les virements bancaires électroniques

et le service téléphonique automatique. Bien que ces services permettent de gagner du temps et d'économiser de l'argent, ils peuvent parfois rebuter les utilisateurs. Par ailleurs, le système, la tâche ou le répondant peuvent tous être la source de problèmes et d'erreurs. Les problèmes liés au système engendrent principalement une erreur due à la non-réponse, tandis que la complexité de la tâche et la qualité de son exécution par le répondant sont la cause d'une erreur de mesure.

Quoiqu'ils ne traitent pas directement des enquêtes, les ouvrages sur l'ergonomie indiquent plusieurs facteurs interdépendants pouvant contribuer aux erreurs d'exécution découlant de l'utilisateur d'une interface homme-machine. D'une part, il se peut que les répondants ne soient ni familiers ni à l'aise avec la technologie. D'après Waterworth (1984), le langage employé dans l'interface homme-machine est différent de la communication humaine étant donné que les actes sont exécutés suivant un ordre reflétant la logique de la programmation informatique. Comme il n'est pas facile d'adopter un mode de pensée épousant cette logique, les personnes ayant peu d'expérience de l'informatique peuvent éprouver de la difficulté à comprendre la tâche qu'elles doivent exécuter et à utiliser le système. D'autre part, la parole synthétique est plus difficile à comprendre que la parole naturelle et pose des exigences plus élevées sur la fonction traitement de la mémoire immédiate (Schwab, Nusbaum et Pisoni 1985). Ainsi, les problèmes de compréhension et de mémoire associés à ce mode de collecte des données peuvent être la source d'erreurs.

La parole synthétique comprend à la fois la parole numérisée, où une voix humaine est échantillonnée, codée numériquement et enregistrée, et la parole synthétisée suivant des règles, laquelle est produite à l'aide d'un texte servant d'intrant (Marics et Williges 1988). Le système IDTC utilise la parole numérisée, laquelle est moins difficile à comprendre que la parole synthétisée suivant des règles. Toutefois, la numérisation soulève des problèmes de compréhension du fait qu'elle provoque une distorsion de la parole naturelle (Cox et Coope 1981). Certains travaux de recherche ont toutefois démontré qu'il est possible d'améliorer la compréhension de la parole synthétique par la formation. Lors d'une expérience sur la perception de la parole synthétique, Schwab et ses collaborateurs (1985) ont découvert que le fait de recevoir une formation sur la parole synthétique permet d'améliorer la perception. Ainsi, la compréhension de cette parole peut s'améliorer à mesure qu'on acquiert de l'expérience dans l'utilisation du système. Le débit constitue un autre facteur qui peut influer sur la compréhension. Marics et Williges (1988) ont découvert, en mesurant les erreurs de transcription et le temps de latence (temps écoulé entre la fin de la question et le début de la réponse correspondante), que le débit de la parole synthétique a une incidence considérable sur son intelligibilité. Toutefois, les sujets qui ont reçu des renseignements contextuels avant d'écouter la parole ont commis moins d'erreurs de transcription.

Par conséquent, les erreurs susceptibles de découler de l'utilisation d'une interface homme-machine peuvent être attribuables à un manque d'expérience avec la technologie et la tâche à exécuter, ainsi qu'à des problèmes de compréhension et de mémoire reliés au débit et à la clarté de la voix. Ces problèmes peuvent néanmoins être surmontés, puisqu'il a été démontré que, avec de l'expérience et une formation adéquate, il est possible d'améliorer le rendement.

## 4. LES DONNÉES

Plusieurs objectifs ont été pris en considération au moment de l'évaluation des problèmes soulevés par l'IDTC et de l'erreur due au mode de collecte, ainsi que de la détermination des données à utiliser ou à recueillir. Premièrement, il était nécessaire de déterminer si des problèmes se posaient et, le cas échéant, à quelle étape du processus ils se posaient. Deuxièmement, nous estimions que les répondants devaient cerner et interpréter eux-mêmes les problèmes, mais nous voulions aussi obtenir des mesures indépendantes des évaluations des répondants. Troisièmement,

il nous fallait étudier les erreurs et les problèmes découlant de la complexité de la tâche et de sa compréhension par les répondants, y compris la possibilité d'améliorer le rendement des répondants avec le temps.

Nous avons décidé d'évaluer les problèmes soulevés par l'IDTC et l'erreur due à ce mode de collecte à partir de trois sources de données différentes, portant toutes sur un groupe commun d'environ 465 entreprises de la Pennsylvanie. Ces entreprises déclaraient leurs données d'enquête mensuelles par IDTC au laboratoire des techniques de collecte automatisée (ACT), au siège national du BLS, à Washington. Un petit nombre de ces entreprises ont commencé à déclarer leurs données au laboratoire ACT par IDTC en avril 1989 et les autres se sont graduellement jointes à ce groupe de départ d'un mois à l'autre, jusqu'en novembre 1989. La plupart des entreprises ont continué de déclarer des données au laboratoire ACT jusqu'en avril 1990. La majorité de ces entreprises ont passé de la méthode de collecte des données par la poste à la méthode IDTC.

Le premier ensemble de données se divise en deux composantes: les données IDTC qui ont été enregistrées mécanographiquement d'avril à décembre 1989 et les mêmes données consignées par les entreprises sur une formule d'enquête. Tous les répondants reçoivent chaque année une formule d'enquête sur laquelle on leur demande de consigner leurs données pour chaque mois. Les entreprises qui répondent par la poste remplissent la formule chaque mois et l'envoient par la poste à l'organisme d'État responsable de la sécurité de l'emploi. Ce dernier enregistre les données, puis renvoie la formule par la poste pour la collecte des données du mois suivant. Les entreprises qui répondent par ITAO et par IDTC reçoivent la formule par la poste, mais elles ne la renvoient pas. Toutefois, nous avons demandé aux répondants employant la méthode IDTC de renvoyer leur formule d'enquête de 1989 et nous avons enregistré un taux de retour de 96%. Nous avons ensuite comparé les données obtenues par IDTC et celles qui avaient été consignées sur la formule afin de détecter les divergences entre les deux. Les deux sous-ensembles de données comprennent 1,930 observations faites sur une période de neuf mois. Étant donné qu'on a introduit lentement et progressivement la méthode IDTC auprès des entreprises, le nombre d'observations par entreprise varie. Nous disposons de données couvrant une période de six à neuf mois pour environ 75 entreprises, une période de quatre à six mois pour environ 200 entreprises et une période de deux à trois mois pour environ 190 entreprises. Ces données sont désignées sous le nom de données de vérification des enregistrements.

Le deuxième ensemble de données comprend des renseignements enregistrés mécanographiquement qui se rapportent au rendement des répondants durant l'appel téléphonique IDTC. En janvier 1990, on a reprogrammé les ordinateurs utilisés pour l'IDTC afin qu'ils totalisent et enregistrent automatiquement le nombre de fois qu'une question a été répétée en raison d'une non-réponse (catégorie des questions répétées), le nombre de fois qu'un répondant a réintroduit des données (catégorie des données réintroduites) pour chaque question, ainsi que le nombre de fois qu'une entreprise a appelé et raccroché avant d'introduire des données. Malheureusement, les seules questions pour lesquelles il a été possible d'effectuer des relevés distincts pour les catégories des questions répétées et des données réintroduites sont celles concernant le mois et le nombre total d'employés. Dans le cas des questions portant sur le nombre d'employés de sexe féminin, sur le nombre d'employés à la production, sur la rémunération et sur les heures de travail, la structure du programme machine initial nous a forcés à combiner les données sur les répétitions de questions et sur les réintroductions de données. Ces données sont désignées sous le nom de données enregistrées mécanographiquement.

La troisième source de données est une enquête d'évaluation téléphonique qui a été réalisée de janvier à avril 1990 auprès des entreprises de la Pennsylvanie et qui portait sur leur expérience avec le système IDTC. Environ 411 entreprises ont répondu à l'interview, pour un taux de réponse de 88%. Les questions portaient sur des sujets tels que la qualité de la voix, le débit, les problèmes relatifs à la tâche, l'utilisation des fonctions du système, l'à-propos des instructions et une évaluation du système.

## 5. RÉSULTATS

### 5.1 Données de vérification des enregistrements

Lorsque nous avons choisi de demander aux entreprises qui répondaient par IDTC de renvoyer leurs formules d'enquête, nous nous sommes d'abord demandé combien de répondants avaient vraiment utilisé les formules. Nous supposions qu'une des sources de l'erreur due au mode de collecte était le fait que les répondants ne remplissaient pas la formule avant d'introduire leurs données au téléphone à clavier, ce qui plaçait un fardeau accru sur leur mémoire. Selon nous, ces répondants couraient plus de risques d'introduire et/ou de confirmer des données inexactes. C'est pourquoi, lorsque nous avons demandé aux répondants de nous renvoyer les formules d'enquête, nous leur avons indiqué qu'ils devaient nous les retourner peu importe s'ils les avaient remplies ou non. Toutefois, parmi les 96% d'entreprises qui ont renvoyé leur formule d'enquête, une seule a retourné une formule vierge. Bien qu'il soit possible que les non-répondants travaillent de mémoire, la plupart des répondants avaient rempli leur formule, ce qui nous amène à croire que les problèmes de mémoire dus au fait qu'on n'avait pas rempli la formule ne constituaient pas une source majeure d'erreurs.

En comparant les données obtenues par IDTC et celles figurant sur les formules d'enquête, nous avons détecté et codé certaines divergences. Les données figurant sur la formule d'enquête sont celles que nous aurions obtenues si les répondants avaient participé à l'enquête par la poste. On trouve les résultats de cette comparaison au tableau 1. Le premier type de divergence était le suivant: les données IDTC indiquaient qu'on n'avait pas répondu à une question, mais une réponse figurait sur la formule d'enquête de l'entreprise. Cette non-réponse à certaines questions intervenait pour le plus grand nombre de divergences (82 sur 177) et était répartie assez uniformément sur toutes les questions de l'enquête.

Par ailleurs, la non-réponse aux questions variait selon le mois et l'entreprise. Ainsi, le taux de non-réponse était de 40% supérieur le premier mois où une entreprise déclarait les données à l'aide de la méthode IDTC. De plus, certaines entreprises éprouvaient plus de difficultés que d'autres (comme l'indique le fait qu'elles aient laissé deux questions ou plus sans réponse). Presque la moitié des cas de non-réponse à certaines questions ont été enregistrés auprès de 18 entreprises, au moment où elles ont commencé à utiliser la méthode IDTC ou peu après. Cette constatation nous a permis de conclure que la première utilisaton de la méthode soulevait certains problèmes susceptibles de s'atténuer avec le temps. Comme le problème était circonscrit à un petit groupe d'entreprises, nous croyons qu'il découlait d'un manque de connaissance des processus automatisés. Les autres cas de non-réponse ne suivaient aucune tendance identifiable; nous avons présumé que certains répondants oubliaient tout simplement de répondre à une question, peut-être en raison d'une source de distractions dans leur bureau, et passaient à la question suivante.

**Tableau 1**

Données de vérification des enregistrements - Nombre et type de divergences obtenues avec la méthode IDTC

| | |
|---|---|
| Non-réponse à une question par IDTC | 82 |
| 1 à 2 chiffres manquants ou de trop | 18 |
| Glissement sur le clavier | 17 |
| Divergences/confirmation "1", erreur "0" | 14 |
| Formule corrigée, mais pas les données IDTC | 12 |
| Aucune raison apparente pour les erreurs | 26 |
| Autres raisons | 8 |
| Total | 177 |

Le deuxième type de divergence était l'inscription de chiffres additionnels ou, dans quelques cas, de trop peu de chiffres (18 erreurs sur 177 tombaient dans cette catégorie). Ce problème s'est surtout produit à la question sur la rémunération, où quatre répondants ont essayé d'indiquer les cents au lieu d'arrondir au dollar près. Plusieurs des mêmes répondants ont semblé inscrire une demi-heure, 50, comme nombre d'heures de travail des employés à la production, plutôt que d'arrondir. Dans le cas du troisième type de divergence, les chiffres introduits au moyen du téléphone à clavier étaient identiques à ceux figurant sur la formule, à une exception près. Le nombre inscrit incorrectement nous a laissé croire que le répondant avait peut-être éprouvé des problèmes avec le clavier (ses doigts ayant pu glisser sur la touche se trouvant directement à côté ou au-dessous de la touche correspondant au bon chiffre). Nous avons relevé 17 divergences de ce type. Pour ce qui est du quatrième type de divergence, on l'a détecté principalement à la question portant sur le nombre total d'employés: huit entreprises ont inscrit un "1" en réponse à cette question en utilisant la méthode IDTC, alors qu'un nombre plus élevé d'employés figurait sur leur formule d'enquête. Nous supposons que les répondants ont appuyé deux fois sur le "1" lorsqu'ils ont confirmé leur réponse à la question précédente (le mois).

Enfin, quelques répondants avaient corrigé certaines données sur leur formule, sans les corriger au moment de la déclaration au téléphone à clavier. Nous avons aussi relevé d'autres divergences que nous n'avons pu expliquer. En outre, plusieurs répondants ont transposé les chiffres déclarés ou les ont déclaré avec un décalage d'une catégorie ("autres raisons"). Dans la plupart des cas, il était difficile d'établir explicitement si les erreurs avaient été commises au moment de l'introduction des données ou au moment de la vérification, par suite d'une mauvaise compréhension de la question ou des chiffres lors de leur relecture par l'ordinateur. Bien que nous ayons soupçonné qu'elles avaient été commises au moment de l'introduction des données, il nous a été impossible de vraiment écarter la possibilité que les problèmes de compréhension aient eu une incidence, sauf dans le cas du deuxième type de divergence.

Comme l'indique le tableau 2, les taux d'erreurs enregistrés pour les différentes questions de l'enquête variaient de 1.2% à 2.5%. De même, on observe un taux d'erreurs moins élevé pour les questions portant sur le nombre total d'employés, sur le nombre d'employés de sexe féminin et sur le nombre d'employés à la production que pour les questions portant sur la rémunération et sur les heures de travail. Cette situation n'a rien de surprenant, puisqu'il faut habituellement inscrire de quatre à six chiffres en réponse aux questions portant sur la rémunération et sur les heures de travail, comparativement à deux ou trois dans le cas des autres questions. Ainsi, les séries de chiffres plus longues posent davantage de problèmes aux répondants. Ces problèmes peuvent être attribuables au fait que les répondants éprouvent des difficultés à introduire les données au clavier ou à se rappeler de séries de chiffres plus longues durant la validation, ou encore au fait qu'ils ne sont pas motivés à apporter les corrections nécessaires.

**Tableau 2**

Données de vérification des enregistrements - Nombre de divergences et pourcentage d'erreurs par élément d'information

| | Nombre total d'employés | Nombre d'employés de sexe féminin | Nombre d'employés à la production | Rémunération | Heures de travail des employés à la production | Total |
|---|---|---|---|---|---|---|
| Divergences | 23 | 29 | 28 | 48 | 49 | 177 |
| Erreur en pourcentage | 1.2 | 1.5 | 1.5 | 2.5 | 2.5 | 1.8 |
| (Estimations d'enquête) | (.2) | (.3) | (.3) | (.4) | (.4) | (.3) |

(N = 1,930 pour chaque question).

**Tableau 3**

Données de vérification des enregistrements - Erreur due au mode de collecte par IDTC avant et après les corrections découlant du contrôle

| | Nombre total d'employés | Nombre d'employés de sexe féminin | Nombre d'employés à la production | Rémunération | Heures de travail des employés à la production |
|---|---|---|---|---|---|
| Erreur due au mode de collecte en pourcentage, avant les corrections découlant du contrôle | 0.0 | 0.5 | 7.3 | 7.3 | 4.4 |
| (Estimations de l'enquête) | (.4) | (.3) | (5.2) | (3.8) | (3.7) |
| Erreur due au mode de collecte en pourcentage, après les corrections découlant du contrôle | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| (Estimations de l'enquête) | (0.0) | (0.0) | (0.0) | (0.0) | (0.0) |

(N = 1,886 à 1,930 pour chaque question).

Le tableau 3 montre les effets possibles des divergences sur les éléments d'information de la CESS, effets déterminés en faisant la somme des écarts entre les valeurs obtenues par la méthode IDTC et celles figurant sur la formule, puis en divisant le résultat par la somme de ces dernières. Dans le cas de la CESS, les estimations publiées sont établies à l'aide d'un estimateur fondé sur une chaîne de rapports. Les estimations contenues dans le tableau 3 ne tiennent pas compte de cet estimateur. Elle permettent toutefois de mesurer indirectement l'erreur due au mode de collecte par IDTC entachant les estimations de l'enquête. Au seuil de 5%, aucune des erreurs ne s'écarte de façon significative de zéro. Toutefois, les risques d'erreurs dues au mode de collecte semblent être plus élevés dans le cas du nombre d'employés à la production, de la rémunération et des heures de travail des employés à la production. Dans cette étude, on a surestimé le nombre des employés à la production et la rémunération de 7.3%, et les heures de travail de 4.4%.

Par ailleurs, presque toutes les réponses divergentes auraient été rejetées au contrôle (selon les paramètres utilisés dans la CESS) et corrigées. Comme l'illustre le tableau 3, les corrections effectuées au contrôle ont pour résultat d'annuler l'effet des réponses divergentes sur les données. On peut citer comme exemples de divergences importantes le cas d'un répondant qui a inscrit par erreur la rémunération à la place du nombre d'employés à la production, augmentant ainsi ce dernier de plus de 10,000, et celui de deux répondants qui ont inscrit le nombre d'heures de travail des employés à la production à la place du nombre d'employés à la production, augmentant ce dernier de plusieurs milliers. On relève de grossières divergences du même ordre dans les réponses aux questions portant sur la rémunération et sur les heures de travail, notamment le cas de répondants qui indiquent des cents au lieu d'arrondir au plus proche dollar.

### 5.2 Données enregistrées mécanographiquement

Le système IDTC permet d'évaluer les problèmes qu'éprouvent les répondants à utiliser ce mode de collecte. Il peut enregistrer le nombre de fois que les répondants réintroduisent leurs données, ainsi que le nombre de fois qu'une question est relue une deuxième ou une troisième fois avant d'obtenir une réponse. Il peut aussi conserver des renseignements sur les répondants qui raccrochent avant d'introduire des données.

Les données enregistrées mécanographiquement ont été recueillies pour un échantillon total de 1,203 observations durant une période de trois mois, en 1990. Les données ont été recueillies auprès d'environ 474 répondants uniques, dont un bon nombre ont fourni des données pendant

deux ou trois mois. Comme les données enregistrées mécanographiquement variaient peu d'un mois à l'autre, toutes les données sont présentées pour les trois mois combinés.

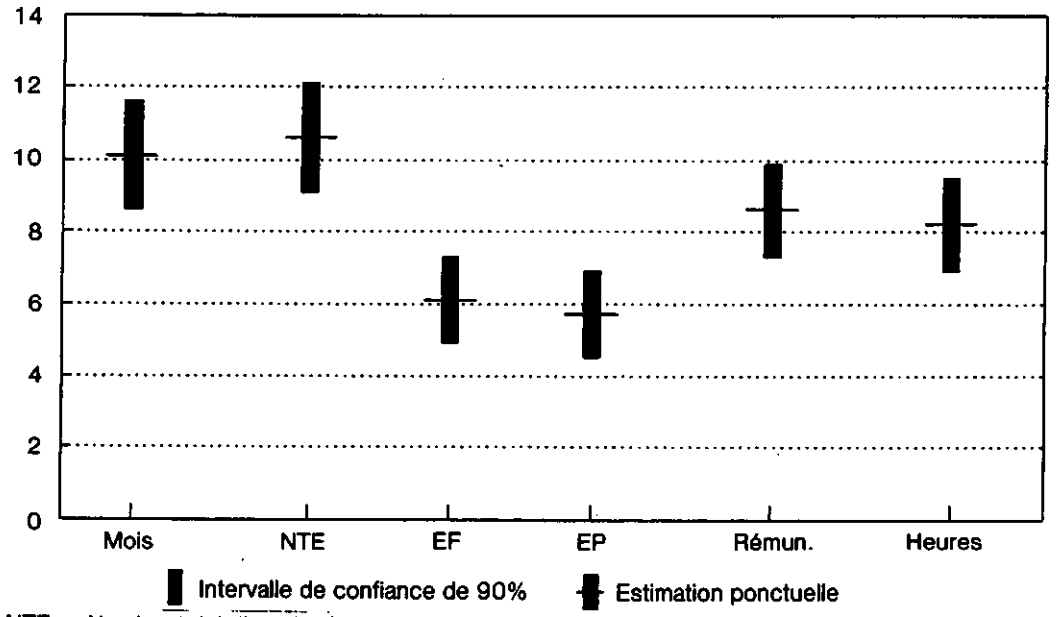
La figure 1 indique, pour chaque question, le pourcentage d'appels pour lesquels la question a été posée au répondant plus d'une fois. La question pouvait être répétée une deuxième fois si le répondant ne répondait pas dans les deux secondes (répétition), ou s'il ne confirmait pas sa réponse en appuyant sur le "1" après que celle-ci lui eut été relue (réintroduction). La figure indique que les taux de répétitions et de réintroductions sont plus élevés dans le cas des questions portant sur le mois, sur le nombre total d'employés, sur la rémunération et sur les heures de travail. Il est possible que les taux plus élevés enregistrés pour les deux premières questions (plus de 10%) soient attribuables au fait que les répondants aient eu besoin de quelques questions pour se familiariser avec le système. Les questions sur la rémunération et sur les heures de travail nécessitant en général l'introduction d'un plus grand nombre de chiffres, nous supposons que les risques d'erreurs d'introduction des données sont plus élevés, ce qui explique pourquoi la question doit être répétée et la réponse réintroduite.

La figure 2 donne séparément les pourcentages de questions répétées faute de réponse (répétitions) et à défaut de confirmation (réintroductions) pour les deux premières questions ("mois" et "nombre total d'employés"). Il n'a pas été possible d'obtenir des données distinctes pour les autres éléments d'information. Lorsqu'ils utilisent le système IDTC, les répondants à la CESS sont tenus d'introduire au moins leur numéro d'identification de déclarant, le mois et le nombre total d'employés. Le système accepte une non-réponse aux autres questions. Le fait que les répondants doivent au moins indiquer le mois et le nombre total d'employés explique pourquoi nous avons pu obtenir des données distinctes sur les répétitions et sur les réintroductions dans le cas de ces deux questions. Presque tous les problèmes que posait la question relative au mois étaient dus à des répétitions, c'est-à-dire au fait que deux secondes se soient écoulées sans qu'une réponse soit donnée. Par ailleurs, on comptait parmi les problèmes posés par la question relative au nombre total d'employés presque autant de répétitions faute de réponse que de réintroductions à défaut de confirmation de la réponse précédente.

Seulement 2% des appels reçus par le système IDTC chaque mois comprenaient uniquement un numéro d'identification du déclarant. Lors de ces appels, les répondants avaient tout simplement raccroché ou avaient pu voir leur communication avec le système coupée.

En outre, le système IDTC enregistre tous les appels reçus qui comprennent au moins le numéro d'identification du déclarant, le mois et le nombre total d'employés. À partir des données de vérification des enregistrements dont il a été question précédemment, nous avons pu repérer les répondants qui ont effectué plus d'un appel durant un mois et codé les raisons des rappels. En tout, environ 4% des répondants ont appelé le système plus d'une fois au cours d'un mois donné. La plupart de ces répondants ont fourni des éléments d'information qu'ils n'avaient pas communiqués lors de l'appel initial (2%). Une proportion additionnelle de 1% des répondants ont corrigé certains éléments d'information en plus d'en fournir de nouveaux. Bon nombre de ces répondants ont semblé avoir éprouvé des problèmes à introduire les données la première fois. Une dernière proportion de 1% des répondants n'ont rappelé que pour apporter des corrections aux éléments d'information fournis antérieurement ou pour transmettre des données identiques. Ces appels étaient souvent effectués plusieurs jours plus tard, ce qui veut peut-être dire que les répondants avaient obtenu de nouvelles données de leurs dossiers. Pour ce qui est des données identiques, les répondants ne se souvenaient peut-être pas s'ils avaient déjà déclaré leurs données ou non. Actuellement, le système accepte les données transmises à la date et à l'heure les plus récentes, même si les analystes obtiennent la liste des répondants qui ont rappelé à des fins de révision et, au besoin, de correction.

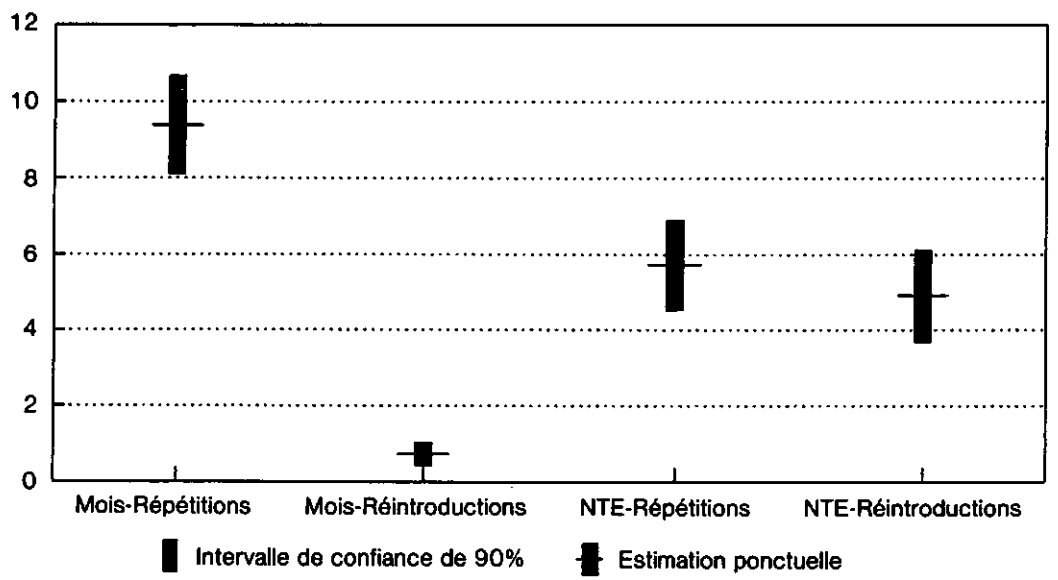
Il semble que la nécessité d'introduire un "1" pour confirmer chaque élément d'information introduit ait constitué une fréquente raison de rappel. Avant de rappeler, bon nombre des répondants qui ont corrigé leurs données avaient introduit dans la zone de données un "1" suivi d'une autre réponse. Les répondants qui utilisaient le système IDTC pour la première fois ont rappelé deux fois plus souvent que les utilisateurs "expérimentés".




NTE = Nombre total d'employés
EF = Nombre d'employés de sexe féminin
EP = Nombre d'employés à la production

(N = 1,203 par question)

**Figure 1.** Données enregistrées mécanographiquement - Pourcentage de questions répétées et de réintroduction des données




NTE = Nombre total d'employés
Répétitions = Aucune réponse après 2 secondes
Réintroductions = Correction des réponses

(N = 1,203 par question)

**Figure 2.** Données enregistrées mécanographiquement - Pourcentage de questions répétées par rapport aux réintroductions de données pour les deux premiers éléments d'information

Quelques répondants ont appelé trois fois pour introduire leurs données pour un mois donné, et un répondant a même appelé quatre fois. Ces répondants semblaient éprouver des difficultés avec le système, mais ils ont quand même réussi à déclarer toutes leurs données correctement.

En général, ces données laissent supposer que les répondants éprouvent certaines difficultés à utiliser le système (plus qu'ils ne l'admettent durant l'interview d'évaluation). Certaines mesures pourraient être prises pour aider à atténuer une partie de ces problèmes. On pourrait, entre autres, donner plus de temps pour répondre, fournir de meilleures instructions et essayer d'améliorer la méthode de confirmation des données introduites. En outre, il se peut qu'il suffise, pour résoudre quelques-uns des problèmes, de donner aux répondants la possibilité de revenir à une question.

### 5.3 Enquête d'évaluation auprès des répondants

Au cours de l'année 1990, les intervieweurs du BLS ont mené une enquête d'évaluation téléphonique auprès des répondants qui ont utilisé le système IDTC. Compte tenu des ouvrages sur l'ergonomie que nous avons étudiés précédemment, certaines des questions portaient sur la compréhension et le débit de la voix numérisée. Comme les données enregistrées mécanographiquement ont révélé un nombre substantiel de répétitions et de réintroductions, certaines questions ont aussi été élaborées afin d'étudier ce problème. En outre, on a demandé aux répondants d'évaluer le système IDTC et de répondre à des questions portant sur la conception du système.

Les résultats de l'enquête d'évaluation sont présentés dans le tableau 4. Les répondants n'ont eu que peu de difficulté à comprendre la voix numérisée. Environ 97% d'entre eux ont déclaré que la voix était très compréhensible, et ils ont tous indiqué qu'il était facile de comprendre les chiffres à mesure qu'ils leur étaient relus pour fin de confirmation. Au cours des deux premiers mois de l'enquête, nous avons interrogé les répondants au sujet du débit de l'interview. La plupart des répondants ont indiqué que le débit était à peu près correct (88%), bien que 10% d'entre eux aient estimé qu'il était trop lent. Tout au long de l'étude, nous avons supposé que la compréhension de la voix constituait un problème moins important que les difficultés à exécuter la tâche. Bien qu'il ait été difficile de séparer les deux types de problèmes dans les données de vérification des enregistrements, les interviews d'évaluation sont venues confirmer cette hypothèse.

**Tableau 4**

Résultats de l'enquête d'évaluation*

| | |
|---|---|
| Voix compréhensible | 97% |
| Chiffres relus faciles à comprendre | 100% |
| Débit assez bon | 88% |
| Jamais réintroduit de chiffres | 60% |
| Jamais fait répété de questions | 83% |
| Jamais eu une mauvaise communication téléphonique | 93% |
| Utilisé la fonction d'accélération | 63% |
| Instructions adéquates | 98% |
| Expérience avec l'IDTC très favorable | 93% |

* N = 411, à l'exception des rubriques relatives au débit et aux questions répétées. Environ 177 répondants ont été interrogés au sujet du débit et 209 concernant la répétition des questions.

Quant à la difficulté d'exécution de la tâche, 60% des répondants ont indiqué qu'ils n'avaient jamais eu à réintroduire de chiffres, tandis que la plupart des autres ont fait savoir qu'ils avaient parfois dû le faire. Lorsqu'on leur a demandé les raisons pour lesquelles ils devaient réintroduire des chiffres, une majorité ont répondu qu'ils avaient accidentellement appuyé sur une mauvaise touche. D'autres ont indiqué qu'ils n'avaient pas eu assez de temps, qu'ils avaient été distraits ou qu'ils avaient inscrit leurs chiffres trop rapidement. Dans les derniers mois de l'enquête d'évaluation, nous avons interrogé les répondants concernant la répétition des questions (sans qu'il n'y ait eu réintroduction des données). Environ 83% d'entre eux ont indiqué qu'ils n'avaient jamais jugé nécessaire de faire répéter les questions. La majorité des 17% de répondants qui ont fait répéter les questions ont déclaré qu'ils avaient été distraits, tandis que d'autres ont mentionné qu'ils avaient manqué de temps.

La plupart des répondants ont éprouvé peu de problèmes avec le système de télécommunications, puisque 93% d'entre eux ont indiqué que la communication téléphonique avait été bonne lorsqu'ils avaient utilisé le système IDTC. La plupart des répondants qui ont déclaré avoir éprouvé des problèmes de communication ont mentionné que cela ne s'était produit qu'une seule fois. Un grand nombre des répondants (63%) ont utilisé le symbole de la livre, fonction du système conçue pour accélérer la déclaration des données.

Presque tous les répondants ont indiqué que les instructions qu'ils avaient reçues lorsqu'ils ont commencé à utiliser le système IDTC étaient à-propos. En général, les répondants ont semblé satisfaits du système, environ 93% d'entre eux ayant évalué leur expérience avec le système IDTC comme étant très favorable.

## 6. ANALYSE

Selon les données recueillies, le mode de collecte par IDTC pose peu de problèmes graves. Les données de vérification des enregistrements indiquent l'existence d'une certaine erreur due à la non-réponse aux questions, erreur surtout associée aux répondants qui utilisent le système pour la première fois. L'inscription de chiffres additionnels ou inexacts semble être le plus grave problème qui touche les éléments d'information. Toutefois, dans une enquête par panel, les vérifications de contrôle longitudinales pourraient permettre de réduire le nombre de ces erreurs, comme pourraient le permettre les vérifications de contrôle logiques dans toutes les enquêtes. En outre, il est essentiel d'aborder la question de l'arrondissement des données dans les instructions transmises aux répondants. Tant les données de vérification des enregistrements que les données enregistrées mécanographiquement indiquent que les séries de chiffres plus longues posent plus de problèmes, probablement autant à l'étape de l'introduction des données qu'à celle de la vérification des données inexactes. Les répondants éprouvent peut-être de la difficulté à se rappeler des longues séries de chiffres durant la vérification, puisqu'ils ne semblent pas avoir de problèmes à comprendre les chiffres qui leur sont relus pour fin de confirmation.

D'après les données de vérification des enregistrements, il se peut que les répondants aient introduit le numéro de confirmation du mois dans la zone relative à la question sur le nombre total d'employés. En outre, les données enregistrées mécanographiquement indiquent qu'il arrive souvent que les répondants n'inscrivent pas le mois la première fois que celui-ci est demandé. Comme les répondants semblent se servir de leurs formules d'enquête lorsqu'ils introduisent des données, il est probable qu'ils aient besoin d'un peu plus de temps pour passer du numéro d'identification au haut de la formule, au mois et aux éléments d'information qui se trouvent plus bas. On pourrait résoudre ce problème en mettant tous les renseignements devant être introduits à un seul endroit sur la formule. Ainsi, on pourrait diminuer le nombre de fois que la question du "mois" doit être répétée et, potentiellement, réaliser des économies en réduisant la durée des appels. On pourrait réduire le nombre de fois que les autres questions doivent être répétées en donnant aux participants à l'enquête plus de temps pour répondre,

puisqu'ils ont déclaré qu'ils étaient dérangés lorsqu'ils devaient répondre à l'enquête. Toutefois, étant donné que la plupart des répondants jugent le débit assez bon, et que bon nombre d'entre eux se servent de la fonction d'accélération, le fait de donner plus de temps pour répondre pourrait être une source de frustration. On ne peut probablement pas faire grand chose pour réduire le nombre de réintroductions, puisque les répondants indiquent qu'ils avaient appuyé sur la mauvaise touche et qu'ils devaient corriger les données.

Les données indiquent que le nombre d'erreurs diminue à mesure que les répondants s'habituent au système. Les enquêtes par panel sont donc peut-être celles qui se prêtent le mieux à ce mode de collecte des données. Par ailleurs, nous croyons que le système IDTC présente beaucoup de promesses pour les enquêtes auxquelles il faut répondre par oui ou non ou par des chiffres. Les erreurs relevées ne sont pas très graves, et les répondants ont évalué leur expérience de l'utilisation du système comme étant très favorable. La méthode IDTC peut être particulièrement attirante pour les entreprises participant à des enquêtes, car elle leur permet d'appeler au moment qui leur convient, plutôt que d'être dérangées par des appels téléphoniques leur demandant de répondre à des questions. Toutefois, pour certaines enquêtes, l'auto-initiation et l'absence de contacts humains peuvent constituer des problèmes qui contribueraient à l'erreur due à la non-réponse.

Même si les répondants ont réservé un accueil très favorable à la collecte des données par IDTC, certaines mesures peuvent être prises afin d'améliorer le système, notamment:

- accorder suffisamment de temps aux répondants pour introduire leurs données au clavier, surtout dans le cas des premières questions et de celles nécessitant l'inscription d'une longue série de chiffres;
- chercher des moyens d'améliorer le mode de confirmation des réponses; enfin,
- prévoir des vérifications de contrôle longitudinales pour détecter les cas de déclaration de cents et d'autres grossières erreurs; on pourrait intégrer ces contrôles au système IDTC en posant aux répondants des questions appropriées afin de corriger ou de confirmer leurs réponses.

Le BLS a eu recours à la méthode IDTC dans une autre enquête, consistant en un suivi auprès d'un petit échantillon d'entreprises ayant participé, en 1988, à une enquête sur les programmes d'aide en matière de drogues mis en oeuvre par les employeurs. Réalisée en 1990, cette enquête de suivi avait pour objet de déterminer si le pourcentage d'entreprises qui offrent à leurs employés un programme d'aide en matière de drogues avait beaucoup changé au cours des deux années précédentes. Ces entreprises ont reçu par la poste un court questionnaire auquel elles devaient répondre par des données numériques ou par oui ou non, et on les a encouragées à déclarer leurs réponses au moyen d'un téléphone à clavier. À la fin des premières semaines de l'enquête, près de 20% des entreprises avaient répondu à l'aide du système IDTC, et une proportion égale par la poste. Le système IDTC n'a pas été utilisé après le début des activités de suivi des cas de non-réponse, soit environ deux semaines après l'envoi initial des questionnaires par la poste. Les autres données ont été recueillies par téléphone (méthode ITAO).

Compte tenu du temps que l'IDTC permet de sauver et des économies qu'elle permet de réaliser au titre des activités de perforation, nous croyons que la collecte des données au moyen d'un téléphone à clavier pourrait être utilisée avec profit pour d'autres enquêtes où la collecte des données et assujettie à des contraintes de temps. Il est possible que l'utilisation de ce mode de collecte communique au participant à l'enquête l'importance qu'il y a à respecter les délais de réponse. Nous avons démontré dans le présent article qu'on peut réduire les erreurs de mesure en utilisant le mode de collecte des données par IDTC.

Comme la collecte par IDTC permet de réduire les délais de réponses et les coûts de collecte, nous nous attendons à ce qu'elle soit beaucoup plus utilisée dans l'avenir. Nous savons que deux projets sont en cours afin de mettre à l'essai la reconnaissance des données introduites au moyen d'un téléphone à clavier dans le cadre d'une enquête. Statistique Canada procède actuellement à l'essai d'un système de collecte des données par IDTC pour l'enquête sur

l'emploi, la rémunération et les heures de travail, tandis que les laboratoires Bell procèdent actuellement à la mise au point d'un système IDTC pour une enquête auprès des clients de la société AT & T (Wendler 1990).

Le BLS utilise également la technologie de la reconnaissance de la parole à titre expérimental aux fins de la collecte des données de la CESS (voir Winter et Clayton 1990). Bien que les postes téléphoniques à clavier soient de plus en plus accessibles, nous estimons qu'entre 10% et 20% de nos répondants ne disposent pas de tels appareils. Dès que la technologie de la reconnaissance de la parole multilocuteur aura atteint un niveau acceptable pour les dix chiffres dont on a besoin pour déclarer les données de la CESS, nous nous attendons à ce que les utilisateurs préfèrent l'employer à la place du système de collecte par IDTC. Il convient donc d'entreprendre d'autres travaux de recherche sur les erreurs de mesure associées à cette technologie.

## REMERCIEMENTS

Les auteurs remercient Darrell Philpot et Henry Chiang de leur aide dans cette étude.

## BIBLIOGRAPHIE

CLAYTON, R., et HARRELL, L.J., Jr. (1990). Developing a cost model for alternative data collection methods: Mail, CATI and TDE. Présenté à l'Annual Meeting of the American Statistical Association, Anaheim, California.

COX, A.C., et COOPE, M.B. (1981). Selecting a voice for a specified task: The example of telephone announcements. *Language and Speech*, 24, 233-243.

MARICS, M.A., et WILLIGES, B.H. (1988). The intelligibility of synthesized speech in data inquiry systems. *Human Factors*, 30, 719-732.

PONIKOWSKI, C.H., COPELAND, K.R., et MEILY, S.A. (1989). Applications for touch-tone recognition technology in establishment surveys. Présenté à l'American Statistical Association Winter Conference, San Diego, California.

SCHWAB, E.C., NUSBAUM, H.C., et PISONI, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, 27, 395-408.

WATERWORTH, J.A. (1984). Interaction with machines by voice: A telecommunications perspective. *Behaviour and Information Technology*, 3, 163-177.

WENDLER, E.R. (1990). Respondent-initiated computer-directed surveys. Présenté à l'Annual Conference of the American Association of Public Opinion Research, Lancaster, Pennsylvania.

WERKING, 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., TUPEK, A., et CLAYTON, R. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics*, 4, 349-362.

WINTER, D.L.S., et CLAYTON, R.L. (1990). Speech data entry: Results of the first test of voice recognition for data collection. Bureau of Labor Statistics, Washington, D.C.

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 31–46
Statistique Canada



# Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire

J. MICHAEL BRICK et JOSEPH WAKSBERG[1]

RÉSUMÉ

La technique de Mitofsky et Waksberg est une méthode efficace pour prélever par composition aléatoire un échantillon autopondéré de ménages. Il s'agit d'une méthode progressive qui exige le prélèvement d'un nombre constant de ménages dans chaque grappe. Dans le présent article, les auteurs décrivent une version modifiée de la méthode de Mitofsky et Waksberg qui n'est ni autopondérée, ni progressive. Après avoir analysé le biais et la variance des estimations obtenues à l'aide de la méthode modifiée, ils indiquent dans quelles circonstances il peut être avantageux d'utiliser la version modifiée plutôt que la version originale de la méthode de Mitofsky et Waksberg.

MOTS CLÉS: Composition aléatoire; échantillonnage téléphonique; échantillonnage par grappes; troncation.

## 1. INTRODUCTION

La méthode élaborée par Mitofsky et Waksberg pour tirer des échantillons de ménages par composition aléatoire (Waksberg 1978) est souvent utilisée pour prélever un échantillon dans les enquêtes téléphoniques. Comme Waksberg le dit dans son article, il s'agit d'une méthode efficace de production d'un échantillon autopondéré, c'est-à-dire un échantillon dans lequel tous les ménages ayant le téléphone ont la même probabilité de tirage (à l'exception des ménages ayant plus d'un numéro de téléphone). L'efficacité de la méthode de Mitofsky et Waksberg est attribuable au fait qu'elle permet de réduire de façon marquée la proportion de numéros de téléphone non-résidentiels devant être composés avant qu'il soit possible de joindre les ménages-échantillon.

La méthode de Mitofsky et Waksberg suit un plan de sondage à deux degrés. Au premier degré, on prélève un échantillon de grappes consistant chacune en un bloc de 100 numéros de téléphone ou en un multiple de tels blocs. Les grappes (ou blocs de 100 numéros de téléphone) sont d'abord tirées avec une probabilité égale. On choisit au hasard un numéro de téléphone dans chaque grappe et on le compose. S'il s'agit du numéro d'un ménage, on retient la grappe. Sinon, on la rejette. Au deuxième degré, on prélève un échantillon de ménages à l'intérieur des grappes-échantillon que l'on a retenues. Pour s'assurer d'obtenir un échantillon autopondéré, il faut prélever un nombre constant de ménages par grappe. Certaines organisations (dont Westat Inc.) vont en général un peu plus loin et précisent un nombre constant de ménages interviewés par grappe (ou de ménages soumis à une sélection préliminaire, si la première partie du processus de collecte des données consiste en une présélection). On estime que le fait de substituer un autre ménage tiré au hasard à l'intérieur de la même grappe à chaque ménage non répondant constitue une façon raisonnable de réduire le biais dû à la non-réponse.

Ce système présente un inconvénient sur le plan opérationnel. Il faut parfois recomposer un numéro un assez grand nombre de fois pour déterminer s'il s'agit d'un numéro résidentiel ou non, surtout dans le cas des numéros auxquels on n'obtient pas de réponse. Le nombre de rappels nécessaires pour déterminer quels ménages acceptent de collaborer est encore plus élevé. On doit procéder ainsi pour chaque échantillon initial afin de déterminer dans quelles grappes il faut ajouter des numéros de téléphone pour obtenir la taille voulue et quel nombre de numéros on doit ajouter. De fait, il faut utiliser dans chaque grappe un échantillonnage progressif selon lequel on doit régler tous les cas précédents avant de savoir s'il est nécessaire d'augmenter la

[1] J. Michael Brick et Joseph Waksberg, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850, U.S.A.

taille de l'échantillon. Ce processus est particulièrement incommode lorsque l'on ne dispose que d'un temps limité pour la collecte des données.

Plusieurs auteurs ont tenté de modifier la méthode de Mitofsky et Waksberg afin d'atténuer ou d'éliminer le caractère progressif de l'échantillonnage. Potthoff (1987) a élaboré une généralisation de la technique de Mitofsky et Waksberg selon laquelle on choisit $c$ numéros de téléphone par grappe, plutôt qu'un seul selon la méthode de Mitofsky et Waksberg, pour déterminer si l'on doit retenir celle-ci ou non. On s'assure d'obtenir un échantillon autopondéré en soumettant les grappes dans lesquelles un seul des $c$ numéros de téléphone a été composé à un plan de sondage prévoyant le tirage d'un nombre fixe de ménages par grappe, et en soumettant les autres grappes à un plan prévoyant le tirage d'un nombre fixe de numéros de téléphone par grappe. Il n'est pas nécessaire de soumettre le dernier groupe de grappes à un échantillonnage progressif. Selon Potthoff, en pratique, la plupart des grappes tomberont dans la seconde catégorie, de sorte que le nombre d'opérations séquentielles nécessaire s'en trouvera nettement réduit, sans toutefois que ces dernières soient complètement éliminées.

Lepkowski et Groves (1986) décrivent une méthode d'échantillonnage selon laquelle on tire des blocs de numéros de téléphone, dont un nombre assez important figurent dans les annuaires (et dans d'autres sources, s'il y a lieu), avec une probabilité de tirage proportionnelle au nombre de numéros inscrits. Les blocs de numéros qui ne contiennent aucun numéro de téléphone ou qui n'en contiennent que très peu sont échantillonnés à l'aide de la méthode de Mitofsky et Waksberg. Sudman (1973) avait précédemment proposé l'échantillonnage de blocs de numéros avec une probabilité de tirage proportionnelle au nombre de numéros inscrits dans les annuaires, mais il n'avait rien prévu en ce qui a trait aux blocs vides (qui pourraient contenir des numéros non inscrits). Drew et Jaworski (1986) décrivent une enquête téléphonique à composition aléatoire qui a été menée au Canada et dans laquelle on a utilisé des listes de numéros résidentiels (inscrits dans les annuaires ou non) obtenues à titre onéreux pour servir de mesures de l'effectif. Étant donné que les listes étaient considérées comme étant pratiquement complètes, il n'était pas nécessaire de prélever des blocs vides. Pour autant que nous le sachions, il est impossible d'obtenir des dénombrements pratiquement complets des numéros résidentiels aux États-Unis.

Ni le plan de sondage de Potthoff ni celui de Lepkowski et Graves n'éliminent entièrement la nécessité d'un échantillonnage progressif, bien que tous deux semblent réduire la partie de l'échantillon devant faire l'objet d'un tel échantillonnage. Ces deux méthodes présentent aussi d'autres inconvénients. La technique de Potthoff semble plutôt complexe: pour autant que nous le sachions, elle n'a pas été beaucoup utilisée pour les enquêtes téléphoniques à composition aléatoire. Dans le cas des enquêtes nationales, la technique de Lepkowski et Graves nécessite, aux fins de l'obtention de mesures de l'effectif, l'achat et le dépouillement d'une liste-annuaire englobant tous les États-Unis. Bien qu'on puisse se procurer de telles listes commerciales, elles sont coûteuses. En outre, selon un certain nombre de rapports récents, le pourcentage de l'ensemble des numéros résidentiels qui sont inscrits dans les annuaires n'est pas très élevé, et il diminue rapidement. Tucker (1989) fait état d'une analyse des numéros inscrits dans un groupe de comtés et de villes des États-Unis, analyse qui révèle des taux d'inscription variant de 48 à 62%. Selon Linda Piekarski (1990), si le nombre de numéros non inscrits continue d'augmenter au rythme actuel, il se pourrait que, "d'ici l'an 2000, la proportion de ménages dont le numéro de téléphone ne sera pas inscrit s'élève jusqu'à 62 %" (traduction). Ainsi, les mesures de l'effectif ne correspondent probablement que plus ou moins au nombre réel de ménages dans un bloc de numéros actifs.

Waksberg a proposé d'apporter à la technique de Mitofsky et Waksberg, une autre modification qui élimine complètement la nécessité de procéder à un échantillonnage progressif. La société Westat a eu recours à cette méthode dans un grand nombre d'études réalisées à l'aide de la composition aléatoire. Cummings (1979) avait déjà utilisé les mêmes techniques à la suite d'une erreur d'application de la méthode de Mitofsky et Waksberg. Ce dernier n'a toutefois pas reconnu que la méthode pouvait être utile pour éliminer l'échantillonnage progressif, ni examiné ses caractéristiques en vue de l'utiliser dans d'autres enquêtes. Nous allons maintenant décrire cette méthode et en exposer les propriétés mathématiques et statistiques.

## 2. MÉTHODE DE RECHANGE POUR ÉTABLIR LA TAILLE DES GRAPPES AVEC LA TECHNIQUE DE MITOFSKY ET WASKBERG

Comme on l'a indiqué précédemment, la technique de Mitofsky et Waksberg exige qu'on prélève un nombre constant de numéros résidentiels par grappe (ou blocs de numéros) pour s'assurer d'obtenir un échantillon autopondéré. La solution de rechange proposée consiste à prélever dans chaque grappe un échantillon ($K$) comportant un nombre constant de numéros de téléphone. Le premier degré d'échantillonnage reste inchangé et consiste toujours à prélever des grappes dont la probabilité de tirage est proportionnelle au nombre de ménages. Le nombre de numéros de téléphone par grappe étant constant, il est possible de désigner les numéros-échantillon à l'avance et d'éliminer ainsi le processus séquentiel. Il convient de noter qu'il est quand même nécessaire d'effectuer un suivi afin de déterminer quels numéros-échantillon sont des numéros résidentiels, et ce aux deux degrés d'échantillonnage. Toutefois, ce suivi ne s'applique qu'à un ensemble fixe de numéros de téléphone et ne nécessite pas la mise en oeuvre d'un processus séquentiel.

En revanche, la méthode de rechange ne permet pas d'obtenir un échantillon autopondéré. Comme, au premier degré, les grappes sont tirées avec une probabilité proportionnelle à la taille, la probabilité qu'une grappe soit sélectionnée est de $rN_i/100$, où $r$ est le taux de sondage utilisé pour le tirage des grappes, c'est-à-dire la fraction de sondage qui s'applique au premier degré, et $N_i$ est le nombre de numéros résidentiels compris dans la grappe $i$. Le poids devrait être proportionnel à $N_i^{-1}$, mais, comme $N_i$ n'est pas connu, on considère qu'il est proportionnel à $n_i^{-1}$, le nombre de ménages-échantillon dans la grappe.

Cette version modifiée de la méthode de Mitofsky et Waksberg présente des caractéristiques intéressantes du point de vue opérationnel. Elle est simple. Elle permet virtuellement de prélever l'échantillon d'avance sans qu'aucune opération de contrôle coûteuse ne soit nécessaire. Enfin, bien qu'elle exige qu'on effectue une pondération, on peut obtenir les poids directement à partir des données-échantillon et les produire mécaniquement sans contrôle systématique par des spécialistes.

Cette version présente cependant quelques graves inconvénients. Premièrement, on introduit un biais lorsqu'on estime $N_i^{-1}$ à l'aide de $K/100n_i$, où $K$ est le nombre de numéros de téléphone sélectionnés par grappe (un nombre constant dans toutes les grappes). Bien que ce biais soit relativement faible, il existe quand même. On ne peut ni l'éliminer ni l'atténuer en apportant des modifications mineures aux poids en utilisant, par exemple, $1/(n_i + t)$ au lieu de $n_i^{-1}$, où "$t$" désigne une constante fixe. Deuxièmement, l'introduction de poids variables augmente considérablement les variances des estimations. (L'augmentation n'est pas tant causée par les poids mêmes que par le fait qu'ils expriment des probabilités d'échantillonnage variables). Enfin, en employant la version modifiée, on perd l'une des caractéristiques utiles de la méthode de Mitofsky et Waksberg: la possibilité d'établir la taille exacte de l'échantillon. En effet, comme la méthode prévoit le prélèvement d'un nombre constant de ménages par grappe, il suffit de prélever un échantillon comportant le nombre approprié de grappes pour obtenir un échantillon ayant la taille désirée. Dans le cas de la version modifiée, la taille de l'échantillon devient une variable aléatoire qui, en règle générale, ne sera pas exactement égale à la valeur désirée. Bien que les écarts soient habituellement petits, il est utile de pouvoir obtenir exactement les tailles visées lorsque, aux termes de contrats ou d'engagements budgétaires, l'organisation responsable de l'enquête doit satisfaire à certaines exigences en la matière. Nous aborderons ces questions aux sections 3 et 4.

Avant de passer à l'étude des variances et des biais, il est utile d'examiner la distribution des grappes selon la taille pour l'ensemble des États-Unis. Les tableaux 1 à 3 présentent les estimations de ces distributions qui ont été calculées à partir des données recueillies lors de deux grandes enquêtes téléphoniques nationales à composition aléatoire réalisées aux États-Unis par la société Westat. Lors de ces deux enquêtes, on a employé la version modifiée de la méthode de Mitofsky et Waksberg qui a été décrite précédemment. L'échantillon de l'enquête sur laquelle porte le tableau 1, qui a été prélevé en 1986, se composait de 2,427 grappes (retenues après

**Tableau 1**

Nombre de ménages par grappe qui ont répondu à toutes les questions de présélection par grappe dans l'enquête de 1986
(Selon un échantillon de 2,427 grappes, comportant chacune 15 numéros de téléphone)

| Nombre d'interviews effectuées par grappe | Poids moyen de la grappe[1] | Répartition des ménages | | | Répartition des grappes | | |
|---|---|---|---|---|---|---|---|
| | | Fréquence | Pourcentage | Pourcentage cumulatif | Fréquence | Pourcentage | Pourcentage cumulatif |
| 0 | xx | 0 | 0 | 0.0 | 62 | 2.6 | 2.6 |
| 1 | 7.93[2] | 54 | 0 | 0.3 | 54 | 2.2 | 4.8 |
| 2 | 3.97[2] | 106 | 0.6 | 0.9 | 53 | 2.2 | 7.0 |
| 3 | 2.64 | 258 | 1.4 | 2.2 | 86 | 3.5 | 10.5 |
| 4 | 1.98 | 440 | 2.3 | 4.6 | 110 | 4.5 | 15.0 |
| 5 | 1.59 | 810 | 4.3 | 8.9 | 162 | 6.7 | 21.7 |
| 6 | 1.32 | 1,290 | 6.9 | 15.8 | 215 | 8.9 | 30.6 |
| 7 | 1.13 | 1,960 | 10.5 | 26.2 | 280 | 11.5 | 42.1 |
| 8 | 0.99 | 2,656 | 14.2 | 40.4 | 332 | 13.7 | 55.8 |
| 9 | 0.88 | 2,862 | 15.3 | 55.6 | 318 | 13.1 | 68.9 |
| 10 | 0.79 | 2,990 | 15.9 | 71.6 | 299 | 12.3 | 81.2 |
| 11 | 0.72 | 2,717 | 14.5 | 86.1 | 247 | 10.2 | 91.4 |
| 12 | 0.66 | 1,548 | 8.3 | 94.3 | 129 | 5.3 | 96.7 |
| 13 | 0.61 | 780 | 4.2 | 98.5 | 60 | 2.5 | 99.2 |
| 14 | 0.57 | 210 | 1 | 99.6 | 15 | 0.6 | 99.8 |
| 15 | 0.53 | 75 | 0 | 100.0 | 5 | 0.2 | 100.0 |
| Total | | 18,756 | 100.0 | xx | 2,427 | 100.0 | xx |
| Taille moyenne des grappes[3] | | | 7.93 | | | | |
| Effet du plan de sondage[4] | | | 1.31 | | | | |

[1] Le poids de la grappe correspond à la taille moyenne des grappes (c.-à-d. 7.93) divisée par le nombre de ménages qui ont répondu à toutes les questions de présélection dans la grappe $i$.
[2] La troncation des poids ramènerait ceux-ci à 3.
[3] La taille moyenne des grappes correspond à la taille moyenne des 2,365 grappes comprenant au moins un ménage qui a répondu à toutes les questions de présélection.
[4] L'effet du plan de sondage est réduit à 1.12 lorsque le poids maximal est de 3.

le premier degré d'échantillonnage), comportant chacune 15 numéros de téléphone, pour un total de 36,405 numéros. Au total, 18,756 ménages ont répondu à toutes les questions de présélection et on a enregistré 2,396 cas de refus, 1,727 cas de non-réponse pour d'autres raisons, et 13,526 cas où le numéro composé était un numéro non résidentiel ou hors service, où l'appel n'a pas obtenu de réponse et où il a été impossible de classer le cas dans une des catégories. L'analyse porte uniquement sur les 18,756 ménages qui ont répondu à toutes les questions. Les données qui figurent dans les tableaux 2 et 3 ont été recueillies auprès d'un échantillon, prélevé en 1989, qui se composait de 1,000 grappes comportant chacune 30 numéros de téléphone, pour un total de 30,000 numéros de téléphone, desquels 19,586 étaient des numéros résidentiels appartenant à des ménages ayant répondu à toutes les questions de présélection. Le tableau 2 montre la distribution des 15,030 ménages qui ont répondu à toutes les questions et le tableau 3, la distribution des 19,586 numéros résidentiels relevés dans les 1,000 grappes. Les poids attribués aux grappes sont exprimés par $\bar{n}/n_i$, $\bar{n}$ étant le nombre moyen de ménages par grappe. Il semble utile d'exprimer le poids sous cette forme, car ils indiquent les écarts observés par rapport à un échantillon autopondéré. Le seul effet du plan de sondage qui cause l'augmentation des variances est la variabilité des fractions d'échantillonnage. Les autres aspects du plan de sondage n'ont pas d'effet.

**Tableau 2**

Nombre de ménages par grappe qui ont répondu à toutes les questions de présélection dans l'enquête de 1989

(Selon un échantillon de 1,000 grappes, comportant chacune 30 numéros de téléphone)

| Nombre d'interviews effectuées par grappe | Poids moyen de la grappe[1] | Répartition des ménages | | | Répartition des grappes | | |
|---|---|---|---|---|---|---|---|
| | | Fréquence | Pourcentage | Pourcentage cumulatif | Fréquence | Pourcentage | Pourcentage cumulatif |
| 0 | xx | 0 | 0 | 0.0 | 8 | 0.8 | 0.8 |
| 1 ou | 7.57[2] | 6 | 0 | 0.0 | 3 | 0.3 | 1.1 |
| 3 ou 4 | 4.33[2] | 37 | 0.2 | 0.3 | 10 | 1.0 | 2.1 |
| 5 ou 6 | 2.75 | 126 | 0.8 | 1.1 | 22 | 2.2 | 4.3 |
| 7 ou 8 | 2.02 | 403 | 2.7 | 3.8 | 53 | 5.3 | 9.6 |
| 9 ou 10 | 1.59 | 688 | 4.6 | 8.4 | 72 | 7.2 | 16.8 |
| 11 ou 12 | 1.32 | 1,325 | 8.8 | 17.2 | 115 | 11.5 | 28.3 |
| 13 ou 14 | 1.12 | 1,987 | 13.2 | 30.4 | 147 | 14.7 | 43.0 |
| 15 ou 16 | 0.98 | 2,636 | 17.5 | 50.0 | 170 | 17.0 | 60.0 |
| 17 ou 18 | 0.85 | 2,692 | 17.9 | 65.9 | 154 | 15.4 | 75.4 |
| 19 ou 20 | 0.78 | 2,387 | 15.9 | 81.8 | 123 | 12.3 | 87.7 |
| 21 ou 22 | 0.70 | 1,673 | 11.1 | 92.9 | 78 | 7.8 | 95.5 |
| 23 ou 24 | 0.64 | 816 | 5.4 | 98.3 | 35 | 3.5 | 99.0 |
| 25 ou 26 | 0.55 | 254 | 1.7 | 100.0 | 10 | 1.0 | 100.0 |
| 27 ou 28 | xx | 0 | 0 | 100.0 | 0 | 0 | 100.0 |
| 29 ou 30 | xx | 0 | 0 | 100.0 | 0 | 0 | 100.0 |
| Total | xx | 15,030 | xx | xx | 1,000 | xx | xx |
| Taille moyenne des grappes[3] | | | | 15.11 | | | |
| Effet du plan de sondage[4] | | | | 1.33 | | | |

[1] Le poids de la grappe correspond à la taille moyenne des grappes (c.-à-d. 15.15) divisée par le nombre de ménages qui ont répondu à toutes les questions de présélection dans la grappe *i*.
[2] La troncation des poids ramènerait ceux-ci à 3.
[3] La taille moyenne des grappes correspond à la taille moyenne des 992 grappes comprenant au moins un ménage qui a répondu à toutes les questions de présélection.
[4] L'effet du plan de sondage est réduit à 1.12 lorsque le poids maximal est de 3.

Il convient de souligner que le tableau 1 a été compilé à partir d'un échantillon de 15 numéros de téléphone par grappe, tandis que les tableaux 2 et 3 sont fondés sur un échantillon de 30 numéros de téléphone par grappe. À l'évidence, les estimations du pourcentage de numéros résidentiels dans une grappe composée de 15 numéros de téléphone seront sujettes à une erreur d'échantillonnage plus élevée que les estimations établies à partir de 30 numéros de téléphone. Toutefois, le nombre de grappes utilisées dans le tableau 1 est plus de deux fois supérieur au nombre utilisé dans les tableaux 2 et 3, ce qui devrait largement compenser l'effet de l'écart entre la taille des grappes.

Il existe deux différences entre les tableaux 2 et 3. D'une part, le tableau 2 montre la distribution des ménages qui ont répondu à toutes les questions de présélection (comme dans le cas du tableau 1), tandis que les données du tableau 3 sont fondées sur tous les ménages-échantillon. Le fait que seules les données des ménages ayant répondu à toutes les questions soient utilisées dans le tableau 2 se traduit par une réduction du nombre estimatif moyen de ménages par grappe et modifie toute la distribution. En outre, ce fait accroît la variabilité des estimations de la distribution puisque les chiffres reflètent les erreurs d'échantillonnage découlant à la fois de la distribution des ménages par grappe et de celle des taux de non-réponse par grappe. D'autre part, le tableau 2 (tout comme le tableau 1) indique le nombre de ménages par grappe, tandis que le tableau 3 indique le pourcentage de numéros résidentiels par grappe. Il était pratique d'exprimer le tableau 3 sous cette forme aux fins des analyses dont il est fait état ultérieurement dans le présent article.

**Tableau 3**

Proportion de numéros résidentiels par grappe dans l'enquête de 1989
(Selon un échantillon de 1,000 grappes, comportant chacune 30 numéros de téléphone)

| Proportion de numéros résidentiels par grappe | Poids moyen de la grappe[1] | Répartition des ménages | | | Répartition des grappes | | |
|---|---|---|---|---|---|---|---|
| | | Fréquence | Pourcentage | Pourcentage cumulatif | Fréquence | Pourcentage | Pourcentage cumulatif |
| 0 | xx | 0 | 0.0 | 0.0 | 5 | 0.5 | 0.5 |
| .001 à .049 | 21.76[2] | 5 | 0.0 | 0.0 | 3 | 0.3 | 0.8 |
| .05 à .099 | 8.70[2] | 18 | 0.1 | 0.1 | 6 | 0.6 | 1.4 |
| .10 à .149 | 5.22[2] | 41 | 0.2 | 0.3 | 9 | 0.9 | 2.3 |
| .15 à .199 | 3.73[2] | 48 | 0.2 | 0.6 | 8 | 0.8 | 3.1 |
| .20 à .249 | 2.90 | 53 | 0.3 | 0.8 | 7 | 0.7 | 3.8 |
| .25 à .299 | 2.37 | 144 | 0.7 | 1.6 | 16 | 1.6 | 5.4 |
| .30 à .349 | 2.01 | 178 | 0.9 | 2.5 | 17 | 1.7 | 7.1 |
| .35 à .399 | 1.74 | 408 | 2.1 | 4.6 | 34 | 3.4 | 10.5 |
| .40 à .449 | 1.54 | 459 | 2.3 | 6.9 | 34 | 3.4 | 13.9 |
| .45 à .499 | 1.37 | 840 | 4.3 | 11.2 | 56 | 5.6 | 19.5 |
| .50 à .549 | 1.24 | 1,040 | 5.3 | 16.5 | 63 | 6.3 | 25.8 |
| .55 à .599 | 1.14 | 1,926 | 9.8 | 26.3 | 107 | 10.7 | 36.5 |
| .60 à .649 | 1.04 | 2,126 | 10.9 | 37.2 | 109 | 10.9 | 47.4 |
| .65 à .699 | 0.97 | 3,255 | 16.6 | 53.8 | 155 | 15.5 | 62.9 |
| .70 à .749 | 0.90 | 2,610 | 13.3 | 67.1 | 116 | 11.6 | 74.5 |
| .75 à .799 | 0.84 | 3,022 | 15.4 | 82.6 | 126 | 12.6 | 87.1 |
| .80 à .849 | 0.79 | 1,556 | 7.9 | 90.5 | 61 | 6.1 | 93.2 |
| .85 à .899 | 0.75 | 1,458 | 7.4 | 98.0 | 54 | 5.4 | 98.6 |
| .90 à .949 | 0.71 | 399 | 2.0 | 100.0 | 14 | 1.4 | 100.0 |
| .95 à .999 | xx | 0 | 0.0 | 100.0 | 0 | 0.0 | 100.0 |
| Total | xx | 19,586 | 100.0 | xx | 1000 | 100.0 | xx |
| Taille moyenne des grappes[3] | | | 19.68 | | | | |
| Effet du plan de sondage[4] | | | 1.28 | | | | |

[1] Le poids de la grappe correspond à la proportion moyenne de numéros résidentiels dans une grappe (c.-à-d. 0.653) divisée par la proportion de numéros résidentiels dans la grappe *i*.
[2] La troncation des poids ramènerait ceux-ci à 3.
[3] La taille moyenne des grappes correspond à la taille moyenne des 995 grappes comprenant au moins un numéro résidentiel.
[4] L'effet du plan de sondage est réduit à 1.12 lorsque le poids maximal est de 3.

Une autre particularité des pourcentages contenus dans les tableaux 1 à 3 est digne de mention. En effet, ces pourcentages reflètent la distribution selon la taille des grappes qui ont été sélectionnées dans l'échantillon, et non pas la distribution des grappes dans l'ensemble des États-Unis. L'utilisation d'un échantillonnage se faisant avec une probabilité proportionnelle à la taille entraîne un suréchantillonnage des grappes comportant une proportion élevée de numéros résidentiels et une sous-représentation des grappes ne comportant qu'un petit nombre de ces numéros. Il est possible de convertir une distribution qui représente l'échantillon en une distribution qui représente l'ensemble de la population en multipliant chaque pourcentage par les poids des grappes et en calculant la distribution en pourcentage des chiffres ainsi obtenus. Comme les poids sont exactement proportionnels à la réciproque du nombre de ménages par grappe qui ont répondu à toutes les questions, la conversion de la distribution des ménages en une distribution représentant l'ensemble de la population permet d'obtenir les pourcentages indiqués dans la distribution des grappes. La distribution des grappes dans l'échantillon est donc identique à celle des ménages dans la population.

Nous avons choisi de présenter tant les distributions de tous les ménages-échantillon que celles des ménages qui ont répondu à toutes les questions, parce qu'elles présentent toutes deux un intérêt pour les chercheurs. Les analyses dont il est fait état dans les sections 3 et 4 sont fondées sur les données du tableau 3.

## 3. RÉPERCUSSIONS SUR LA VARIANCE DE L'UTILISATION DE LA VERSION MODIFIÉ DE LA MÉTHODE DE MITOFSKY ET WASKSBERG

Dans la version originale de la méthode de Mitofsky et Waksberg, la variance d'une estimation fondée sur l'échantillon est fonction du nombre de ménages tirés par grappe ainsi que de l'homogénéité des ménages à l'intérieur des grappes et entre celles-ci. La variance pour un échantillon en grappes peut s'exprimer comme le produit de la variance pour un échantillon aléatoire simple par $[1 + \rho(\bar{n} - 1)]$, où $\rho$ est le coefficient de corrélation interne intra-classe et $\bar{n}$ est le nombre moyen de ménages par grappe. Étant donné que les grappes de numéros de téléphone sont souvent reliées à des régions géographiques et tendent à être assez homogènes, le prélèvement d'un grand nombre de ménages par grappe peut s'avérer inefficace.

Lorsqu'on utilise la version modifiée de la méthode de Mitofsky et Waksberg, on introduit une autre source de variance en permettant au nombre de ménages prélevés de varier d'une grappe à l'autre. Comme on l'a souligné à la section 2, le dénominateur de la fraction de sondage utilisée au deuxième degré ne s'annule pas avec le nombre de ménages dans la grappe (qui est proportionnel à la probabilité de sélection au premier degré) et les probabilités globales d'échantillonnage des ménages varient d'une grappe à l'autre.

En raison de la variabilité des taux globaux d'échantillonnage des ménages d'une grappe à l'autre, les variances des estimations sont plus grandes que lorsqu'on utilise la version originale de la méthode de Mitofsky et Waksberg, où la probabilité de tirage de chaque ménage est la même. Kish (1965) et Waksberg (1973) ont exposé des méthodes permettant d'évaluer l'augmentation de la variance d'une estimation causée par des probabilités de tirage inégales. On peut obtenir une approximation simple de la variance d'une estimation obtenue selon une méthode de pondération inégale (où les poids ne traduisent pas les fractions de sondage variables utilisées dans des strates choisies exprès pour réduire les variances d'échantillonnage) en multipliant la variance d'échantillonnage qu'on obtiendrait avec un échantillon autopondéré par un facteur d'inflation de la variance (FIV), défini par FIV $= \{1 +$ Varrel(poids)$\}$. Nous allons utiliser cette approximation ci-après pour examiner les répercussions sur la variance de l'emploi de la version modifiée de la méthode de Mitofsky et Waksberg.

La variance relative des poids a été calculée en deux étapes. Premièrement, la moyenne et la variance des poids ont été calculées étant donné un échantillonnage effectué à partir d'une distribution hypergéométrique tronquée (puisqu'il n'est pas possible d'obtenir un nombre nul de ménages si la grappe est échantillonnée au premier degré) définie par la densité de ménages dans la grappe et par la taille de l'échantillon prélevé dans la grappe. La moyenne et la variance inconditionnelles des poids ont ensuite été calculées pour l'ensemble de la distribution des ménages dans les grappes échantillonnées, qu'on trouve au tableau 3. La distribution des ménages dans l'échantillon est essentielle à l'évaluation du FIV.

Le poids naturel attribué à un ménage dans la version modifiée de la méthode de Mitofsky et Waksberg est proportionnel à $n_i^{-1}$, où $n_i$ est le nombre de ménages observés dans la grappe-échantillon $i$. Ce poids peut varier selon des facteurs s'échelonnant d'aussi peu que $1/K$ jusqu'à 1, où $K$ est le nombre de numéros de téléphone tirés dans une grappe. Le poids moyen est d'environ $1.5/K$, puisque près de 65% des numéros qu'on retrouve dans les grappes échantillonnées sont des numéros résidentiels.

Si le nombre de numéros de téléphone échantillonnés par grappe se situe entre 5 et 30, l'augmentation de la variance due à la pondération est alors d'environ 30%. Le FIV diminue légèrement à mesure que le nombre de numéros échantillonnés par grappe s'accroît au-dessus de 30, pour atteindre environ 17% lorsque tous les numéros de la grappe sont échantillonnés.

Le FIV ou la variance relative des poids est une fonction de la distribution du nombre de ménages entre les grappes et de la variabilité de l'échantillonnage aléatoire à l'intérieur des grappes. Nous avons rendu cette décomposition explicite en exprimant la variance des poids comme la somme de la moyenne de la variance conditionnelle des poids et de la variance de leur moyenne conditionnelle, où la variable de condition est la densité de ménages dans la grappe.

**Tableau 4**

Facteurs d'inflation de la variance (FIV) approximatifs pour les échantillons obtenus par composition aléatoire suivant la version modifiée de la méthode de Mitofsky et Waksberg

| Poids | Taille de l'échantillon prélevé dans chaque grappe ($K$) | | | | |
|---|---|---|---|---|---|
| | 5 | 10 | 30 | 60 | 100 |
| $1/n_i$ | 1.31 | 1.34 | 1.29 | 1.23 | 1.17 |
| $1/(n_i + .5)$ | 1.18 | 1.21 | 1.20 | 1.18 | 1.16 |
| $1/(n_i + 1)$ | 1.12 | 1.15 | 1.16 | 1.15 | 1.14 |
| $1/(n_i + 2)$ | 1.07 | 1.09 | 1.11 | 1.12 | 1.13 |

Lorsque la taille de l'échantillon prélevé dans chaque grappe est peu élevée, la moyenne de la variance conditionnelle est la composante dominante de la variance globale. À mesure que la taille de l'échantillon prélevé dans chaque grappe s'accroît, la variance de la moyenne conditionnelle devient plus dominante. C'est la raison pour laquelle la variance relative des poids, indiquée sur la première ligne du tableau 4, n'est pas une fonction monotone de la taille de l'échantillon prélevé dans chaque grappe.

**Variances obtenues en utilisant des poids différents**

Nous avons aussi examiné des poids autres que les poids inversement proportionnels au nombre de ménages afin de déterminer leurs effets sur l'erreur systématique et sur la variance des estimations. Bon nombre des poids de rechange qui ont été étudiés ont été obtenus à partir des transformations de stabilisation de la variance proposées pour les variables binomiales.

De toutes les solutions de rechange qui ont été examinées, les estimateurs comportant le plus petit biais et la plus petite variance étaient obtenus en apportant de simples corrections au poids naturel. En particulier, l'addition d'une petite constante au nombre observé de ménages (estimateurs de la forme $(n_i + t)^{-1}$, où $t$ est égal à .5, 1 ou 2) a permis de réduire les augmentations de la variance dues à la pondération différentielle. L'addition de la constante permet de réduire l'intervalle de variation des poids en tronquant les valeurs des poids les plus grands tout en ne modifiant que légèrement les poids des grappes où l'on trouve un plus grand nombre de ménages échantillonnés.

Le tableau 4 montre les FIV des estimateurs de la forme $(n_i + t)^{-1}$ dans le cas où des nombres différents de numéros de téléphone sont échantillonnés dans chaque grappe. Les données contenues dans ce tableau sont aussi fondées sur les distributions des ménages et des grappes figurant au tableau 3. Ce tableau démontre clairement qu'il est possible de réduire considérablement la variance due à une pondération inégale en utilisant un estimateur de la forme $(n_i + 1)^{-1}$ plutôt que l'estimateur naturel. Ceci est particulièrement vrai dans le cas des plans de sondage par composition aléatoire où l'on prélève 30 numéros de téléphone ou moins par grappe. En pareil cas, l'augmentation de la variance due à la pondération différentielle n'est que de 16% lorsqu'on utilise l'estimateur $(n_i + 1)^{-1}$, tandis qu'elle s'élève à 29% lorsqu'on utilise l'estimateur naturel.

**Variances obtenues avec des poids tronqués**

Un moyen souvent employé pour atténuer l'inflation de la variance associée à la variation des poids est la troncation des poids très élevés. Cette troncation, ou réduction des poids, est habituellement pratiquée à un poids au-dessus duquel on trouve relativement peu d'observations. Dans nombre d'échantillons prélevés au cours des enquêtes téléphoniques à composition

**Tableau 5**

Facteurs d'inflation de la variance (FIV) approximatifs pour les échantillons obtenus par composition aléatoire suivant la version modifiée de la méthode de Mitofsky et Waksberg, avec des poids tronqués*

| Poids | Taille de l'échantillon prélevé dans chaque grappe ($K$) | | | |
|---|---|---|---|---|
| | 10 | 30 | 60 | 100 |
| $1/n_i$ | 1.12 | 1.11 | 1.09 | 1.09 |
| $1/(n_i + .5)$ | 1.11 | 1.10 | 1.09 | 1.09 |
| $1/(n_i + 1)$ | 1.09 | 1.10 | 1.09 | 1.09 |
| $1/(n_i + 2)$ | 1.07 | 1.09 | 1.09 | 1.08 |

* Tous les poids ont été tronqués à une valeur trois fois supérieure au poids moyen, sauf dans le cas des échantillons composés de dix numéros, où ils ont été tronqués à une valeur deux fois supérieure au poids moyen.

aléatoire menées par la société Westat, on a procédé à une troncation des poids à une valeur qui était de deux à trois fois supérieure au poids moyen. Pour les fins de la présente recherche, nous avons examiné les poids qui ont été tronqués à une valeur environ trois fois supérieure au poids moyen. Dans le cas des échantillons de dix numéros par grappe, les poids ont été tronqués à une valeur deux fois supérieure au poids moyen afin d'éviter que le nombre d'observations touchées soit trop faible.

Le tableau 5 indique les FIV des estimateurs pour différents nombres de numéros échantillon par grappe, lorsque le seuil de troncation est établi à une valeur trois fois supérieure au poids moyen pour $n_i^{-1}$. On n'y donne pas les FIV pour les échantillons composés de cinq numéros par grappe, parce que le seuil de troncation dans les échantillons de cette taille est presque égal à l'unité, le poids le plus élevé possible.

Les valeurs figurant dans le tableau montrent que la troncation permet de réduire considérablement les effets défavorables de la pondération différentielle sur la variance des estimations. C'est pour l'estimateur naturel qu'on observe la réduction la plus marquée du FIV, ce dernier diminuant de plus de 50% par suite de la troncation. Bien que les FIV des autres estimateurs diminuent quelque peu, ces baisses sont moins frappantes étant donné que ces FIV étaient déjà moins élevés que celui de l'estimateur naturel. La troncation peut toutefois introduire des biais susceptibles de neutraliser l'avantage relatif à la réduction de la variance. Nous aborderons cette question dans la section 4.

### Variances obtenues par suite d'une augmentation de la taille de l'échantillon

Un troisième moyen de réduire la variabilité des poids consiste à augmenter la taille de l'échantillon. Les poids sont élevés lorsque le nombre de ménages prélevés dans la grappe est faible par rapport au nombre prévu de ménages par grappe. On peut faire diminuer la probabilité d'occurrence d'une telle situation en augmentant la taille de l'échantillon. Si le nombre de ménages dans une grappe est inférieur à une valeur déterminée (disons inférieur au tiers du nombre moyen de ménages par grappe), la taille de l'échantillon prélevé dans la grappe peut alors être doublée ou accrue dans une autre proportion.

On peut répéter cette opération afin de s'assurer que le nombre de ménages prélevés dans la grappe atteint une limite préétablie ou jusqu'à ce que tous les numéros contenus dans la grappe soient utilisés. Ce plan itératif a pour inconvénient évident de nécessiter le contrôle du rendement de l'échantillon par grappe et de constituer un échantillonnage progressif. Cette méthode a aussi pour inconvénient d'entraîner l'échantillonnage d'un plus grand nombre de numéros de téléphone dans les grappes ayant une plus faible densité de ménages (celles où l'on risque le plus de devoir augmenter la taille de l'échantillon), et donc de provoquer une baisse de productivité.

Malgré les lacunes opérationnelles que comporte l'augmentation de la taille de l'échantillon, nous l'avons quand même soumis à une examen limité. Comme elle n'a pas permis d'obtenir de meilleurs résultats que la troncation des poids, nous n'avons pas cru bon d'en faire une étude plus poussée.

## 4. EFFETS DE L'UTILISATION DE LA VERSION MODIFIÉE DE LA MÉTHODE DE MITOFSKY ET WAKSBERG SUR LE BIAIS

L'augmentation de la variance n'est qu'une des conséquences de l'utilisation de la version modifiée de la méthode d'échantillonnage de Mitofsky et Waksberg. Cette méthode a aussi pour conséquence d'introduire un biais dans les estimations résultantes. Si on prélève un nombre fixe de numéros dans chaque grappe sans apporter aucune correction aux poids, la variance des estimations n'augmente pas, mais le biais risque d'être très élevé.

Le poids non biaisé ($W_u$) pour la version modifiée de la méthode est défini par

$$W_u = \frac{100}{rN_i} \times \frac{100}{K},$$

où les termes correspondent aux termes définis plus haut. La difficulté tient au fait que $N_i$ est inconnue et ne s'annule pas avec le terme utilisé au deuxième degré, comme dans le cas de la version originale de la méthode de Mitofsky et Waksberg. Il faut donc avoir recours à la pondération pour tenter de réduire le biais.

Nous désignons l'estimateur qui utilise un poids de $n_i^{-1}$ comme l'estimateur naturel parce que, pour un échantillonnage à partir d'une distribution binomiale ou hypergéométrique, $n_i/K$ est un estimateur non biaisé de $N_i/100$. (Nous définissons le poids par $n_i^{-1}$, bien qu'il corresponde en fait à $K/100n_i$. Étant donné que $K/100$ est une constante, l'utilisation de $n_i^{-1}$ n'a aucun effet sur le rapport entre les poids.) Ce poids semble être l'estimateur naturel, malgré le fait que $n_i^{-1}$ constitue une estimation biaisée de $N_i^{-1}$, à moins que les 100 numéros contenus dans une grappe soient tirés. Le biais introduit par $n_i^{-1}$ a été étudié dans divers ouvrages; à cet égard, on peut se reporter à l'étude sur la stratification postérieure à l'échantillonnage dans Hansen, Hurwitz et Madow (1953). Il est peu probable qu'il existe un estimateur simple non biaisé, certainement pas de la forme $(n_i + t)^{-1}$, pour toutes les tailles d'échantillons prélevés dans les grappes.

Il est possible d'évaluer le biais susceptible d'être introduit en comparant la valeur prévue des estimateurs (le poids moyen calculé à l'aide d'estimateurs de la forme $(n_i + t)^{-1}$) avec le poids non biaisé, $W_u$. Étant donné que tant le poids non biaisé que la valeur prévue des estimateurs sont des fonctions de $N_i$, nous étudierons d'abord ces quantités étant donné $N_i$.

Le graphique de la figure 1 montre les courbes du poids non biaisé et des poids moyens obtenus en utilisant les estimateurs $n_i^{-1}$ et $(n_i + 1)^{-1}$, lorsque $K = 10$ numéros de téléphone sont sélectionnés dans chaque grappe. La fraction de sondage constante par grappe, $r$, a été omise de tous les poids. Étant donnée l'étendue de $W_u$, nous avons représenté la variation des poids moyens selon une échelle logarithmique.

Le graphique montre clairement qu'on enregistre les plus grands écarts entre $W_u$ et les poids moyens obtenus à l'aide des deux estimateurs lorsque $N_i/100$ est peu élevé. Dès que la densité de numéros résidentiels dépasse 20% (lorsqu'on utilise $(n_i + 1)^{-1}$) et 10% (lorsqu'on utilise $n_i^{-1}$), les écarts sont relativement mineurs. Bien que le graphique indique que le poids obtenu à l'aide de l'estimateur $(n_i + 1)^{-1}$ est toujours inférieur à $W_u$, il en ira autrement si l'on procède à une stratification a posteriori. Les poids stratifiés a posteriori ne sont pas illustrés dans le graphique parce que la stratification a posteriori a réellement effet sur les poids inconditionnels plutôt que sur les poids conditionnels illustrés ici. Nous aborderons la question du biais inconditionnel ci-après.

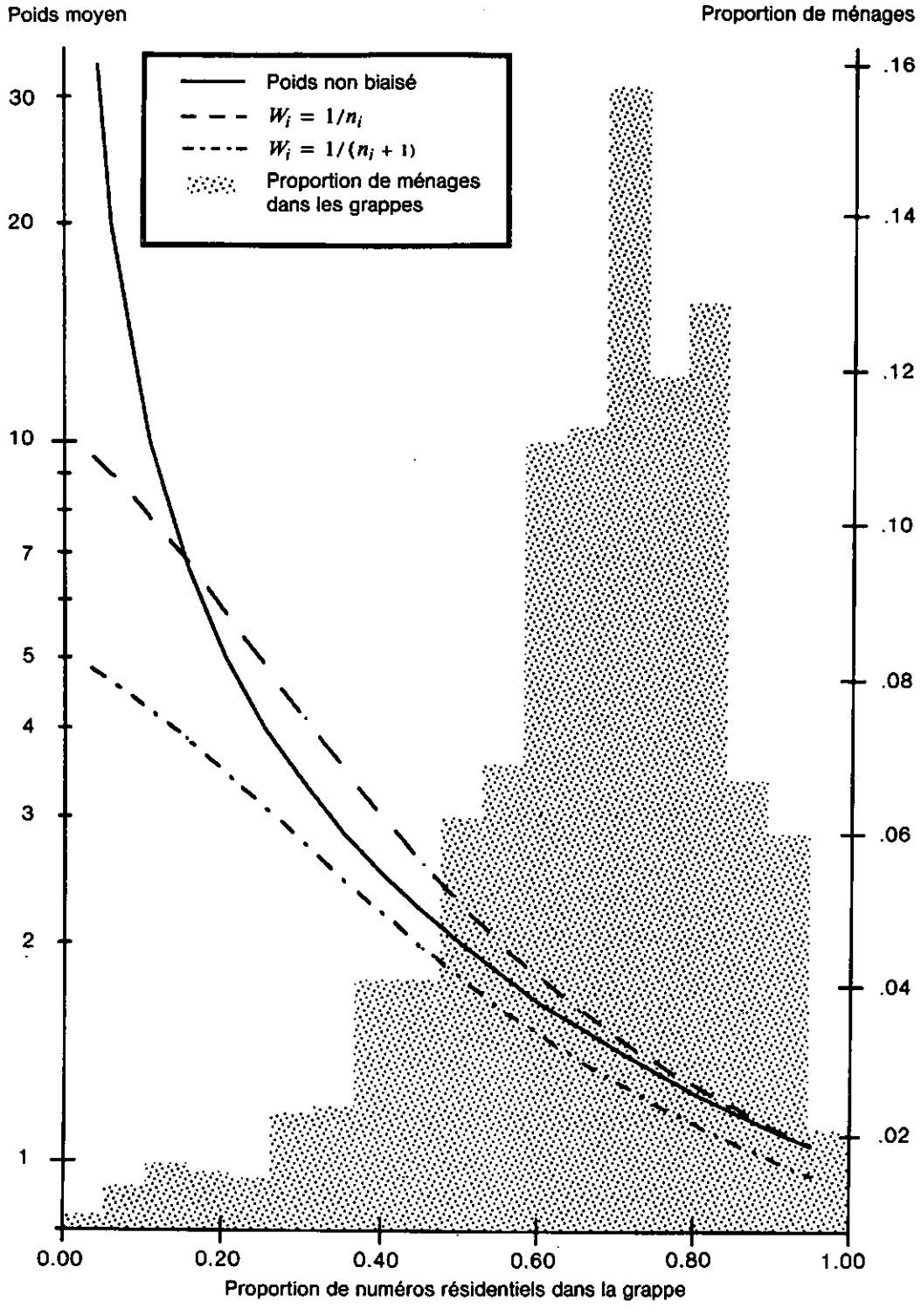



**Figure 1.** Poids moyens des estimateurs étant donné la proportion de numéros résidentiels, avec histogramme ombré de la proportion de ménages dans la grappe

L'histogramme ombré de la figure 1 illustre la distribution des ménages dont il est fait état au tableau 3. Nous l'avons superposé aux courbes des poids afin d'illustrer le fait que c'est dans les grappes qui ne contiennent qu'une très petite fraction des ménages échantillonnés que l'on observe les écarts les plus marqués entre les poids.

**Biais dans la taille de l'échantillon et dans les estimations**

Dans presque toutes les enquêtes téléphoniques à composition aléatoire, y compris celles dans lesquelles on utilise le plan de sondage de Mitofsky et Waksberg, on procède à une stratification a posteriori de l'échantillon selon les totaux connus de personnes ou de ménages. L'un des principaux objets de cette mesure consiste à rectifier les estimations en fonction de l'ensemble de la population plutôt que de les faire porter uniquement sur les ménages ayant le téléphone. Massey et Botman traitent de cette question et exposent d'autres avantages de l'emploi de la stratification a posteriori dans les enquêtes téléphoniques à composition aléatoire au chapitre 9 de Groves et coll. (1988).

Peu importent les raisons de l'utiliser, la stratification a posteriori permet d'obtenir des estimations qui sont égales aux totaux connus, quels que soient les poids appliqués aux ménages individuels. Comme ce biais, qu'on peut considérer comme un biais dans la taille de l'échantillon, est toujours nul, il est difficile de trouver une seule statistique qui mesure directement le biais inconditionnel. Nous allons tenter de surmonter cette difficulté en examinant de quelle façon le biais dans la taille de l'échantillon varie en fonction de la densité de ménages dans la grappe.

Nous avons évalué cette contribution au biais dans la taille de l'échantillon en suivant les étapes ci-après. D'abord, nous avons calculé les différents estimateurs ou fonctions de pondération à l'aide de la densité empirique de ménages indiquée au tableau 3. Ensuite, nous avons soumis les estimations à une stratification a posteriori afin de les rendre égales à l'unité et nous avons déterminé la contribution au total pour différentes valeurs de $N_i/100$. Enfin, nous avons défini le biais relatif dans la taille de l'échantillon comme étant la différence entre la contribution au total pour un estimateur donné et la contribution au total lorsqu'on utilise $W_u$ comme poids.

Cette mesure tient donc compte à la fois de la différence entre les poids pour des valeurs fixes de $N_i$ et de la distribution des ménages pour toutes les valeurs de $N_i$. Par conséquent, les ménages échantillonnés dans des grappes ayant des valeurs de $N_i$ qui sont rares ne contribueront pas beaucoup au biais relatif dans la taille de l'échantillon, même s'ils sont associés à de grands écarts entre les poids.

Pour illustrer ces calculs, la figure 2 montre le biais relatif dans la taille de l'échantillon pour certains estimateurs utilisés pour des échantillons de 30 numéros par grappe. L'un des estimateurs utilise le poids non corrigé, c'est-à-dire que le poids est constant pour tous les ménages, quel que soit le nombre de ménages relevés dans une grappe. Le biais relatif dans la taille de l'échantillon pour l'estimateur utilisant des poids non corrigés est beaucoup plus important que pour l'estimateur utilisant d'autres poids. Lorsqu'on utilise le poids non corrigé, les biais relatifs dans la taille de l'échantillon varient d'environ $-2\%$ à $+3\%$.

L'importance du biais entachant l'estimation d'une caractéristique est limitée par l'amplitude du biais dans la taille de l'échantillon. Autrement dit, le biais relatif entachant l'estimation ne peut pas être plus élevé que celui dans la taille de l'échantillon. Cette limite supérieure est atteinte seulement lorsqu'il existe une corrélation parfaite entre la variable et la densité de numéros résidentiels, ce qui est très rare. Il est peu probable qu'il existe des corrélations très élevées dans les échantillons tirés à l'échelle nationale, mais il y a plus de chances d'en retrouver dans les échantillons prélevés dans des régions géographiques restreintes.

On peut constater à la figure 2 que certaines tendances se dégagent du graphique des biais: ainsi, l'estimateur utilisant des poids non corrigés est uniformément trop bas dans les grappes à faible proportion de numéros résidentiels et trop élevé, dans les grappes où la proportion de ménages est grande. Lorsque les valeurs prises par les variables varient entre les grappes

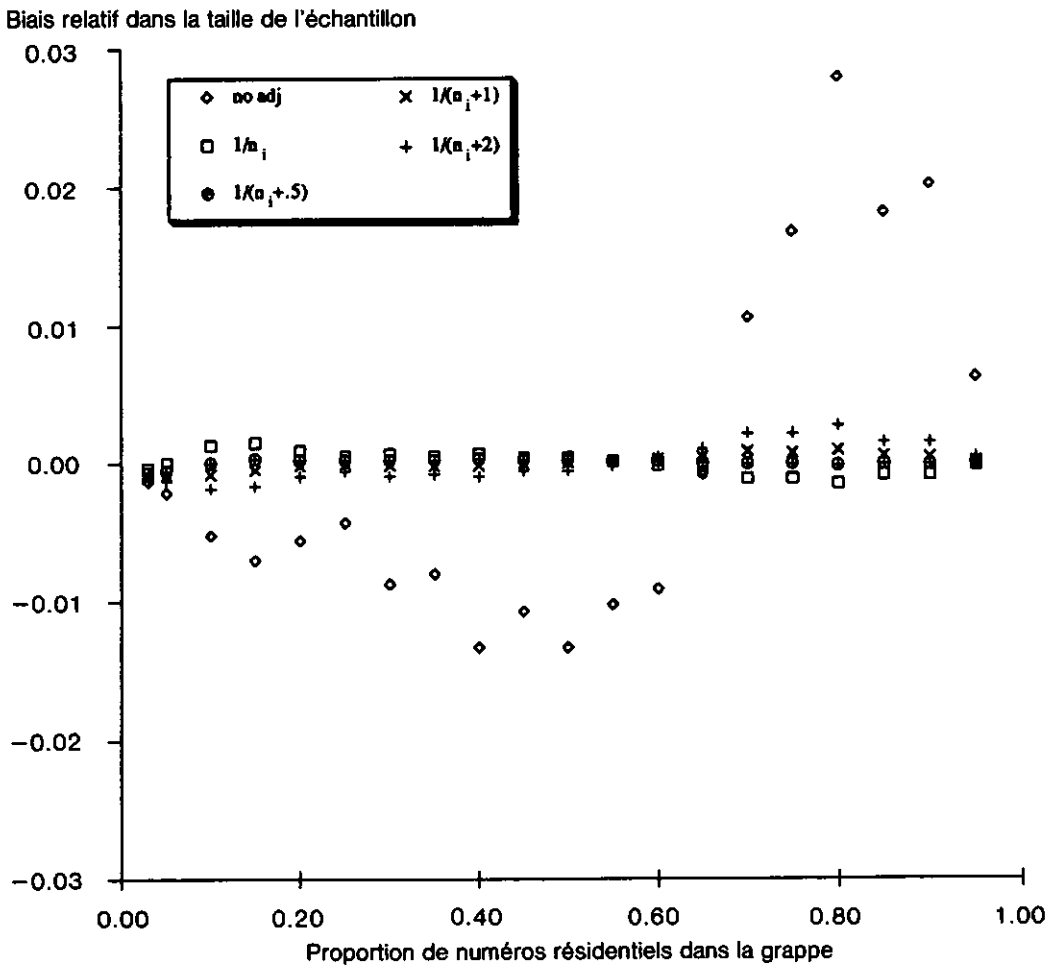



**Figure 2.** Biais relatif dans la taille de l'échantillon pour des échantillons de 30 numéros par grappe

à faible et à haute densité, les biais peuvent être assez élevés. Le tableau 1 de Cummings (1979) illustre, pour certaines variables, le biais introduit dans les estimations par suite de l'utilisation de poids non corrigés. Les biais ne sont pas très élevés, mais une pondération appropriée permettra de les éliminer.

En général, le rapport entre l'estimation et le nombre de ménages dans une grappe sera inconnu et incohérent pour toutes les variables à estimer. Par conséquent, il est raisonnable de choisir un estimateur pour lequel le biais relatif dans la taille de l'échantillon est peu élevé pour toutes les valeurs de $N_i$. Lorsque le biais relatif dans la taille de l'échantillon pour un ensemble d'estimateurs est peu élevé, le choix de ces derniers peut alors être dicté par la variance.

**Biais obtenus en utilisant des poids différents**

Nous avons calculé le biais relatif dans la taille de l'échantillon en utilisant des estimateurs différents pour des échantillons de 5, 10, 30 et 60 numéros de téléphone par grappe. Dans le cas des échantillons de 30 et 60 numéros, le biais est négligeable sauf lorsqu'on utilise les poids non corrigés. On pourrait employer pour les échantillons de cette taille n'importe quel des estimateurs utilisant des poids corrigés sans introduire de biais dans les estimations. Dans le cas

des échantillons de dix numéros par grappe, tous les estimateurs, à l'exception de celui utilisant un poids non corrigé, sont raisonnablement efficaces. Le biais obtenu en utilisant le poids $(n_i + .5)^{-1}$ est particulièrement encourageant.

Pour ce qui est de l'échantillon le plus petit étudié, soit celui de cinq numéros par grappe, les risques de biais sont un peu plus élevés. L'utilisation du poids naturel, $n_i^{-1}$, permet d'obtenir un biais dans la taille de l'échantillon un peu moins élevé que celle du poids $((n_i + .5)^{-1})$, mais la différence entre les deux biais n'est pas très grande. Le biais relatif dans la taille de l'échantillon qu'on obtient lorsqu'on utilise ces deux poids est toujours inférieur à 1%. Le biais est positif dans le cas de densités de numéros résidentiels qui se situent entre 45 et 80% et négatif pour les densités en-dehors de cet intervalle. Cette tendance ne saurait se révéler problématique que pour quelques variables dont la corrélation avec la densité de numéros résidentiels est très élevée.

**Biais obtenus en utilisant des poids tronqués**

La troncation des poids peut entraîner l'introduction de biais importants, selon le rapport entre les variables à estimer et les poids qui sont tronqués. Dans certains cas, le biais dû à la troncation peut limiter l'ampleur de celle-ci et, par conséquent, restreindre son utilité comme méthode de réduction de la variance.

Nous avons également calculé le biais relatif dans la taille de l'échantillon pour des échantillons de 10, 30 et 60 numéros par grappe, en utilisant des poids tronqués à une valeur environ trois fois supérieure au poids moyen. Dans le cas des échantillons de dix numéros par grappe, la troncation a été faite selon un facteur de 2 plutôt que de 3, comme nous l'avons vu précédemment.

La différence entre le biais relatif dans la taille de l'échantillon observé lorsqu'on utilise des poids tronqués et lorsqu'on se sert de poids non tronqués est sans importance pour toutes les tailles d'échantillons prélevés dans les grappes et pour la plupart des valeurs de $N_i/100$. La seule différence perceptible a été remarquée lorsque la densité de numéros résidentiels était inférieure à 10 à 15%. Les risques de biais sont légèrement plus élevés dans ces régions. Toutefois, le biais relatif dans la taille de l'échantillon observé lorsqu'on utilise les poids tronqués est encore bien inférieur à 1%, pour toutes les densités de numéros résidentiels.

## 5. CONCLUSION

La version originale de la méthode de Mitofsky et Waksberg constitue une méthode efficace pour obtenir par composition aléatoire un échantillon autopondéré de taille fixe. Toutefois, la nécessité de procéder à un contrôle séquentiel du nombre de ménages échantillonnés par grappe représente une caractéristique opérationnelle peu commode. En raison de ce contrôle, il est difficile de terminer la collecte des données dans des délais serrés. La période de collecte des données doit être suffisamment souple pour permettre l'obtention du nombre approprié d'unités dans chaque grappe. Le prolongement de la période de collecte des données et le contrôle du nombre d'unités échantillonnées se traduisent aussi par une augmentation des coûts. Étant donné la nécessité d'effectuer un contrôle fréquent du nombre d'unités échantillonnées, l'échantillonnage progressif peut également engendrer une certaine frustration chez les membres de l'équipe d'enquête en raison des complications relatives à la combinaison des opérations d'échantillonnage et de collecte des données.

La version modifiée de la méthode de Mitofsky et Waksberg élimine le caractère progressif du plan de sondage et, par le fait même, la nécessité de contrôler les travaux par grappe. Suivant cette méthode, on attribue un nombre fixe de numéros de téléphone à chaque grappe échantillonnée. Ainsi, on évite les dépenses afférentes au contrôle du nombre d'unités échantillonnées et à la mise en oeuvre d'une période de collecte des données prolongée. Toutefois, la version modifiée de la méthode de Mitofsky et Waksberg introduit de nouveaux éléments de biais et de variance dans les estimations. Il faut régler ces problèmes statistiques avant d'employer la méthode modifiée.

Il est difficile de formuler des recommandations précises quant au moment où il faut employer la version originale ou la version modifiée de la méthode de Mitofsky et Waksberg, car tout dépend de circonstances qui varient d'une enquête à l'autre. Nous proposons ci-après des lignes directrices concernant le choix de la méthode appropriée.

En règle générale, il est préférable d'utiliser la version originale de la méthode dans le cas des enquêtes nécessitant un contrôle très rigoureux de la taille de l'échantillon ou le prélèvement d'un échantillon pratiquement autopondéré. Bien que la méthode modifiée permette d'estimer avec relativement de précision la taille de l'échantillon, on peut s'attendre à une certaine variation de cette taille, surtout en raison de l'incertitude relative aux taux de non-réponse. Le prélèvement d'un échantillon autopondéré, qu'on ne peut réussir à l'aide de la version modifiée de la méthode, présente par ailleurs certains avantages découlant de la simplification de l'analyse statistique standard.

Comme les coûts d'utilisation des deux méthodes sont différents, il serait utile de disposer de modèles d'estimation de l'écart des coûts pour nous aider à les évaluer. Malheureusement, il n'est pas facile de quantifier les différences entre les coûts inhérents aux versions originale et modifiée de la méthode de Mitofsky et Waksberg. De fait, l'absence de modèles de coûts raisonnables constitue un des principaux problèmes qui restreint notre capacité d'établir des plans de sondage optimaux.

En l'absence de modèles raisonnables d'estimation de l'écart des coûts, nous proposons certaines conditions susceptibles de favoriser l'utilisation d'une méthode plutôt que de l'autre. Ainsi, il est préférable d'utiliser la version modifiée de la méthode lorsque la durée de l'interview est relativement brève. À mesure que l'interview se prolonge, les économies réalisées par suite de l'utilisation de la version modifiée de la méthode risquent de devenir moins élevées, tandis que les variances des estimations s'accroissent.

La durée de l'interview est particulièrement importante dans le cas des enquêtes où l'on soumet les ménages à une présélection afin de trouver ceux qui possèdent certaines caractéristiques. Ainsi, dans certaines enquêtes téléphoniques à composition aléatoire, on soumet les ménages à une sélection préliminaire et on n'effectue l'interview que si un des membres du ménage fait partie d'un groupe cible donné. En pareil cas, l'interview de présélection est souvent très brève. Il peut alors être très avantageux d'utiliser la version modifiée de la méthode de Mitofsky et Waksberg. Les enquêtes au cours desquelles on soumet les ménages à une présélection tendent aussi à avoir un nombre plus élevé d'unités échantillonnées par grappe, ce qui rend la version modifiée de la méthode encore plus efficace. Lorsque dix numéros ou plus sont échantillonnés par grappe (ce qui équivaut à environ six ménages par grappe), les biais dans les estimations obtenues en utilisant la version modifiée de la méthode de Mitofsky et Waksberg sont pratiquement sans importance, et l'utilisation de poids tronqués permet de réduire l'augmentation de la variance à environ 10%. Il est souvent acceptable de prélever des échantillons de dix numéros ou plus par grappe aux fins de la sélection préliminaire, bien qu'il soit généralement inefficace de prélever des échantillons de cette taille dans chaque grappe pour les fins de l'interview, même lorsque le coefficient de corrélation interne intra-classe est peu élevé.

Compte tenu de ces facteurs, on peut de façon générale recommander l'utilisation de la version modifiée de la méthode de Mitofsky et Waksberg lorsque les ménages à l'intérieur des grappes doivent être soumis à une sélection préliminaire. De façon plus explicite, on devrait considérer l'emploi de la méthode modifiée avec des poids tronqués dans les cas suivants: 1) lorsque dix numéros ou plus sont échantillonnés par grappe, et 2), lorsque le coût total de l'utilisation de la méthode modifiée est inférieur d'au moins 10% à celui de l'emploi de la version originale, ou lorsque la période de collecte des données est relativement courte. Si l'on ne satisfait pas à ces deux conditions, il faut fonder le choix de la méthode sur des évaluations relatives à d'autres exigences de l'enquête.

Lorsque le nombre de numéros échantillonnés par grappe est inférieur à dix, le biais et la variance provoqués par l'utilisation de la version modifiée de la méthode de Mitofsky et Waksberg constituent alors des problèmes plus graves. Toutes les caractéristiques corrélées avec la proportion de numéros résidentiels dans une grappe pourraient être touchées lorsque

le nombre de numéros échantillonnés par grappe est si faible. De plus, la variance des estimations obtenues avec la méthode modifiée sera de 20 à 30% plus élevée que si on utilisait la méthode originale, car la troncation n'est pas très efficace avec les échantillons de petite taille. Par conséquent, dans la plupart des enquêtes où il y a moins de dix numéros échantillonnés par grappe, il faut que la méthode originale pose de très graves problèmes d'application avant qu'on décide de l'abandonner et d'employer la méthode modifiée.

## REMERCIEMENTS

Les auteurs tiennent à remercier les arbitres de leurs observations, qui ont été très utiles et qui ont permis de clarifier le contenu du présent article.

## BIBLIOGRAPHIE

CUMMINGS, K.M. (1979). Random digit dialing: A sampling technique for telephone surveys. *Public Opinion Quarterly*, 233-244.

DREW, J.D., DICK, P., et SWITZER, K. (1989). Development and testing of telephone survey methods for household surveys at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-127.

GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., et WAKSBERG, J. (éds.) (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.

HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory*, 2. New York: John Wiley and Sons, 138-139.

KISH, L.(1965). *Survey Sampling*. New York: John Wiley and Sons, 429-430.

LEPKOWSKI, J.M., et GROVES, R.M. (1986). A two phase probability proportional to size design for telephone sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 73-98.

PIEKARSKI, L. (1990). Working block density declines. *The Frame*, une publication de Survey Sampling Inc.

POTTHOFF, R.F. (1987). Generalizations of the Mitofsky-Waksberg technique for random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.

SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.

TUCKER, C. (1989). Characteristics of commercial residential telephone lists and dual frame designs. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-137.

WAKSBERG, J. (1984). Efficiency of alternative methods of establishing cluster sizes in RDD sampling. Note de service non publiée de Westat Inc.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subjects of the population. *Proceedings of the Social Statistics Section, American Statistical Association*.

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 47–61
Statistique Canada



# Le système Blaise ou comment élaborer un système de traitement intégré des données d'enquête

JELKE G. BETHLEHEM et WOUTER J. KELLER[1]

RÉSUMÉ

Grâce aux progrès récents de l'informatique, les organismes nationaux de statistique sont en mesure de produire efficacement des données de première qualité. Le Bureau central de la statistique des Pays-Bas (BCS) recourt de plus en plus au micro-ordinateur pour les diverses étapes de la production de données statistiques. Dans cet article, nous examinons le rôle que jouent les logiciels et les machines dans la collecte et le contrôle des données ainsi que dans la totalisation et l'analyse. Afin de sensibiliser le lecteur aux conséquences néfastes d'une gestion anarchique d'un système décentralisé de traitement de données, nous faisons ressortir toute l'importance de l'intégration. Celle-ci rend la production de données statistiques plus facile à gérer et plus efficace. Cet article est aussi une occasion de présenter le système Blaise, élaboré par le BCS, comme un outil informatique favorisant l'intégration du traitement des données d'enquête. À partir d'une description du questionnaire d'enquête, ce système peut générer automatiquement divers programmes destinés à la collecte des données (IPAO ou ITAO) ou à la saisie et au contrôle des données (IDAO). Il peut aussi être en liaison avec d'autres progiciels. À cet égard, nous décrivons le lien qui existe entre Blaise et les progiciels "maison" Bascula (pour la pondération) et Abacus (pour la totalisation). Ainsi, le système Blaise contrôle et coordonne, et par conséquent intègre, la majeure partie des opérations d'enquête.

MOTS CLÉS: Intégration; traitement des données d'enquête; IPAO; ITAO; micro-ordinateurs; décentralisation; uniformisation.

## 1. INTRODUCTION

Le Bureau central de la statistique des Pays-Bas (BCS) utilise de plus en plus des micro-ordinateurs pour le traitement des données d'enquête. L'arrivée du micro-ordinateur a eu des effets considérables sur les méthodes de travail de l'organisme statistique néerlandais. Les statisticiens des divisions spécialisées prennent conscience graduellement des possibilités de ce nouvel appareil et c'est pourquoi ils l'utilisent de plus en plus fréquemment dans le quotidien.

Dans cet article, nous allons examiner le rôle des nouvelles techniques informatiques dans la collecte et le contrôle des données ainsi que dans la totalisation et l'analyse. Nous allons insister sur l'importance de l'uniformisation et de l'intégration. Ces mesures présentent trois avantages concrets: elles nous épargnent les effets néfastes d'une gestion anarchique d'un système décentralisé de traitement de données, elles facilitent la gestion de la production de données statistiques et favorisent une plus grande efficacité.

Nous présentons de plus le système Blaise, une création du BCS, comme le pivot d'un système intégré de traitement des données d'enquête. D'une part, Blaise se caractérise plus particulièrement par la cohérence qu'il introduit dans les diverses étapes de la collecte et du traitement des données. D'autre part, il favorise l'uniformisation des méthodes entre les divisions. Si toutes les divisions utilisent le même logiciel pour traiter les données de leurs enquêtes, tous parleront le même "langage" et l'échange de renseignements entre divisions sera plus simple et moins sujet à l'erreur.

[1] Jelke G. Bethlehem et Wouter J. Keller, Bureau central de la statistique, division de l'informatique, C.P. 959, 2270 AZ Voorburg, Pays-Bas.

## 2. LA PRODUCTION DE DONNÉES STATISTIQUES

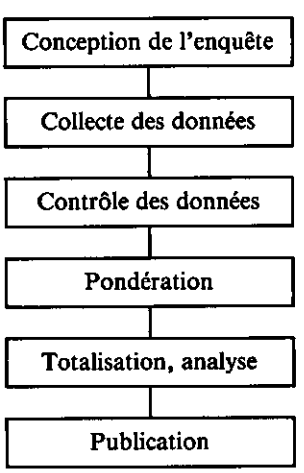
Les organismes nationaux de statistique recueillent des données sur les particuliers, les ménages et les établissements et transforment cette information en des statistiques utiles. La production de données statistiques est un processus souvent complexe, long et coûteux. Cette section vise à décrire les étapes que doit franchir l'organisme de statistique pour mener à bien une enquête, les problèmes qui peuvent se poser en cours de route ainsi que les décisions qui doivent être prises. Le tableau 1 donne un aperçu des diverses étapes d'une enquête.

La première étape est évidemment la conception de l'enquête; le statisticien doit alors déterminer quelle population sera étudiée, quelles données seront recueillies et quelles caractéristiques seront estimées. Comme les organismes statistiques recueillent la plupart de leurs données au moyen d'enquêtes par sondage, il faut élaborer un questionnaire qui contiendra toutes les questions qui seront posées aux répondants. Ce questionnaire est la première description pratique des données à recueillir. De plus, comme il s'agit de sondages, le statisticien doit aussi élaborer un plan d'échantillonnage et veiller à ce que l'échantillon soit prélevé selon les normes.

La deuxième étape du processus est la *collecte des données*. Dans de nombreuses enquêtes, les questionnaires sont remplis généralement à l'occasion d'une interview sur place: un intervieweur se rend à domicile, pose les questions et inscrit les réponses sur un formulaire. La qualité des données recueillies est généralement bonne. Cependant, comme cette méthode nécessite un grand nombre d'intervieweurs, qui peuvent tous être appelés à effectuer de nombreux déplacements, l'interview sur place est une formule qui peut s'avérer coûteuse et chronophage. C'est pourquoi on a recours parfois à l'interview téléphonique. Les intervieweurs appellent les répondants depuis les bureaux de l'organisme; ils n'ont donc plus besoin de se déplacer. En revanche, il n'est pas toujours possible de réaliser des interviews téléphoniques: seules les personnes qui ont le téléphone peuvent être rejointes de cette manière, et le questionnaire ne doit pas être trop long ni trop compliqué. L'enquête postale est une solution encore moins coûteuse car elle n'exige pas les services d'intervieweurs. On envoie un questionnaire aux répondants éventuels et ceux-ci sont priés de le retourner dûment rempli. Bien que l'on dispose de lettres de rappel dans ces conditions, cela n'équivaut pas au pouvoir de persuasion d'un intervieweur; par conséquent, le taux de réponse et la qualité des données recueillies en souffrent.

**Tableau 1**

La production de données statistiques



Si la collecte des données se fait au moyen de questionnaires, ceux-ci, une fois remplis, doivent faire l'objet d'un traitement approfondi. Pour obtenir des données de première qualité, il est indispensable d'extraire de ces questionnaires toutes les erreurs que l'on peut y trouver. Nous en sommes à l'étape du *contrôle des données*. On y distingue trois types d'erreur: - l'*erreur d'inscription* (lorsque la réponse donnée est en dehors de l'intervalle des réponses possibles, par ex.: âge = 348 ans); - l'*erreur de logique* (lorsqu'il y a incohérence entre des réponses à certaines questions, par ex.: il est raisonnable de croire qu'une personne peut être âgée de 8 ans, et il n'y a rien d'anormal à ce qu'une personne se dise mariée, mais lorsque ces deux caractéristiques se rapportent au même individu, il y a sûrement quelque chose qui cloche, du moins en ce qui concerne les Pays-Bas); - l'*erreur de cheminement* (lorsque l'intervieweur ou le répondant ne respecte pas les instructions relatives à la manière de remplir le questionnaire, c'est-à-dire lorsqu'il ne tient pas compte de l'enchaînement logique des questions: ainsi, des réponses sont données à des questions qui ne concernent pas l'individu, et vice-versa).

Les erreurs qui ont été décelées doivent être corrigées, mais cette tâche peut s'avérer très ardue si elle doit exécutée postérieurement, dans les bureaux de l'organisme statistique. Dans beaucoup de cas, particulièrement en ce qui concerne les enquêtes-ménages, il est impossible de joindre de nouveau le répondant; il faut donc trouver d'autres moyens de résoudre la difficulté. Parfois, on réussira à déterminer une valeur approximative acceptable à l'aide d'une méthode d'imputation mais dans d'autres circonstances, on remplacera la valeur inexacte par un code spécial indiquant que la valeur est "inconnue".

À l'étape du contrôle des données, on procède aussi parfois à ce qu'on appelle le *codage des réponses libres*. Un exemple typique de question à réponse libre est la question relative à la profession du répondant. Il est évidemment plus facile de traiter les questions qui renferment une liste de réponses pré-codées. Toutefois, dans le cas de la profession, cette liste serait interminable et le répondant aurait beaucoup de difficulté à faire le bon choix. Pour éviter ce problème, il suffit de laisser le répondant inscrire en toutes lettres sa profession sur le questionnaire. Ces réponses doivent être ensuite classées si l'on veut être en mesure d'analyser ce genre d'information. La classification des réponses est une opération longue et coûteuse qui doit être confiée à des spécialistes chevronnés.

Le contrôle des données produit un fichier "épuré", c.-à-d. un fichier ne contenant pas d'erreur. Cependant, il est encore trop tôt pour procéder à la totalisation et à l'analyse des données. En premier lieu, l'échantillonnage peut avoir été fait avec des probabilités inégales; par exemple, des établissements ont été échantillonnés avec une probabilité proportionnelle à la taille. Un choix judicieux des probabilités d'échantillonnage permet d'obtenir des estimations plus précises des paramètres de la population mais à la condition que l'on utilise une méthode d'estimation qui tient compte des différences de probabilité. En second lieu, la représentativité de l'échantillon peut être affectée par la non-réponse, c'est-à-dire le phénomène selon lequel il manque des données pour certains éléments de l'échantillon. Si le comportement des non-répondants varie en fonction des caractéristiques démographiques étudiées, les résultats de l'enquête seront biaisés.

Afin de compenser les effets de la non-réponse et de l'inégalité des probabilités d'échantillonnage, on a souvent recours à la *pondération*. À cette étape, on attribue un poids à chaque enregistrement. Les poids sont calculés de telle manière que la distribution empirique pondérée de caractères comme le sexe, l'âge, l'état matrimonial et la région reflète la distribution connue de ces caractères dans la population.

Dans les cas de non-réponse partielle, c'est-à-dire les cas où un enregistrement ne contient pas toutes les réponses voulues, on peut aussi procéder à une *imputation*. En se servant d'un modèle quelconque, on calcule des estimations pour les valeurs manquantes et on les inclut dans l'enregistrement.

À l'issue de cette étape, nous avons un fichier épuré prêt pour l'*analyse*. Celle-ci débute presque toujours par la totalisation des caractéristiques de base. La construction de tableaux n'est pas une opération aussi simple qu'on pourrait le croire. En effet, des facteurs tels la

composition des lignes et des colonnes (construites souvent à partir de plusieurs variables), les quantités figurant dans les cases (effectifs, moyennes, pourcentages), le mode de calcul des pourcentages, le traitement des variables à réponse multiple, la position des totaux et des sous-totaux et beaucoup d'autres facteurs encore peuvent compliquer grandement la tâche des statisticiens.

De nombreux organismes statistiques soumettent aussi leurs données à des analyses qui visent à en révéler la structure fondamentale; ils peuvent ainsi faire une meilleure lecture de ces données. Les résultats de ces analyses peuvent contribuer à perfectionner les enquêtes ultérieures et par conséquent à améliorer la qualité des données ou à réduire les coûts des opérations.

Enfin, les résultats des analyses sont *publiés* sous forme de rapport. Ce rapport renferme normalement des tableaux et des graphiques. Il est essentiel de présenter l'information statistique de manière à ce qu'elle soit comprise de la bonne façon. Les graphiques, les tableaux et les textes doivent être simples et limpides. La construction des graphiques mérite une attention particulière car un graphique équivoque donnera lieu très facilement à de fausses interprétations.

## 3. NÉCESSITÉ DE L'INTÉGRATION

L'ordinateur a toujours occupé une place importante dans le traitement des données statistiques. À l'origine, il servait uniquement à des opérations comme le tri, le dénombrement et la totalisation. Dans les années soixante et soixante-dix, l'apparition des gros ordinateurs et des progiciels statistiques a permis l'exécution d'analyses plus poussées. L'ordinateur servait aussi de plus en plus au contrôle des données, à la pondération et à l'imputation. Ce n'est que plus tard que l'ordinateur commencera à servir à la collecte des données. La première occasion sera les interviews téléphoniques (ITAO). La dernière décennie a vu naître l'ordinateur portatif; désormais, l'intervieweur peut se rendre au domicile du répondant avec un ordinateur sous le bras. Cette forme d'interview est désignée par le sigle IPAO (interview sur place assistée par ordinateur).

Il ne fait plus de doute que l'ordinateur est utilisé à des fins de plus en plus nombreuses. Il existe du matériel et des logiciels pour presque toutes les étapes du processus de production. En outre, de plus en plus de gens utilisent les outils informatiques. Au début, les ordinateurs étaient réservés à l'usage des spécialistes de l'informatique mais aujourd'hui, les statisticiens et les experts des divisions spécialisées sont rompus aux techniques de l'informatique et réclament avec toujours plus d'intérêt le matériel et les logiciels voulus pour accomplir leur tâche. À l'heure actuelle, le traitement électronique des données peut être confié aux divisions spécialisées pourvu qu'il n'implique pas d'opérations complexes; de fait, les divisions se chargent déjà de cette opération, laissant aux spécialistes de la division de l'informatique l'élaboration et la maintenance des systèmes d'information complexes. Ces changements ont eu pour conséquence de modifier le travail des statisticiens et des experts des divisions spécialisées. Auparavant, ces gens étaient des spécialistes de leur domaine propre, sans plus; mais aujourd'hui, ils ont élargi leur champ de connaissances et diversifié leur expérience en touchant aussi bien à la méthodologie statistique et à l'informatique qu'à des sujets spécialisés. Ainsi, le groupe de spécialistes a fait place à un nouveau groupe qui est plus au fait des divers aspects du traitement des données d'enquête.

Il est bien d'automatiser la production de données statistiques mais il faut être conscient des risques que cela implique. Même si l'utilisation de l'ordinateur est le gage d'une plus grande efficacité et d'une meilleure qualité des données, un usage incontrôlé des nouvelles techniques peut facilement mener à la confusion et, partant, à une productivité moindre. Voici les facteurs qui influent sur l'efficacité de la production de données statistiques:

- **Multiplicité des étapes**
  L'information passe entre plusieurs mains: le répondant remplit le questionnaire, l'agent spécialisé en vérifie le contenu et corrige les erreurs, l'agent de saisie introduit les données dans l'ordinateur et le programmeur élabore des logiciels de contrôle. L'échange de documents entre des personnes ou des divisions peut être une source d'erreur, de confusion et de retard.

- **Multiplicité des systèmes informatiques**
  Les diverses opérations liées au traitement des données peuvent ne pas être exécutées sur le même système. Le transfert de fichiers provoque des retards; en outre, une spécification et une documentation fautives peuvent être à l'origine d'erreurs.

- **Répétition des spécifications relatives aux données**
  La structure des données doit être spécifiée à presque toutes les étapes. Les questions qui reviennent le plus souvent dans les divisions sont les suivantes: Quelle est la signification des variables? Quelles valeurs peuvent-elles prendre? Des conditions s'appliquent-elles au processus de cheminement? Quelles relations entre les variables faut-il vérifier? Bien que les spécifications soient essentiellement les mêmes d'une fois à l'autre, le mode de spécification, lui, peut être très différent. Chaque système a son propre "langage". La première spécification concerne le questionnaire proprement dit. Une deuxième spécification peut être nécessaire pour la saisie des données, et une autre pour le programme de contrôle, la totalisation et l'analyse, *etc.* De toute évidence, ce n'est pas la façon la plus efficace de traiter l'information.

Devant ces difficultés, le BCS propose comme solution l'*intégration*, une intégration sur trois plans: tâches, matériel et logiciel. Examinons tout d'abord l'intégration des tâches.

Dans sa forme classique, le traitement des données consiste en ce que nous appelons des *macro-cycles*. L'ensemble des données d'une enquête passe par des cycles: d'une division à une autre et d'un système informatique à un autre. En premier lieu, la division spécialisée s'occupe d'"épurer" manuellement les questionnaires; ensuite, la section de la saisie des données voit à l'introduction des données de ces questionnaires; en troisième lieu, les fichiers sont transférés dans un gros ordinateur. Un programme soumet les données à une vérification de cohérence; les erreurs détectées sont inscrites sur une liste que l'on envoie à la division spécialisée pour qu'elle fasse les corrections nécessaires. Ce processus de saisie et de contrôle doit être répété un certain nombre de fois avant que l'on puisse dire que les données sont "épurées".

Ce que vise au fond l'intégration des tâches est le remplacement des macro-cycles par des *micro-cycles*. Autrement dit, on devrait songer à faire circuler un seul enregistrement à la fois plutôt que tout le fichier de données. La notion de micro-cycle signifie que toutes les opérations successives se font à l'intérieur du même système informatique et qu'elles sont sous le contrôle d'une seule division. Substituer le micro-cycle au macro-cycle revient à concentrer toutes les opérations de traitement des données dans une seule division, soit la division spécialisée. Comme les statisticiens des divisions spécialisées sont les personnes qui connaissent le mieux le domaine sur lequel porte une enquête, ils sont le plus en mesure de traiter les données, de résoudre les problèmes et de produire des statistiques de première qualité. Pour s'acquitter de ces tâches, il faut, bien sûr, qu'ils aient les outils appropriés, c.-à-d. des systèmes puissants et conviviaux.

L'idée de confier les opérations de traitement des données d'enquête à des spécialistes de l'informatique est révolue. Les statisticiens des divisions spécialisées se rendent compte de plus en plus des possibilités et de l'utilité de l'ordinateur pour leur travail. C'est pourquoi les divisions spécialisées devraient songer à prendre en charge les opérations liées au traitement des données d'enquête, dans la mesure où il s'agit d'opérations peu complexes. Évidemment, il revient à la division de l'informatique de mettre en place l'infrastructure nécessaire. Par ailleurs, cette division demeure responsable de l'élaboration et de la maintenance des systèmes d'information complexes.

Le second volet de la solution du BCS est l'intégration du matériel. Celle-ci vise à faire exécuter le maximum d'opérations sur un seul type d'ordinateur. Compte tenu de ce qu'un grand nombre de statisticiens peu initiés à l'informatique devront se servir d'un ordinateur,

il va de soi que l'on opte pour le micro-ordinateur. On peut ainsi s'offrir la convivialité à un prix relativement raisonnable et avoir accès à toute une série de logiciels utiles.

Sachant que les organismes statistiques traitent régulièrement d'énormes quantités de données, on peut se demander si les micro-ordinateurs peuvent à eux seuls suffire à la tâche et remplacer les gros ordinateurs. Pour être en mesure de répondre à cette question, nous devons distinguer deux types d'opérations liées au traitement des données. Premièrement, les opérations axées sur l'enregistrement, c'est-à-dire celles qui portent sur un seul enregistrement à la fois (par ex.: saisie et contrôle des données). Les opérations de ce genre se prêtent habituellement très bien au traitement interactif. Deuxièmement, les opérations axées sur le fichier, c'est-à-dire celles qui ne peuvent être exécutées convenablement que si l'on dispose de tout le fichier (par ex.: calcul des poids et totalisation). En raison de leur envergure, les opérations axées sur le fichier sont souvent exécutées suivant le principe du traitement par lots.

Il y a quelques années, on était d'avis que les opérations axées sur l'enregistrement pouvaient être faites sur micro-ordinateur mais que celles axées sur le fichier devaient être exécutées sur un gros ordinateur. Comme on fabrique des micro-ordinateurs de plus en plus puissants, l'attention est tournée de plus en plus vers le micro-ordinateur. À l'heure actuelle, la position du BCS est d'utiliser le micro-ordinateur pour toutes les opérations axées sur l'enregistrement et pour un grand nombre des opérations axées sur le fichier (pour des fichiers de moins de 50 méga-octets par exemple). En revanche, nous conservons le gros ordinateur pour le stockage des données et les gros travaux en traitement par lots.

Les personnes qui doivent utiliser un ordinateur devraient être le moins possible mises en contact avec un gros ordinateur. C'est pourquoi le BCS favorise de plus en plus l'utilisation de systèmes frontal et dorsal. Le système frontal est constitué de micro-ordinateurs et c'est ce qu'utilisent les statisticiens pour traiter les problèmes courants. Le système dorsal consiste en un gros ordinateur ou en un mini-ordinateur et sert aux gros travaux; les opérations pertinentes peuvent même se faire à l'insu de l'utilisateur. En particulier, le serveur de base de données est une option très intéressante en ce qui concerne la gestion de bases de données. Selon cette option, les opérations portant sur la base de données s'effectuent sur un mini-ordinateur spécialisé tandis qu'elles sont spécifiées, déclenchées et contrôlées au moyen d'un micro-ordinateur placé sur le bureau de l'utilisateur.

## 4. UNIFORMISATION

Le BCS utilise de plus en plus le micro-ordinateur (avec système MS-DOS) à beaucoup d'étapes de la production de données statistiques. D'une part cette initiative favorise un traitement plus efficace de l'information, de l'autre elle crée de nouveaux problèmes qu'il faut chercher à résoudre. Si on laisse à chaque division la liberté d'acheter le matériel et le logiciel de son choix, on risque fort de créer une situation chaotique. Les divisions ne parleront plus le même "langage" parce qu'elles n'utiliseront pas les mêmes structures de données ni les mêmes logiciels. De toute évidence, cette perspective commande une politique d'uniformisation vigoureuse. Le BCS a adopté une telle politique et en conséquence, les divisions disposent tout au plus de deux progiciels pour une tâche particulière.

Un autre avantage de l'uniformisation est qu'elle limite le nombre de cours de formation que doivent suivre les utilisateurs. Afin d'assurer la formation de nombreux nouveaux utilisateurs, le BCS organise en moyenne chaque mois 50 cours d'une journée (ce qui suffit à occuper trois salles de cours, dotées de tout le matériel nécessaire, à chaque jour ouvrable).

On doit aussi porter attention à la manière dont sont utilisés les micro-ordinateurs dans l'organisation. La distribution d'un grand nombre de micro-ordinateurs autonomes peut sembler une solution commode mais cela soulève des problèmes qu'il faut chercher à résoudre. Par exemple, il est très facile de copier des fichiers de données (confidentielles) sur le disque dur d'un micro-ordinateur autonome; il y a donc là un problème de sécurité informatique.

De plus, les utilisateurs dans les divisions spécialisées négligent souvent d'effectuer des opérations comme la sauvegarde de fichiers et l'archivage. Par ailleurs, les divisions ne peuvent se partager l'utilisation de fichiers autrement que par l'échange de disquettes. Enfin, pour les grandes organisations qui possèdent beaucoup de micro-ordinateurs autonomes, la distribution des nouvelles versions de progiciels, y compris la documentation pertinente, est une tâche souvent contraignante.

Pour éviter toutes ces difficultés, le BCS a installé quelque 60 réseaux locaux (RL). Chaque division a son propre RL. De dix à soixante micro-ordinateurs sont reliés à un serveur de fichiers haut de gamme articulé sur un microprocesseur 386 et dont la capacité de mémoire peut atteindre 600 méga-octets. Ainsi, près de 2,300 micro-ordinateurs sont reliés de cette manière à un serveur de fichiers et la moitié d'entre eux sont articulés autour du microprocesseur 386SX d'Intel. La sécurité est assurée de trois façons: - protection par mot de passe dans la procédure d'entrée en communication, - chiffrement et - postes de travail sans unité de disquette (seulement 60 des 2,300 micro-ordinateurs sont équipés d'une unité de disquette ou de disque dur). Les opérations d'archivage et de sauvegarde des RL s'effectuent de façon centralisée sous la responsabilité de la division de l'informatique. Chaque nuit, on procède à une sauvegarde complète de plus de 15 giga-octets. Il est clair qu'un tel contexte d'utilisation facilite des opérations comme le contrôle de la version et la mise à jour du logiciel. Avec un RL, la distribution et l'installation des nouvelles versions de logiciels deviennent des tâches simples puisque par une seule commande, on peut télécharger de nouvelles versions dans tous les serveurs de fichiers à la fois. Tous les permis d'utilisation de logiciel prévoient un usage en commun, qui est contrôlé par des logiciels "maison".

Le micro-ordinateur joue un rôle de plus en plus grand dans la production de données statistiques, mais le gros ordinateur ou le mini-ordinateur sont encore indispensables pour certaines applications, comme l'exploitation de grosses bases de données. Dans ce contexte, le BCS a fait de Oracle son système de gestion de bases de données standard. L'élaboration d'un programme de gestion de base de données se fait de préférence sur un micro-ordinateur alors que l'exécution de ce programme se fait sur un mini-ordinateur. Dernièrement, le BCS a élaboré une architecture client-serveur fondée sur un système de gestion de base de données réparties. Les micro-ordinateurs du réseau constituent le système frontal et les mini-ordinateurs, le système dorsal.

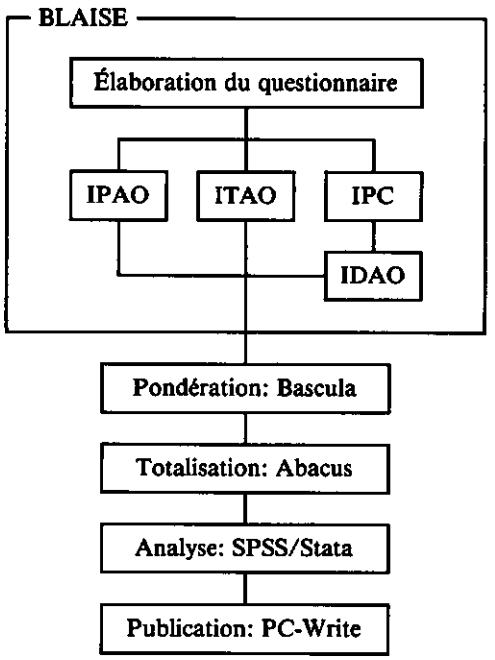
Ainsi, plus les outils informatiques sont à la portée des agents spécialisés des divisions (décentralisation), plus il est nécessaire de centraliser les opérations d'uniformisation et de coordination de l'univers informatique des utilisateurs. Pour plus de détails sur l'infrastructure informatique, voir Keller, Metz et Bethlehem (1990).

## 5. INTÉGRATION DU TRAITEMENT DES DONNÉES D'ENQUÊTE

Dans la section précédente, nous avons discuté de la nécessité d'une intégration du traitement des données d'enquête. Il a été question plus particulièrement de la concentration des tâches dans les divisions spécialisées et de l'uniformisation du matériel et du logiciel. Or, l'uniformisation du logiciel ne suffit pas. On peut accroître davantage l'efficacité de la production de données statistiques en intégrant les logiciels nécessaires dans un système. Cette section vise à décrire la manière dont le BCS a mis en oeuvre un système intégré de traitement des données d'enquête.

Un système intégré de traitement des données d'enquête doit reposer sur un langage puissant pour la spécification des questionnaires. Cette spécification est la "base de connaissances", qui contient tous les faits relatifs au questionnaire et aux données. Le système intégré doit pouvoir exploiter ces connaissances, c'est-à-dire qu'il doit pouvoir générer automatiquement toutes les applications requises pour le traitement des données, soit d'une part, les logiciels pour la collecte, la saisie et le contrôle des données et d'autre part, les interfaces avec d'autres

**Tableau 2**
Système intégré de traitement des données d'enquête




logiciels de traitement de données (par ex.: logiciels pour la totalisation et l'analyse). Ainsi, la répétition des spécifications relatives aux données n'a plus raison d'être et la cohérence des données est observée à toutes les étapes du traitement.

Le pivot du système intégré de traitement des données d'enquête mis au point par le BCS est le système Blaise. À l'étape de la conception de l'enquête, le questionnaire est spécifié en langage Blaise et c'est cette spécification qui sert, durant tout le processus, à extraire les données requises pour les diverses étapes du traitement. Le tableau 2 décrit schématiquement le système intégré de traitement des données d'enquête.

Le système Blaise peut générer trois types de programmes: IDAO, IPAO et ITAO (le sigle IDAO signifie "introduction de données assistée par ordinateur"). Ce système intègre les opérations de saisie et de contrôle des données en permettant le traitement interactif des questionnaires. Il peut aussi produire le logiciel nécessaire pour les IPAO ou les ITAO. Le système Blaise est décrit plus en détail dans la section 6.

Quel que soit le mode de collecte des données, nous obtiendrons toujours un fichier "épuré", c.-à-d. un fichier où on ne peut plus trouver d'erreurs. Viendra ensuite le calcul de facteurs de pondération. Le programme Bascula est conçu spécialement à cette fin. Il peut lire directement les fichiers Blaise et extraire de la spécification Blaise l'information relative aux variables, c'est-à-dire les métadonnées. L'exécution de Bascula aura pour effet d'ajouter une variable dans le fichier contenant les facteurs de pondération pour les divers cas. Le programme Bascula est décrit plus en détail dans la section 7.

Le fichier est maintenant prêt pour la totalisation et à cette fin, le système intégré met à notre disposition le programme Abacus. Ce programme peut, lui aussi, lire et comprendre les fichiers créés à l'étape précédente. Voir la section 8 pour plus de détails. La totalisation peut être suivie d'une analyse plus approfondie des données. À cette fin, le système Blaise peut générer des interfaces pour les progiciels statistiques SPSS et Stata. La section 9 contient plus de détails à ce sujet.

Enfin, on préparera une publication à l'aide du logiciel de traitement de texte PC-Write. Comme ce logiciel peut tourner sur le même système que les autres logiciels, il est facile d'importer dans le texte des tableaux et des résultats d'analyses statistiques.

## 6. LE SYSTÈME BLAISE

Le système Blaise a été élaboré par le BCS et il doit son nom au célèbre théologien et mathématicien français Blaise Pascal (1623-1662). La base de ce système est le langage Blaise, qui sert à spécifier de façon formelle la structure et le contenu du questionnaire d'enquête. Le langage Blaise tire son origine, en majeure partie, du langage de programmation Pascal.

Le système Blaise tourne sur des micro-ordinateurs (ou des réseaux de micro-ordinateurs) dont le système d'exploitation est MS-DOS. Il est le pivot du système intégré de traitement des données d'enquête et comme il est destiné à l'usage des employés des divisions spécialisées, nul n'a besoin d'être informaticien pour l'utiliser. On a conçu le système Blaise dans le but d'offrir aux spécialistes des divisions un outil puissant mais convivial qui leur permette de saisir les données relatives à une enquête et de voir à l'exécution de toutes les étapes subséquentes.

Selon le système Blaise, la réalisation d'une enquête débute par l'élaboration d'un questionnaire en langage Blaise. Un questionnaire élaboré de cette façon contient plus de renseignements qu'un questionnaire classique. On y décrit non seulement les questions, les réponses possibles et les conditions de cheminement dans le questionnaire, mais aussi les relations entre les réponses qui doivent faire l'objet d'un contrôle.

Le système Blaise peut générer des programmes pour l'IDAO ainsi que les IPAO et les ITAO. Un programme IDAO est un système intelligent et interactif destiné à la saisie et au contrôle des données recueillies à l'aide de formulaires. L'agent de la division spécialisée traite un à un les formulaires au moyen d'un micro-ordinateur. Il inscrit les réponses à l'endroit approprié et une fois le questionnaire terminé, il actionne la fonction contrôle pour vérifier s'il n'y aurait pas d'erreur de cheminement ou d'erreur de logique. Les erreurs détectées sont affichées à l'écran avec une explication. L'agent peut corriger ces erreurs soit en se reportant au formulaire ou bien en appelant la personne qui a fourni les renseignements en question. Une fois que toutes les erreurs ont été corrigées, l'enregistrement "épuré" est versé dans un fichier.

Le programme IDAO peut aussi servir à une forme de traitement de données dont on n'a pas fait mention jusqu'à maintenant. En effet, les organismes statistiques ne recueillent pas toujours eux-mêmes les données dont ils ont besoin; ils sont appelés parfois à construire des statistiques à l'aide de fichiers provenant de l'extérieur. Ces données doivent tout de même être vérifiées. Le système Blaise comporte une fonction qui permet d'importer les fichiers de ce genre. Grâce au programme IDAO, on peut soumettre tous les enregistrements à un contrôle intégral par lots. Les enregistrements sont alors classés en deux groupes: "sans erreur" et "avec erreur". Les seconds peuvent être corrigés en mode interactif grâce, une fois de plus, au programme IDAO.

Les programmes IPAO et ITAO peuvent servir aux interviews assistées par ordinateur. Dans ce cas, le questionnaire classique est remplacé par un programme informatique qui contient les questions à poser. Ce programme contrôle entièrement l'interview. Il détermine au fur et à mesure l'ordre dans lequel les questions doivent être posées et vérifie les réponses dès qu'elles ont été saisies. En ce qui concerne l'IPAO, le programme est chargé dans un ordinateur portatif que l'intervieweur apporte au domicile du répondant. Pour ce qui a trait à l'ITAO, le programme est chargé dans un ordinateur de bureau. L'intervieweur appelle le répondant depuis une unité centrale et réalise l'interview par téléphone.

La génération d'un programme IDAO, IPAO ou ITAO de Blaise se fait en un certain nombre d'étapes. Premièrement, on introduit la spécification en langage Blaise du questionnaire au moyen d'un programme de mise en forme de texte; on vérifie ensuite s'il n'y a pas d'erreur de syntaxe. Les erreurs détectées doivent être corrigées et à cette fin, le système rappelle le

**Tableau 3**

Exemple d'un questionnaire Blaise

```
QUESTIONNAIRE de "l'enquête sur le travail";

QUEST
  NumOrd     "Numéro d'ordre de l'interview?": 1..1000 (CLÉ);
  Age        "Quel âge avez-vous?": 0..99;
  Sex        "Êtes-vous de sexe masculin ou féminin?": (Masculin, Féminin);
  EtaMatr    "Quel est votre état matrimonial?":
               (Marié     "Marié",
               NonMar     "Non marié")
  Emp        "Occupez-vous un emploi?": (Oui, Non);
  DesEmp     "De quel genre d'emploi s'agit-il?": [20] CARACTÈRES;
  Revenu     "Quel est votre revenu annuel?":
               (Inf20     "Moins de 20,000",
               Jusq40     "Entre 20,000 et 40,000",
               Plus40     "Plus de 40,000");
  Deplac     "Quel moyen de transport utilisez-vous habituellement pour aller au travail?":
               SÉRIE [3] DE
               (Pied      "à pieds",
               Bicycle    "à bicyclette",
               Aut        "en voiture ou en motocyclette",
               TransCom   "autobus, tramway, train ou métro",
               Autre      "aucun de ceux-là");
  AutMoyen   "Quel autre moyen de transport utilisez-vous?": [20] CARACTÈRES;

CHEMINEMENT
  NumOrd; Age; Sex; EtaMatr; Emp;
  SI Emp = Oui ALORS
    DesEmp; Revenu; Deplac;
    SI Autre sous Deplac ALORS AutMoyen FIN
  FIN

CONTRÔLE
  SI Age < 15 "Répondant n'a pas encore 15 ans" ALORS
    EtaMatr = NonMar "trop jeune pour être marié!"
  FIN

FINQUESTIONNAIRE.
```

programme de mise en forme de texte et place le curseur à l'endroit où se trouve approximativement l'erreur. Après la correction des erreurs, on soumet la spécification à un nouveau contrôle. Si aucune autre erreur n'est détectée, la spécification est transformée en code source Pascal, lequel est compilé en un programme exécutable.

Le langage Blaise a un double objet quelque peu contradictoire. D'une part, il doit être assez puissant pour pouvoir être appliqué à toutes les grandes enquêtes complexes et d'autre part, les spécifications de questionnaires en langage Blaise doivent être suffisamment lisibles pour que les agents spécialisés puissent s'en servir. De fait, l'information contenue dans un questionnaire Blaise doit être explicite; autrement dit, cette information représente la description fondamentale de l'enquête et est utilisable par tous les intéressés. Le tableau 3 reproduit un questionnaire simple en langage Blaise.

La première partie du questionnaire est désignée par le sigle QUEST; cette section contient la définition de toutes les questions qui peuvent être posées. Une définition se compose d'un symbole d'identification (pour usage interne dans le questionnaire), du libellé de la question telle qu'elle est posée au répondant et d'une énumération des réponses possibles. La deuxième partie du questionnaire correspond à la section CHEMINEMENT. Cette section indique dans quel ordre les questions doivent être posées et à quelles conditions. La section CONTRÔLE sert à définir les vérifications de cohérence.

La description ci-dessus ne reflète pas toute la puissance du langage Blaise. Pour avoir une idée générale de ce langage, prière de se référer à Bethlehem et coll. (1989b) et pour plus de détails, voir Bethlehem et coll. (1989c).

Le système Blaise comprend un module pour le *codage interactif*; ce module permet d'intégrer des opérations de codage à l'étape de la collecte des données ou à l'étape de la saisie et du contrôle. Le module se compose de deux "fonctions". La première met en application une approche hiérarchique du codage. Pour coder une réponse, l'opérateur commence par entrer le premier chiffre du code en choisissant la catégorie pertinente dans un menu. Une fois le premier chiffre saisi, le programme affiche un nouveau menu qui est une description plus détaillée de la catégorie choisie précédemment. Ainsi, on obtient une description de plus en plus détaillée à mesure qu'on approche du dernier chiffre. La seconde "fonction" du module représente une approche alphabétique du codage. Elle vise à trouver une description dans une liste alphabétique. Si cette description n'est pas repérée, le programme affiche la liste à partir de la description la plus près possible de la description recherchée. La liste peut être dressée de manière que l'on puisse y trouver presque toutes les descriptions possibles, y compris les permutations. Cette méthode a l'avantage d'être simple, rapide et contrôlable. Les deux fonctions de codage peuvent être appliquées simultanément.

## 7. BASCULA

Une fois "épuré", le fichier de données d'enquête produit par le système Blaise n'est pas encore prêt à servir à des inférences concernant la population d'où a été tiré l'échantillon parce que les données disponibles ne constituent pas un échantillon représentatif; il faut donc procéder à certains ajustements.

Afin de tenir compte de la non-réponse et de l'inégalité des probabilités de sélection, il faut souvent calculer des facteurs de pondération. La stratification a posteriori est une méthode de pondération bien connue. Chaque enregistrement reçoit un poids calculé de telle manière que la distribution empirique pondérée de caractères tels le sexe, l'âge, l'état matrimonial et la région reflète la distribution (connue) des mêmes caractères dans la population. Deux facteurs importants peuvent compliquer la stratification a posteriori: l'existence de strates vides et le manque d'information pertinente sur la population. Des recherches ont été faites au BCS afin d'améliorer les méthodes de pondération. Ces recherches ont abouti à la mise en oeuvre d'une nouvelle méthode générale de pondération qui permet de calculer les poids à l'aide d'un modèle linéaire qui établit un rapport entre les variables étudiées dans une enquête et les variables auxiliaires. La stratification a posteriori est un cas particulier de cette méthode. En raison de la généralité de la méthode, on peut appliquer divers schémas de pondération qui exploitent le plus possible l'information existante sur la population sans soulever les problèmes mentionnés plus haut. Voir Bethlehem et Keller (1987) pour plus de détails.

Bascula est un programme général de pondération qui tourne sur des micro-ordinateurs dont le système d'exploitation est MS-DOS. Ce programme réunit plusieurs méthodes de pondération. En premier lieu, il peut exécuter une stratification a posteriori classique. Et si le nombre de strates vides est faible, il peut fondre (c'est-à-dire combiner) ces strates avec des strates avoisinantes. Lorsque le nombre de strates vides est élevé ou qu'il manque d'informations sur la population, Bascula peut exécuter la pondération linéaire décrite plus haut ou effectuer un

ajustement proportionnel itératif (aussi appelé pondération multiplicative ou méthode itérative du quotient). Les poids ainsi calculés peuvent soit être ajoutés au fichier de données ou être stockés dans un fichier à part.

Bascula peut lire directement les fichiers Blaise et extraire de la spécification Blaise l'information pertinente sur les variables. Il revient à l'utilisateur de fournir l'information sur la population. Le programme est à base de menus, donc convivial. Il exécutera une stratification a posteriori complète si cela est possible. Sinon, l'utilisateur doit choisir entre une pondération linéaire ou une pondération multiplicative.

À l'heure actuelle, Bascula ne peut être utilisé que sur un micro-ordinateur. Nous prévoyons mettre au point une version "dorsale" qui tournera sur un gros ordinateur. Bascula a été conçu spécialement pour les enquêtes sociales et démographiques, où la stratification a posteriori se mêle à des méthodes d'estimation relativement simples. En ce qui concerne les enquêtes économiques, il faudra élaborer un logiciel différent qui mettra l'accent sur l'échantillonnage stratifié et des estimateurs plus complexes (estimateur par quotient; estimateur par régression).

## 8. ABACUS

La totalisation est une des principales étapes de la production de données statistiques et elle a été parmi les premières à être informatisées. Il existe déjà de nombreux progiciels de totalisation dans le monde mais peu sont vraiment conviviaux. Cela s'explique en partie par le fait qu'il faille définir un grand nombre de paramètres pour produire un tableau complexe selon les normes; ces paramètres peuvent être: les variables à utiliser pour les diverses dimensions (ligne, colonne, couche); le choix entre l'enchaînement des variables (c.-à-d. présenter toutes les valeurs d'une variable, puis toutes celles d'une autre variable) ou leur emboîtement (pour chaque valeur d'une variable, présenter toutes les valeurs possibles d'une autre variable) pour une dimension donnée; la quantité figurant dans les cases (fréquence, pourcentage, total, moyenne); le choix de présenter ou non les totaux et les sous-totaux; et de nombreuses caractéristiques de présentation. Pour pouvoir maîtriser tous ces paramètres, les progiciels classiques renferment des langages de commande pour définir les tableaux et ces langages ne sont pas très faciles à apprendre ni à utiliser.

Abacus est un progiciel de totalisation qui tourne sur des micro-ordinateurs dont le système d'exploitation est MS-DOS. Bien qu'on puisse le voir comme un progiciel de totalisation parmi tant d'autres, Abacus a été élaboré de façon très particulière. Premièrement, les tableaux ne sont pas définis au moyen de langages de commande. Le programme est plutôt à base de menus. L'utilisateur conçoit son tableau d'une manière simple et intuitive en mode interactif sans avoir à connaître aucun langage de commande. Deuxièmement, Abacus peut lire directement les fichiers créés par le système Blaise, de même que les fichiers ASCII. Les métadonnées, c.-à-d. l'information relative aux variables dans le fichier, peuvent être produites par le système Blaise ou encore (dans le cas de fichiers ASCII indépendants) être introduites en mode interactif par l'utilisateur. Troisièmement, le programme peut produire des tableaux prêts à la photo.

Une autre caractéristique remarquable d'Abacus est sa vitesse. Il prend environ 6 secondes pour produire un tableau que SPSS produit en 3 minutes (sur le même micro-ordinateur de type 386SX). Si Abacus est aussi rapide, c'est à cause de sa taille relativement modeste; il peut donc utiliser une grande partie de la mémoire comme zone de travail, et ainsi produire des tableaux pouvant contenir jusqu'à 90,000 cases.

Les tableaux générés par Abacus peuvent avoir jusqu'à trois dimensions (couches, lignes et colonnes). Le nombre de variables pour chaque dimension peut atteindre 10; ces variables peuvent être emboîtées ou enchaînées. Dans le tableau 4 par exemple, les variables "emploi" et "sexe" (colonnes) sont présentées selon le mode enchaînement tandis que les variables "région" et "ville" (lignes) sont présentées selon le mode emboîtement. Ce tableau ne comporte pas une troisième dimension (couches).

**Tableau 4**

Exemple d'un tableau à deux dimensions

| Nombre d'enregistrements | Population de Samplonie | | | | |
|---|---|---|---|---|---|
| | Total | Emploi | | Sexe | |
| | | Oui | Non | M | F |
| Total | 1,000 | 341 | 659 | 511 | 489 |
| Agria | 293 | 121 | 172 | 145 | 148 |
| Wheaton | 144 | 60 | 84 | 70 | 74 |
| Greenham | 94 | 38 | 56 | 44 | 50 |
| Newbay | 55 | 23 | 32 | 31 | 24 |
| Induston | 707 | 220 | 487 | 366 | 341 |
| Oakdale | 61 | 26 | 35 | 36 | 25 |
| Crowdon | 244 | 73 | 171 | 128 | 116 |
| Smokeley | 147 | 49 | 98 | 80 | 67 |
| Mudwater | 255 | 72 | 183 | 122 | 133 |

Source: Bureau statistique de Samplonie.

Le tableau ci-dessus contient de simples fréquences, mais Abacus peut calculer aussi des totaux de variables quantitatives et produire des tableaux de pourcentages et des tableaux de moyennes. De plus, les cases d'un tableau produit par Abacus peuvent contenir plus d'un élément à la fois (de fait, elles peuvent en contenir jusqu'à 10). L'utilisateur doit alors décider de la présentation du tableau (c.-à-d. classer les éléments par ligne, par colonne ou par couche). Si des données ont été recueillies au moyen d'une enquête par sondage, Abacus peut être appliqué à des données pondérées; il peut s'agir, par exemple, de poids calculés au moyen du programme Bascula. L'utilisateur n'a plus qu'à définir la variable qui contient les poids.

On accorde beaucoup d'attention à la présentation du tableau car les tableaux produits par Abacus doivent être prêts à la photo. Abacus renferme donc de nombreuses options qui permettent de déterminer la présentation. Ainsi, on peut définir jusqu'à 10 lignes de texte pour l'en-tête et le cartouche d'un tableau et choisir soit des lignes horizontales et verticales (comme dans l'exemple ci-dessus), soit des lignes horizontales seulement ou ni l'une ni l'autre. La disposition du texte dans les en-têtes de colonne et la largeur des colonnes peuvent aussi être modifiées.

On peut procéder à un arrondissement afin de conserver le caractère confidentiel des données du tableau. Non seulement les totaux de case, mais aussi les totaux marginaux sont arrondis au multiple d'une constante donnée, par ex.: 5. Abacus exécute aussi bien l'arrondissement normal que l'arrondissement aléatoire. Si l'utilisateur n'est pas satisfait du tableau obtenu, il peut importer le produit d'Abacus dans le tableur Lotus 123 et pousser plus loin le traitement. Une dernière option qui mérite d'être soulignée ici est la possibilité de créer de nouvelles variables en récrivant les variables existantes (par ex.: âge groupe d'âge). Pour plus de détails sur le programme Abacus, voir Bethlehem et coll. (1989a).

## 9. ANALYSE

Le BCS n'a pas élaboré de logiciel pour l'analyse des données statistiques principalement parce que le marché compte suffisamment de bons progiciels statistiques. Le BCS utilise lui-même les progiciels SPSS (sur gros ordinateur et sur micro-ordinateur) et Stata (sur micro-ordinateur). Pour que ces progiciels puissent faire partie du système intégré de traitement des

données d'enquête, nous devons trouver des "fonctions" qui nous permettrons d'exporter des données du système Blaise vers ces progiciels. Cela se fait en deux étapes. Premièrement, le fichier de données, en format Blaise, est converti en format ASCII et deuxièmement, l'information relative aux variables, qui est consignée dans le questionnaire Blaise, est traduite de manière à pouvoir être comprise par le progiciel en question. Un fichier provisoire est donc créé. En exécutant ce fichier à partir du progiciel statistique, on se trouve à créer un fichier informatique. Et en chargeant ce dernier, l'utilisateur peut débuter immédiatement son analyse sans se préoccuper de définir les variables, les étiquettes, *etc.*

La procédure que nous venons de décrire ne peut être réalisée qu'avec les progiciels SPSS et Stata et aucun autre. Nous pourrions, bien sûr, élaborer la même procédure pour chaque progiciel statistique connu mais cela exigerait un long travail de programmation. Nous avons choisi une autre solution. Le système Blaise contient un utilitaire spécial, générateur de fichiers provisoires. L'utilisateur "colore" la structure du fichier provisoire dans un logiciel de traitement de texte et exécute le générateur de fichiers provisoires en utilisant comme données d'entrée la description générale qu'il vient de faire du fichier provisoire et le questionnaire Blaise; il obtient ainsi un vrai fichier provisoire. Le générateur de fichiers provisoires permet donc à l'utilisateur de créer des fichiers provisoires pour les progiciels de son choix.

## 10. CONCLUSION

L'arrivée du micro-ordinateur a eu une incidence considérable sur les opérations des organismes nationaux de statistique. Le statisticien spécialisé recourt de plus en plus au micro-ordinateur et a besoin, pour son travail, d'un système intégré de traitement des données d'enquête comme celui fondé sur le système Blaise. Ce dernier se distingue d'une part par la cohérence qu'il introduit dans les diverses étapes de la collecte et du traitement des données. Il est ainsi plus facile de gérer et de contrôler le processus dans son entier. D'autre part, le système Blaise favorise l'uniformisation entre les diverses divisions. Comme toutes les divisions utilisent le même logiciel pour leurs enquêtes, l'échange de renseignements se fait plus facilement et il y a un moins grand risque d'erreur.

Le système intégré de traitement des données d'enquête est conçu plus spécialement pour des organisations très centralisées, comme le Bureau central de la statistique des Pays-Bas. Dans une organisation de cette envergure, le système intégré de traitement peut amener un accroissement substantiel de l'efficacité. Or, les organismes statistiques n'ont pas tous une structure centralisée. Dans les grands pays en particulier, le traitement des données est souvent décentralisé. Les bureaux régionaux s'occupent du traitement des données dans leurs régions respectives et font parvenir ensuite les fichiers constitués à l'administration centrale. Celle-ci combine les fichiers régionaux en un seul fichier national. Le système intégré de traitement des données d'enquête peut trouver sa place dans un telle structure: l'administration centrale élabore le questionnaire Blaise et envoie à chaque bureau régional des exemplaires du programme de saisie des données ainsi élaboré. De cette manière, on assure la cohérence des opérations de collecte et de contrôle des données dans les diverses régions. Comme les fichiers régionaux auront tous le même format (Blaise), il sera facile de les combiner en un seul fichier national en se servant des fonctions du système Blaise. De plus, comme ces fichiers seront tous "épurés", il ne sera pas nécessaire d'exécuter d'autres contrôles au niveau national. L'administration centrale n'aura plus qu'à analyser les données et les mettre en tableau et à publier les résultats de ses analyses. En outre, les bureaux régionaux ou l'administration centrale pourront publier des données régionales.

Le système Blaise est en usage depuis le milieu de 1986 et sert à un très grand nombre d'enquêtes. On le perfectionne en étroite collaboration avec les utilisateurs et chaque nouvelle version renferme des améliorations. Abacus est en usage depuis plus d'un an et est devenu très populaire auprès des utilisateurs. Quant au programme Bascula, il est encore en voie d'élaboration. On vient juste de lancer le premier prototype.

Nous n'avons pas parlé de la possibilité d'exporter des données et des métadonnées du système Blaise vers le progiciel de base de données Paradox. Nous projetons aussi d'établir un lien entre Blaise et le système de gestion de base de données Oracle. Ainsi, il sera possible de créer une architecture client-serveur pour les utilisateurs de Blaise.

À l'extrémité de la chaîne de production des données statistiques, il y a encore quelques aspects de la publication qui méritent notre attention. En premier lieu, il faudra élaborer un logiciel qui permettra d'évaluer les risques de divulgation d'information confidentielle qui accompagnent la publication de données statistiques. On offrira aussi des "fonctions" qui permettront de mettre les tableaux ou les fichiers à l'abri de ces risques. Enfin, les organismes statistiques se livrent de plus en plus à la publication électronique de données statistiques (par ex.: information sur disquette, sur disque compact-ROM, *etc.*). Pour aider les utilisateurs de ce genre de support à choisir les séries de données dont ils ont besoin, il est nécessaire de mettre à leur disposition des logiciels conviviaux. C'est ce à quoi nous nous occupons présentement.

## BIBLIOGRAPHIE

BETHLEHEM, J.G., VAN BUITENEN, A.A.A., HUNDEPOOL, A.J., ROESINGH, M.J., et VAN DE WETERING, A. (1989a). Abacus 1.0, A Tabulation Package, Compact Guide. CBS report, Voorburg: Netherlands Central Bureau of Statistics.

BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., et VERMEULEN, L.F.M. (1989b). Blaise 2.0/An Introduction. CBS report, Voorburg: Netherlands Central Bureau of Statistics.

BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., et VERMEULEN, L.F.M. (1989c). Blaise 2.0/Language Reference Manual. CBS report, Voorburg: Netherlands Central Bureau of Statistics.

BETHLEHEM, J.G., et KELLER, W.J. (1987). Linear Weighting of Sample Survey Data. *Journal of Offical Statistics* 3, 141-154.

KELLER, W.J., BETHLEHEM, J.G., et METZ, K.J. (1990). The impact of micromputers on survey processing at the Netherlands Central Bureau of Statistics. *Proceedings of 1990 Annual Research Conference, U.S. Bureau of the Census*, 637-645.

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 63–75
Statistique Canada



# Recherche et essais pour les méthodes d'enquêtes par téléphone à Statistique Canada

J. DOUGLAS DREW[1]

RÉSUMÉ

On présente les résultats de la recherche et des essais des méthodes d'enquêtes par téléphone et assistées par ordinateur pour les enquêtes-ménages, que suit une discussion de la façon dont ces conclusions vont influencer le remaniement des enquêtes-ménages à Statistique Canada pendant les années 90. On s'attache plus particulièrement dans la communication à l'enquête sur la population active du Canada.

MOTS CLÉS: Collecte des données; enquêtes-ménages; plan de sondage.

## 1. INTRODUCTION

Les enquêtes ont subi au cours des années 80 d'importants changements en raison des progrès de la technologie et de l'élaboration de méthodes d'enquête par téléphone modernes et il est probable que le rythme de ces changements va s'accélérer pendant les années 90. Dans cette communication, nous décrivons les recherches, les essais et l'élaboration des méthodes qui constitueront l'infrastructure des activités de collecte des données pour les enquêtes-ménages de Statistique Canada pendant les années 90. Cette recherche s'est attachée en particulier à l'enquête sur la population active du Canada et a été menée de 1985 à 1989 afin d'identifier les améliorations à mettre en oeuvre pendant le remaniement post-censitaire de 1991 de l'enquête.

## 2. PROGRAMME DE RECHERCHE ET D'ESSAIS DE L'EPA

L'enquête sur la population active (EPA) du Canada est l'enquête-ménage la plus importante de Statistique Canada, avec un échantillon de 62,300 ménages chaque mois. Elle utilise un panel rotatif dans lequel les ménages restent dans l'échantillon pendant six mois consécutifs, avant d'en être retirés. Elle utilise un sondage aréolaire multiple, avec un personnel décentralisé de 1,000 intervieweurs locaux situés au Canada et faisant rapport à un des cinq bureaux régionaux.

Jusqu'au début des années 70, toutes les interviews se faisaient en personne. En 1972, on a mis en oeuvre les interviews téléphoniques dans les grandes régions urbaines pour les interviews de suivi auprès des ménages, après que ces derniers aient passé l'interview personnelle pendant leur premier mois dans l'échantillon. On appelle ce suivi téléphonique dans la littérature "enquête téléphonique à chaud" pour la distinguer des "enquêtes téléphoniques à froid", lorsque l'interview téléphonique n'est pas précédée d'une interview personnelle (Groves et coll. 1988).

L'enquête téléphonique à chaud était au début limitée aux principales régions urbaines en raison du nombre de lignes groupées dans les régions urbaines et rurales plus petites et des préoccupations entourant la confidentialité des données recueillies. Toutefois, pendant le remaniement de l'enquête de 1981, on a essayé des interviews téléphoniques à chaud pour les petites régions urbaines et rurales, mais on a constaté que les répondants étaient prêts à être interviews par téléphone et que la procédure n'avait pas d'incidence sur les taux de réponse ou les estimations d'enquêtes (Choudhry 1984). L'extension de l'interview téléphonique à ces régions en 1984 s'est traduite par une baisse de 10% des coûts de collecte des données de l'enquête.

[1] J. Douglas Drew, directeur adjoint, Division des enquêtes-ménages, Statistique Canada, Ottawa, Ontario, K1A 0T6.

En 1985, après la mise en oeuvre d'un échantillon remanié, on a entrepris un programme de recherche afin d'étudier quelles améliorations pour la collecte des données pourraient résulter i) d'un accroissement de l'utilisation du téléphone pour la collecte, ii) des méthodes d'enquête par téléphone où le téléphone sert à l'échantillonnage et à la collecte des données, et iii) des méthodes d'interview assistée par ordinateur (IAO).

Dans l'étude des méthodes téléphoniques et IAO, on a constaté qu'il était utile de caractériser le plan de sondage en termes d'un certain nombre de facteurs comme suit:

i) **Mode de collecte.** Comme on l'a déjà dit, le mode de collecte actuel est l'enquête téléphonique à chaud, avec une interview personnelle initiale et des interviews essentiellement téléphoniques au cours des mois suivants. D'autres modes de collecte comprennent les interviews téléphoniques à froid avec un suivi personnel des non-répondants téléphoniques et les interviews téléphoniques à froid sans suivi personnel.

ii) **Organisation des intervieweurs.** L'EPA a actuellement une organisation locale, et les intervieweurs locaux font un mélange d'interviews personnels et d'interviews téléphoniques à partir de leur maison. Une autre possibilité est une organisation centrale, les intervieweurs travaillant à partir d'une ou de plusieurs villes centrales; dans le cas de Statistique Canada, il s'agit de ses cinq bureaux régionaux au pays. Un troisième modèle organisationnel est un modèle combiné, dans lequel l'interview se fait par une combinaison d'intervieweurs locaux et centraux.

iii) **Technologie.** La technologie actuelle de l'enquête est le papier et le crayon. L'autre technologie que l'on considère plus valable pour les enquêtes-ménages est l'interview assistée par ordinateur. L'IAO est habituellement appelée ITAO (interview téléphonique assistée par ordinateur) lorsqu'elle est faite de façon centrale et IPAO (interview personnelle assistée par ordinateur) lorsqu'elle est faite localement en utilisant des ordinateurs portables que les intervieweurs utilisent pour les interviews personnelles et pour les interviews téléphoniques à partir de leur maison.

iv) **Base de sondage et plan de sondage.** L'enquête utilise un plan de sondage aréolaire depuis son délait, en 1945. Les autres possibilités comprennent les bases de sondage téléphoniques utilisant les méthodes de composition aléatoire (Waksberg 1978) ou une combinaison de composition aléatoire et de bases de sondage des numéros de téléphone publiés; les autres bases de sondage, contrairement aux listes téléphoniques, sont conceptuellement complètes, et les méthodes de bases de sondage duales combinant deux ou plusieurs des options ci-dessus.

Dans les sections qui suivent, on examine les résultats du programme de recherche et d'essais pour chacun de ces facteurs.

## 3. MODE DE COLLECTE

On a procédé à un important essai de 1985 à 1989 pour déterminer l'incidence des interviews téléphoniques à froid avec suivi personnel comme solution de remplacement aux interviews téléphoniques à chaud actuelles. Cet essai s'appelle l'essai téléphonique de la première interview. Il a été incorporé dans l'EPA courante dans les régions urbaines du Québec et de l'Ontario. Drew, Choudhry et Hunter (1988) en ont exposé la méthodologie détaillée. En bref, les logements EPA nouvellement échantillonnés sont appariés aux listes achetées des compagnies téléphoniques à partir de l'adresse. On a obtenu des taux d'appariement de 65%. On a choisi des échantillons d'essai et de contrôle à partir des logements appariés de telle sorte que chaque logement d'essai a été couplé à un logement de contrôle à l'intérieur de la même unité d'échantillonnage (îlot urbain). Pour les logements tests, on a donné les numéros de téléphone aux intervieweurs qui devaient essayer de mener une interview téléphonique, mais faire appel à un suivi personnel au besoin. Les intervieweurs n'étaient pas au courant de l'existence d'un

**Tableau 1**

Test de la première interview téléphonique (octobre 1985 - mars 1989)
Estimation du traitement du test en pourcentage du traitement de l'estimation de contrôle

| Caractéristique | Ontario | | Québec | |
|---|---|---|---|---|
| | Pourcentage | Statistique t | Pourcentage | Statistique t |
| Emploi | 98.5 | −1.22 | 97.2 | −1.64 |
| Chômage | 96.3 | −0.94 | 111.8 | 2.27* |
| Inactifs | 101.1 | 0.88 | 98.2 | −0.63 |
| Pop. 15+ | 909.2 | −0.86 | 98.5 | −1.31 |
| Pop. ménages = 3+ | 97.8 | −0.76 | 94.1 | −1.26 |
| Pop. ménages 1 personne | 100.6 | 0.25 | 104.1 | −0.93 |
| Pop. ménages 2 personnes | 101.1 | 0.26 | 101.0 | 0.43 |
| Pop. ménages 3 personnes | 98.6 | −0.56 | 94.7 | −1.28 |
| Hommes occupés 15-24 | 95.2 | −0.75 | 88.7 | −2.00* |
| Hommes occupés 25+ | 98.5 | −1.11 | 96.7 | −1.54 |
| Femmes occupées 15-24 | 99.0 | −0.11 | 111.5 | 1.44 |
| Femmes occupées 25+ | 99.2 | −0.65 | 97.1 | −1.11 |
| Hommes chômeurs 15-24 | 105.6 | 0.53 | 119.0 | 1.53 |
| Hommes chômeurs 25+ | 94.5 | −0.80 | 96.7 | −0.31 |
| Femmes chômeurs 15-24 | 99.6 | −0.15 | 119.4 | 1.30 |
| Femmes chômeurs 25+ | 90.9 | −1.22 | 123.9 | 2.71** |
| Inactifs 15-24 | 99.8 | −0.07 | 99.9 | 0.47 |
| Inactifs 25+ | 105.6 | 1.57 | 101.3 | 0.07 |
| Inactives 15-24 | 101.9 | 0.56 | 95.3 | −0.56 |
| Inactives 25+ | 99.2 | −0.14 | 97.3 | −0.92 |
| Hommes 15-24 | 97.2 | −0.49 | 95.2 | −0.73 |
| Hommes 25+ | 99.9 | −0.09 | 97.7 | −1.49 |
| Femmes 15-24 | 99.9 | 0.16 | 105.4 | 0.95 |
| Femmes 25+ | 98.9 | −1.28 | 98.4 | −1.24 |

* Statistique t significative au niveau de 5%.
** Statistique t significative au niveau de 1%.

échantillon de contrôle et ont suivi les procédures habituelles pour tous les logements dont le numéro de téléphone n'avait pas été communiqué.

On n'a pas trouvé de différence significative des taux de réponse entre les échantillons de test et de contrôle. Pour le Québec, les taux de réponse étaient de 96.1% pour les deux échantillons, tandis que pour l'Ontario, le taux de 96.3% pour l'échantillon test était très légèrement inférieur à celui de 96.5% pour l'échantillon de contrôle.

La comparaison des estimations de la population active obtenues dans les échantillons de test et de contrôle a permis de constater que certaines estimations du Québec au début du test pour la période octobre 1985 - février 1987 étaient significativement différentes. En particulier, le nombre d'hommes occupés et de chômeurs dans les ménages de trois personnes ou plus était sous-estimé dans l'échantillon de test (Drew, Choudhry et Hunter 1988). Le tableau 1 présente les données pour l'ensemble de la période des tests (octobre 1985 - mars 1989). Pour le Québec, quelques différences statistiquement significatives existent, résultant de l'influence de la période antérieure. Lorsqu'on a analysé les données à compter de mars 1987, ces différences ne sont pas significatives. En Ontario, on n'a relevé aucune différence significative. S'agissant des différences pour le Québec, leur coïncidence avec un programme d'inspection des bénéficiaires du Bien-être social effectué par l'administration provinciale semble indiquer que des mesures

**Tableau 2**

Taux de non-réponse: interviews téléphoniques à chaud et interviews téléphoniques à froid avec/sans suivi personnel

| Méthode | Test 1 | Test 2 |
|---|---|---|
| Interviews téléphoniques à chaud avec lettre (EPA courante) | 4.1 | 5.6 |
| Interviews téléphoniques à froid avec lettre | 6.9 | 9.8 |
| Interviews téléphoniques sans lettre | 8.5 | - |

Test 1: Octobre 1985 - septembre 1986, Ontario et Québec.
Test 2: Juillet 1988 - mars 1989, Nouvelle-Écosse et Alberta.

exogènes à l'enquête se sont traduites par un contexte dans lequel les interviews téléphoniques à froid ont paru plus suspectes. Nous avons été heureux d'avoir pu effectuer le test pendant cette période, parce que la conclusion que les résultats de l'enquête obtenus par des interviews téléphoniques à froid sont davantage sujets à des influences extérieures que celles obtenues par des interviews téléphoniques à chaud sera une considération importante dans toute décision concernant l'extension des interviews téléphoniques.

On a également étudié les interviews téléphoniques à froid sans suivi personnel. Deux tests d'échantillonnage par téléphone ont été effectués, au cours desquels l'EPA a été exécutée comme une enquête téléphonique centrale avec des interviews à partir des bureaux régionaux. Le tableau 2 présente les taux de non-réponse pour les deux tests et les taux comparables pour l'EPA courante.

Le premier test a étudié deux méthodes d'échantillonnage: la CTA et une combinaison d'échantillonnage de liste pour les numéros publiés et la CTA pour les numéros non publiés. L'échantillonnage par liste contenait des lettres d'introduction, mais l'échantillonnage CTA non. Les différences des taux de réponse semblent mettre en évidence les effets positifs sur les taux de réponse d'une lettre préalable. Pour les deux tests, la comparaison des deux types d'interviews téléphoniques a révélé que les taux de non-réponse pour les interviews téléphoniques à froid étaient plus élevés, au niveau de signification de 5%. Le deuxième test a utilisé uniquement un échantillon de listes de numéros publiés.

Un problème important est le biais de non-réponse éventuelle, imputable à la non-réponse plus élevée caractéristique des interviews téléphoniques à froid sans suivi personnel. Comme approximation de ces non-répondants supplémentaires, Laflamme (1990) a considéré les ménages autres que ceux du premier mois de l'échantillon de l'EPA courante qui avaient un téléphone mais qui étaient interviewés de façon personnelle. Il avait constaté que la taille du groupe choisi, c'est-à-dire 3.5% des répondants, était proche de celle des cas de non-réponse supplémentaires pour les interviews téléphoniques à froid sans suivi personnel. De plus, il avait constaté que le taux de chômage du groupe de remplacement était de 12.8%, comparativement à 7.4% pour les personnes dans les ménages interviewés par téléphone. L'exclusion du groupe de remplacement de l'échantillon aurait réduit le taux de chômage national de 8.1% à 7.9%. Ceci constitue visiblement un biais sérieux, compte tenu de la précision nécessaire pour les estimations nationales de l'EPA. Comme l'hypothèse de remplacement semble raisonnable, ces conclusions soulèvent des préoccupations graves à propos des interviews téléphoniques à froid sans suivi personnel pour l'EPA.

Le tableau 3 compare les taux de chômage et d'activité pour le premier test d'échantillonnage téléphonique aux estimations correspondantes des ménages téléphoniques de l'EPA. La seule estimation qui devait se révéler significativement différente au niveau de 5% des estimations produites pour la population téléphonique de l'EPA courante était le taux de chômage pour le Québec pour le traitement CTA. Il est intéressant de noter que le test a été effectué en même temps que des problèmes apparaissaient pour les estimations du Québec dans le test de la première interview téléphonique. Un autre point intéressant est que si les autres différences

**Tableau 3**

Test de l'échantillonnage téléphonique (Octobre 1985 - septembre 1986)
Taux de chômage et d'activité

| Province | Plan | Taux de chômage | (D.S.) | Taux d'activité | (D.S.) |
|---|---|---|---|---|---|
| Québec | LISTE | 12.3 | (0.78) | 64.1 | (1.08) |
| | CTA | 13.0* | (0.88) | 62.8 | (1.28) |
| | EPA | 10.9 | (0.27) | 63.4 | (0.29) |
| Ontario | LISTE | 7.3 | (0.59) | 69.0 | (1.11) |
| | CTAI | 7.9 | (0.63) | 69.0 | (1.18) |
| | EPA | 6.9 | (0.16) | 69.0 | (0.20) |

* Différence significative entre les taux de chômage CTA et EPA pour le Québec.

des taux de chômage n'étaient pas statistiquement significatives, les taux pour les interviews téléphoniques à froid étaient plus élevés. D'autres chercheurs ont observé des différences dans le même sens, mais également n'ont pu être en mesure d'attribuer une signification statistique à celles-ci. Ces données pourraient bénéficier d'une méta-analyse.

En résumé, les résultats du test révèlent que les interviews téléphoniques à froid sans suivi personnel donnent des taux de non-réponse plus élevés que la méthode actuelle des interviews téléphoniques à chaud, et même si cela n'est pas concluant, il y a des indices qu'il donne des taux de chômage plus élevés. Par contre, les interviews téléphoniques à froid avec suivi personnel, à l'exception de la seule période pour le Québec, ont donné des résultats comparables à ceux des interviews téléphoniques à chaud.

Compte tenu de ces résultats, on a décidé de mettre en oeuvre les interviews téléphoniques à froid avec suivi personnel pour l'échantillon d'appartements EPA qui représentent approximativement 4% de l'échantillon total. On a estimé que l'existence des numéros de téléphone des unités de la base de sondage des appartements aiderait à surmonter les problèmes de l'accès aux immeubles à étages multiples et permettrait de procéder à plus de tentatives de trouver les personnes à la maison que pour les interviews personnelles. Il semble que ces attentes se soient réalisées. Comme le signale Dufour (1990), alors que les taux de non- réponse du premier mois pour l'échantillon des appartements continuent de dépasser ceux du premier mois pour l'échantillon hors appartements, l'écart s'est réduit, passant d'une différence de 8.7 points au cours de l'année avant le changement à une différence de 5.7 points au cours des cinq premiers mois d'utilisation de la nouvelle procédure.

Un autre changement au mode de collecte pour l'EPA courante a été le suivi téléphonique des ménages du premier mois dans l'échantillon qui ne pouvaient contactés au cours d'une visite initiale. Cette procédure a été instaurée en 1986 et s'est traduite par une économie de $100,000 dans les coûts de collecte des données.

L'effet combiné de la première interview téléphonique pour les appartements et du suivi téléphonique des non-répondants du premier mois a été une hausse du taux des interviews téléphoniques globales de l'enquête, de 80% en 1985 à 83% en 1990.

## 4. ORGANISATION DES INTERVIEWEURS

Pendant le programme des tests, on a étudié deux solutions de rechange à l'organisation locale actuelle des intervieweurs. Les tests des échantillons téléphoniques déjà décrits envisageaient une organisation "centrale" dans laquelle les bureaux régionaux faisaient toutes les interviews. Un autre test portait sur une organisation mixte, dans laquelle le mode de collecte

actuel des interviews téléphoniques à chaud était conservé. L'organisation mixte a fait l'objet d'un essai entre janvier 1988 et mars 1989 dans deux régions métropolitaines de recensement où se trouvaient des bureaux régionaux, Montréal et Halifax. Son premier objectif était de mesurer l'élément coût d'une telle organisation.

La méthodologie des tests consistait en une collecte personnelle par les intervieweurs locaux pour les ménages du premier mois dans l'échantillon et par des interviews téléphoniques par le personnel central travaillant dans les bureaux régionaux pour la plupart des autres cas. Lorsqu'un suivi de non-réponse était nécessaire pour les ménages affectés au départ aux intervieweurs centraux, celui-ci était effectué par les intervieweurs locaux. Cette méthodologie avait été testée au départ pour l'enquête sur la population active par Muirhead et coll. (1975) et a fait l'objet d'études poussées par l'United States Bureau of the Census (1987), où les interviews centrales se font grâce à des interviews téléphoniques assistées par ordinateur (ITAO).

Une des complexités de la méthode était la pratique reconnue pour la première moitié du test de transférer des cas nécessitant un suivi de non-réponse des intervieweurs centraux aux intervieweurs locaux au milieu de la semaine d'interview. Pour la deuxième moitié du test, ce recyclage était limité aux cas où le numéro de téléphone n'était plus valide. Pendant la première moitié du test, les taux de non-réponse étaient de 8.0% pour le traitement de test contre 6.1% pour les procédures de contrôle correspondant à l'interviews décentralisée utilisée pour l'EPA courante. L'écart s'est réduit à 7.3% contre 6.7% pendant la deuxième moitié.

À partir du premier test d'échantillonnage téléphonique, on a estimé que les coûts de l'interview par ménage étaient de $2.72 pour la collecte centrale de données avec un échantillonnage de listes téléphoniques, contre $3.53 pour l'échantillonnage CTA. Les coûts supplémentaires des méthodes CTA s'expliquent par le temps consacré au tri des numéros de téléphone résidentiels. Ces coûts comprennent $0.46 par ménage pour les appels interurbains. On a estimé ce montant à partir des taux interurbains et des données sur la durée des appels, puisque les pratiques de tenue des livres dans les bureaux régionaux ne permettent pas l'obtention des coûts réels. Les coûts comparables pour l'EPA courante sont de $4.76 par ménage pour les frais d'intervieweurs et les dépenses. Le test de l'organisation mixte a donné une économie par rapport à l'EPA courante de $0.78 par ménage en frais d'intervieweurs et dépenses. La comparaison des coûts ci-dessus ne prend pas en compte le coût des locaux à bureaux et du matériel dans les coûts des organisations centralisées et mixtes. Elle ne prend également pas en compte le coût du transfert des documents en direction et en provenance des intervieweurs locaux dans les organisations mixtes et locales, ce qui, dans la technologie actuelle du papier et du crayon, se fait par l'envoi par messagerie des documents, mais se ferait de façon électronique dans les options CTA.

On a envisagé l'organisation mixte pour les villes avec bureau régional seulement, puisque l'extension de l'une de ces dernières se traduirait par une utilisation accrue des appels téléphoniques interurbains. Qui plus est, le plan de sondage des régions urbaines et rurales plus petites est en grappes, de sorte que les unités primaires d'échantillonnage donnent des tailles d'échantillon qui correspondent à une tâche d'intervieweur. La centralisation de la partie téléphonique de l'échantillon nécessiterait une mise en grappes plus forte de l'échantillon afin de conserver une charge de travail suffisante pour les intervieweurs locaux. Par ailleurs, dans les centres urbains de taille moyenne, où il y a actuellement de quatre à cinq intervieweurs, le nombre d'intervieweurs locaux dans une organisation mixte serait ramené à un ou deux, ce qui réduirait sensiblement la marge de manoeuvre quand les intervieweurs devront remplacer les autres pendant les vacances et les maladies.

Tous les trois modèles organisationnels considérés ont des avantages et des inconvénients. L'organisation locale a les taux de non-réponse les plus bas, mais le coût de collecte unitaire des données le plus élevé. L'organisation mixte a une non-réponse un peu plus élevée et des coûts un peu plus bas, et se trouve limitée dans son champ d'application. En fin de compte, on a exprimé que l'organisation mixte entraînait bien des complexités pour un gain tout au plus minime. L'organisation centrale, qui se caractérise par des économies appréciables des

coûts de collecte des données, donne une augmentation de 68-75% de la non-réponse par rapport à l'organisation locale. Comme on l'a vu en 3), il y a des indices que cette non-réponse supplémentaire introduirait un biais de non- réponse sérieux dans les estimations EPA. De plus, on s'inquiète de ce que l'écart des taux de non-réponse entre les organisations locales et centrales pourrait s'élargir à l'avenir, puisque une sollicitation téléphonique plus forte et la disponibilité accrue des techniques de tri téléphonique rendent la population moins réceptive aux interviews téléphoniques. Pour ces raisons, on favorise des stratégies de plan de sondage qui, même si elles sont souples pour ce qui est des appels téléphoniques, peuvent permettre aussi des suivis personnels au besoin. L'organisation locale est celle qui offre la flexibilité la plus grande. Pour toutes ces considérations, on a décidé de conserver l'organisation locale actuelle.

## 5. TECHNOLOGIE

Catlin, Ingram et Hunter (1988) ont effectué une étude contrôlée comparant les méthodes ITAO et celle du papier et du crayon. Dans l'étude, on a soumis le questionnaire EPA à des échantillons CTA de 1,000 ménages par mois par traitement sur une période de neuf mois. Toutes les interviews ont eu lieu à partir du siège central de Statistique Canada à Ottawa.

L'étude faisait partie d'un travail de recherche en collaboration avec le United States Bureau of the Census (USBC) et le logiciel ITAO utilisé a été mis au point par ce dernier. On a délibérément retenu le même libellé des questionnaires pour les deux approches. Les caractéristiques propres à l'approche ITAO sont le branchement automatique, quelques vérifications en direct sur deux bases et l'acheminement automatique des appels.

On a discerné trois améliorations qualitatives pour l'ITAO par rapport à l'autre méthode. D'abord, le taux général de rejet à la vérification pendant le traitement post-collecte des données était inférieur de 50% pour l'ITAO. Ensuite, il y a eu une disparition pratiquement totale des erreurs de branchement. Ce qui est important, c'est que cette baisse s'observe pour certaines parties du questionnaire, qui, même si elles sont rarement utilisées, ont une incidence sur la détermination de la situation vis-à-vis de l'activité, et pour laquelle des intervieweurs utilisant le papier et le crayon font des erreurs de branchement importantes. Enfin, la taille du ménage moyen déclarée pour l'ITAO était supérieure de 3%, ce qui représente en gros une réduction de 50% du sous-dénombrement de l'EPA par rapport au recensement. Cette amélioration semble provenir du mécanisme de contrôle incorporé dans l'instrument ITAO pour les membres supplémentaires du ménage et pour les personnes temporairement absentes.

Compte tenu de ces résultats, on a décidé que la mise en oeuvre des interviews assistées par ordinateur serait une des principales caractéristiques du remaniement post-censitaire de 1991 de l'EPA. En raison de la préférence pour le maintien d'une organisation d'intervieweurs locale, on prévoit une mise en oeuvre IPAO.

## 6. BASE DE SONDAGE ET PLAN DE SONDAGE

### Base de sondage téléphonique

La couverture téléphonique et la mesure dans laquelle des caractéristiques des personnes sans téléphone diffèrent de celles de ceux qui en ont un sont des facteurs importants dans l'étude de méthodes d'enquêtes téléphoniques, en particulier pour ce qui concerne les stratégies de la base de sondage.

Dans une revue internationale de la couverture téléphonique, Trewin et Lee (1988) ont constaté que la couverture téléphonique au Canada est l'une des plus élevées du monde, avec 97-98%. Comme c'est la règle générale dans la plupart des pays qu'ils ont étudiés, les personnes vivant dans des ménages sans téléphone au Canada ont tendance à avoir des revenus plus bas et des taux de chômage plus élevés.

**Tableau 4**

Ménages sans téléphone par province (%)

| | 1976 | 1981 | 1985 | 1987 | 1990 |
|---|---|---|---|---|---|
| Canada | 3.5 | 2.4 | 1.8 | 1.5 | 1.5 |
| Terre-Neuve | 10.0 | 6.0 | 5.1 | 3.6 | 1.9 |
| Île-du-Prince-Édouard | – | – | – | – | 2.8 |
| Nouvelle-Écosse | 7.5 | 4.6 | 3.5 | 3.2 | 1.5 |
| Nouveau-Brunswick | 5.8 | 5.3 | 5.3 | 3.3 | 2.2 |
| Québec | 3.3 | 2.1 | 1.6 | 1.5 | 1.5 |
| Ontario | 2.5 | 1.9 | 1.0 | 1.0 | 1.2 |
| Manitoba | 4.1 | 2.3 | 2.7 | 2.4 | 1.7 |
| Saskatchewan | 3.6 | 2.5 | 2.3 | 2.4 | 2.3 |
| Alberta | 3.0 | 2.4 | 2.0 | 1.8 | 2.0 |
| Colombie-Britannique | 4.2 | 2.8 | 2.4 | 1.3 | 1.5 |

Source: Statistique Canada, estimations provenant de l'enquête sur les installations et l'équipement ménagers.

**Tableau 5**

Caractéristiques de la population active selon le statut téléphonique

| Province | Statut téléphonique | Taux de chômage | Taux d'activité |
|---|---|---|---|
| Nouvelle-Écosse | Publié | 9.0 | 71.9 |
| | Non publiés | 9.8 | 70.2 |
| | Pas de téléphone | 17.2 | 62.3 |
| Alberta | Publiés | 6.3 | 80.7 |
| | Non publiés | 8.2 | 81.5 |
| | Pas de téléphone | 11.1 | 67.0 |

Le tableau 4 donne le pourcentage des ménages sans téléphone au Canada de 1976 à nos jours. La couverture téléphonique, bien que déjà élevée en 1976, a continué de s'étendre, même s'il semble qu'elle se soit stabilisée au cours des dernières années à environ 98.5%.

Laflamme (1990) a entrepris une étude comparative des caractéristiques des univers sans téléphone et avec téléphone. L'enquête comprenait une ventilation de ceux ayant un numéro publié et de ceux n'ayant pas de numéro publié, obtenus par le raccordement des numéros de téléphone fournis par les répondants EPA à des listes de numéros téléphoniques publiés. Deux provinces ont été incluses dans l'étude, la Nouvelle-Écosse et l'Alberta. On devait constater que pour la Nouvelle-Écosse et l'Alberta, 9.7% et 11.9% respectivement des numéros n'étaient pas publiés. Les taux de chômage et d'activité calculés par Laflamme figurent au tableau 5.

Cette étude reproduit les conclusions d'études antérieures selon lesquelles les caractéristiques de la population active des personnes sans téléphone sont très différentes de celles qui en ont un. Les caractéristiques de la population active diffèrent, mais a un degré moindre, entre les personnes qui ont des numéros publiés et celles qui ont des numéros non publiés.

Bien que la population sans téléphone représente 1.0-1.5% de la population seulement, les différences dans les caractéristiques de la population active sont suffisamment importantes pour que la simple exclusion de la population sans téléphone ne soit pas une option valable pour l'EPA compte tenu de la précision nécessaire pour les estimations nationales de l'emploi et du chômage (coefficients de variation de 0.5% et de 2% respectivement).

Un autre inconvénient des bases de sondage téléphoniques, en particulier pour les enquêtes par panel, est qu'elles se détériorent rapidement. Drew, Dick et Switzer (1989) ont trouvé un taux d'addition et de suppression de 0.5 - 1.0% au stock des numéros résidentiels publiés par mois. Les échantillons téléphoniques ne peuvent donc rester représentatifs de l'univers téléphonique pendant très longtemps à moins d'être mis à jour. Les auteurs proposent une stratégie de mise à jour des échantillons d'une enquête par panel en utilisant pour cela des fichiers de numéros publiés acquis sur une base permanente auprès des compagnies de téléphone. Un test opérationnel de la procédure sur une période de neuf mois s'est révélé concluant. Leur procédure ne porte que sur les numéros publiés échantillonnés seulement provenant d'une base de sondage par listes et ne constitue pas une solution au problème de la tenure à jour d'un échantillon utilisant les méthodes de composition téléphonique aléatoire pendant la durée de vie d'un panel.

En raison des problèmes de couverture et de mise à jour des bases de sondage téléphoniques, on estime que les approches utilisant les logements plutôt que les numéros de téléphone comme unités d'échantillonnage sont plus prometteuses pour les grandes enquêtes par panel. Il est intéressant de noter que la situation peut être assez différente pour d'autres enquêtes. Catlin et coll. (1984) ont montré que les biais de couverture des caractéristiques de la population générale étaient moindres que pour les caractéristiques de la population active. De plus, pour les enquêtes plus petites (taille de l'échantillon de 10,000 ou moins), les petits biais sont moins importants compte tenu de l'erreur d'échantillonnage relative plus grande pour ces enquêtes. On a ainsi été amené à établir une enquête auprès des ménages CTA en 1986. Elle a servi à de nombreuses enquêtes uniques et pour l'enquête sociale générale, qui est une enquête annuelle auprès de 10,000 ménages.

**Base de sondage aréolaire**

Comme on l'a déjà dit, il est possible dans les régions urbaines d'apparier certaines adresses tirées d'une base de sondage aréolaire aux listes téléphoniques afin de procéder à des interviews téléphoniques à froid. L'expérience avec l'EPA a permis de constater que l'on pouvait obtenir des numéros de téléphone de cette façon pour environ 60% des ménages. Ces taux d'appariement reposent sur un appariement exact après redressement des renseignements de l'adresse et il est possible de les améliorer grâce à des méthodes de raccordement des enregistrements. En téléphonant à une partie appréciable des cas du premier mois, la mise en grappes de l'échantillon pourrait être réduite quelque peu, mais un échantillon en grappes reste une contrainte de tout plan de sondage aréolaire.

On envisage d'étudier la possibilité d'étendre ces procédures aux régions rurales, ce qui permettrait de changer le type de renseignements recueillis lorsqu'on crée des listes de logements pour l'enquête. Les renseignements actuellement recueillis ont tendance à être une description des caractéristiques physiques des logements, alors que si l'on désire effectuer un appariement avec les listes des abonnés au téléphone, il faudrait avoir des renseignements tels que le nom (qui se trouve souvent sur les boîtes à lettres), le nom et le numéro de la rue, ou en l'absence de ces renseignements, le numéro de la route rurale et le code postal.

**Registre des adresses**

Statistique Canada est en train de construire un registre des adresses dans les régions urbaines du Canada. Il servira pendant le recensement de 1991 à améliorer la couverture en fournissant une vérification indépendante des listes de logements créées par les recenseurs (Drew, Royce et van Baaren 1989). Le registre des adresses sera une liste lisible par machine des adresses obtenue par le raccordement de diverses sources de données administratives, y compris les listes de clients avec les numéros publiés achetés auprès des compagnies de téléphone. Pendant l'utilisation du registre des adresses au cours du recensement de 1991, on mettra à jour sa couverture afin de la faire correspondre à celle du recensement de 1991.

On envisage pendant le remaniement post-censitaire de 1991 de l'enquête sur la population active d'effectuer des études sur l'utilisation du registre des adresses comme base de sondage des enquêtes-ménages dans les régions urbaines. Si l'on en conclut qu'on devrait l'adopter comme base de sondage, le registre des adresses sera mis à jour sur une base permanente après le recensement de 1991.

Un avantage du registre des adresses comme base de sondage sur la base de sondage aréolaire est que les ménages avec et sans téléphone sont connus à l'avance. Les deux par conséquent peuvent être échantillonnés comme des strates distinctes, avec un nombre de grappes réduit pour la strate téléphonique, pour laquelle une partie importante des interviews du premier mois peut être faite par téléphone. La strate non téléphonique comprendrait les ménages avec les numéros non publiés et les ménages sans téléphone. D'après les études antérieures, on peut constater que le taux de refus des interviews téléphoniques à froid des ménages avec un numéro non publié est de 12%, comparativement à 4% pour tous les ménages. Dans le cas des interviews téléphoniques à chaud, il n'y a pas d'augmentation correspondante des taux de refus pour les ménages avec des numéros non publiés, et le taux de conversion de ces ménages pour répondre au téléphone au cours des mois suivants est bon. Cette constatation et le désir de respecter les inquiétudes concernant le respect de la vie privée viennent favoriser les interviews personnelles pour ces ménages, qui représentent le pourcentage estimatif de 10-15% des numéros.

**Base de sondage duale**

Dans les régions urbaines, si la couverture du registre des adresses comme base de sondage unique ne convient pas, on peut envisager un plan de sondage dual dans lequel le registre des adresses est complété par un échantillon régional. Un tel échantillon complémentaire pourrait prendre plusieurs formes. Une option intéressante, qui permet de se passer des dépenses de la construction et de la tenue d'une base de sondage aréolaire ordinaire et d'un registre des adresses, serait d'utiliser une approche par intervalle dans laquelle un échantillon de logements consécutifs dans le registre des adresses serait choisi et vérifié sur le terrain. Tout logement trouvé entre des logements du registre des adresses ferait partie d'un échantillon de logements manquants dans le registre des adresses.

Mian (1990) a étudié les méthodes de bases de sondage duales en considérant une optimisation des coûts et de la variance pour le cas général dans lequel aucune des bases ne doit recouvrir l'univers complet. On a estimé que ceci serait un modèle pratique puisque la base de sondage aréolaire, même si elle est complète sur le plan conceptuel, dans la pratique se caractérise par un sous-dénombrement de 3-4% par rapport au recensement, en plus du 5% de la population qui n'est pas représentée en raison de la non-réponse. L'extension du modèle de Mian afin d'inclure une composante d'erreurs non due à l'échantillonnage permettrait de factoriser dans l'optimisation ce que nous savons ou ce que nous désirons supposer à propos des biais de couverture et de non-réponse avec d'autres bases de sondage. Il peut être utilisé dans le contexte des bases de sondage duales qui combinent le registre des adresses et les bases de sondage aréolaires dans les régions urbaines et la combinaison des bases de sondage aréolaires et téléphoniques dans les régions rurales.

## 7. REMANIEMENT POST-CENSITAIRE DE 1991 DE L'EPA

L'enquête sur la population active est remaniée après chaque recensement décennal de la population. Le remaniement porte habituellement sur le remaniement de l'échantillon, mais dans les années 70, on a effectué une révision importante qui comprenait le remaniement de l'échantillon, des modifications au questionnaire, au libellé des questions et aux résultats de l'enquête et une refonte poussée des systèmes de traitement de l'enquête, y compris la mise en place d'un réseau de mini-ordinateurs dans les bureaux régionaux afin d'appuyer les

opérations d'enquête et les saisies de données régionales. Par contre, les travaux de remaniement pendant les années 80 se sont limités à l'échantillon.

Même si les décisions à propos de l'ampleur du remaniement post-censitaire de 1991 n'ont pas encore été prises, tout programme qui se situerait entre la révision majeure des années 70 et le remaniement minimal des années 80 semble nécessaire. Les travaux du remaniement en sont encore au début de la planification, et ils passeront par quatre sous-projets portant sur: i) le contenu et le questionnaire, ii) la modernisation des systèmes de traitement de l'enquête et l'examen des produits de l'enquête, iii) l'élaboration, les tests et la mise en oeuvre des interviews assistées par ordinateur et iv) le remaniement de l'échantillon. (Drew et coll. 1991).

Les sous-projets iii) et iv) sont ceux qui portent sur les méthodes téléphoniques et IAO. Les plans actuels pour ces sous-projets dans la mesure où ils se rattachent aux facteurs d'étude de l'enquête exposés dans les sections précédentes de cette communication sont résumés brièvement ci-dessous.

**Technologie et organisation**

Compte tenu des résultats positifs des essais des interviews assistées par ordinateur pour l'EPA mentionnés par Catlin et coll. (1988), on a décidé de faire de l'IAO un des principaux éléments du remaniement. De plus, pour des raisons déjà exposées, on a décidé que l'organisation locale actuelle des intervieweurs devrait être conservée, et la mise en oeuvre des méthodes IAO prendra la forme d'interviews personnelles assistées par ordinateur (IPAO). Plus précisément, les intervieweurs locaux recevront des ordinateurs portables légers qu'ils prendront avec eux pour les interviews personnelles et ils feront des interviews téléphoniques à partir de leur maison. La combinaison des interviews personnelles et téléphoniques pourrait rester en gros la même qu'actuellement, soit 80% par téléphone et 17% personnelles, mais on pourrait augmenter le nombre d'interviews téléphoniques si l'on décidait d'interviewer davantage par téléphone à froid les ménages pendant leur premier mois dans l'échantillon.

Le plan de travail du sous-projet des interviews assistées par ordinateur comprend un essai sur le terrain en 1991 d'un ordinateur portable à écran par touches, et au cours des années suivantes, l'élaboration ou l'acquisition du logiciel IAO, un essai combiné des autres méthodes IAO et de questionnaire, l'élaboration de l'édition en direct et une version automatisée du manuel de l'intervieweur incorporée dans un écran HELP accessible pendant l'interview.

**Base de sondage et mode de collecte**

Une conclusion fondamentale des recherches a été que les interviews téléphoniques à froid avec un suivi personnel donnaient des taux de réponse et des estimations de la population active comparables à celles obtenues avec la procédure de collecte téléphonique à chaud actuelle pour l'EPA, qui comprend une interview personnelle pour les ménages du premier mois dans l'échantillon. On a observé une exception au Québec, où, comme on l'a déjà dit, des différences dans les données sont apparues pendant une période. L'importance du suivi personnel dans le maintien des taux de réponse favorise le maintien du logement comme unité de sondage et les numéros de téléphone fournis aux intervieweurs leur donneront plus de souplesse pour l'utilisation du téléphone et les interviews personnelles pour l'obtention des interviews du premier mois.

Dans les régions urbaines, la base aréolaire actuelle et le registre des adresses comme base de sondage sont cohérents avec cette approche. Dans les régions rurales, comme on l'a déjà dit, on étudiera la faisabilité d'apparier les adresses de la base de sondage aréolaire et les listes téléphoniques afin de fournir aux intervieweurs les numéros de téléphone. On prévoit également de poursuivre l'étude des méthodes de bases de sondage duales qui, dans les régions urbaines, pourraient comprendre le registre des adresses et une base de sondage aréolaire, et dans les régions rurales, une base de sondage aréolaire et une base de sondage téléphonique.

## 8. RÉSUMÉ

La méthodologie de collecte des données actuelle de la population active comprend une interview personnelle pour les ménages pendant leur premier mois dans l'échantillon et principalement des interviews téléphoniques au cours des mois suivants. La mise en oeuvre des interviews téléphoniques au cours des mois suivants a eu lieu au cours des années 70 pour les grandes régions urbaines et au cours des années 80 dans les autres régions. Dans les deux cas, les interviews téléphoniques se sont traduites par des économies appréciables des coûts sans aucune incidence sur les taux de réponse ou les estimations d'enquête. Avant le début du programme de recherche et de tests téléphoniques et IAO, en 1985, 80% des interviews EPA se faisaient par téléphone. Cette proportion a légèrement augmenté pour passer à 83% grâce à la mise en oeuvre du suivi téléphonique pour les ménages qui n'ont pu être contactés pendant une visite personnelle initiale et par la fourniture aux intervieweurs des numéros de téléphone de l'échantillon des appartements.

Le principal avantage du programme de recherche et d'essais a été de déterminer les options de la base de sondage et de la collecte des données à conserver pendant le remaniement postcensitaire de 1991 de l'enquête, ce qui comprend le maintien de l'organisation locale actuelle des intervieweurs, l'adoption des interviews personnelles assistées par ordinateur, le maintien des approches de la base de sondage et du plan de sondage dans lesquelles le logement est l'unité de sélection et la fourniture aux intervieweurs des numéros de téléphone afin de disposer d'une marge de manoeuvre dans l'emploi d' une combinaison d'interviews téléphoniques et personnelles pour l'obtention des interviews du premier mois.

## BIBLIOGRAPHIE

CATLIN, G., CHOUDHRY, H., et HOFMANN, H. (1984). Telephone ownership in Canada. Document interne, Statistique Canada.

CATLIN, G., et INGRAM, S. (1988). The effects of CATI on cost and data quality. *Telephone Survey Methodology*, (éds. R. Groves et coll.), 437-450. New York: Wiley.

CHOUDHRY, G.H. (1984). Results from telephone interviewing experiment in the non self representing areas of the Labour Force Survey. Document interne, Statistique Canada.

DREW, J.D., et GAMBINO, J. (1991). Plans for the 1991 post censal redesign of the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, à paraître.

DREW, J.D., CHOUDHRY, H., et HUNTER, L. (1988). Nonresponse issues in government telephone surveys. *Telephone Survey Metholodogy*, (éds. R. Groves et coll.), 233-246. New York: Wiley.

DREW, J.D., DICK, P., et SWITZER, K. (1989). Development and testing of telephone survey methods for household surveys at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

DREW, J.D., et JAWORSKI, R. (1986). Telephone survey development on the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

DREW, J.D., ROYCE, D., et van BAAREN, A. (1989). Address register research at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

DUFOUR, J. (1990). Implantation de la première entrevue par téléphone pour la base d'appartements de l'enquête sur la population active. Document interne, Division des méthodes d'enquêtes sociales, Statistique Canada.

GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., II, et WAKSBERG, J. (éds) (1988). *Telephone Survey Methodology*. New York: Wiley.

LAFLAMME, F. (1990). Étude comparative entre trois différentes populations visées par l'EPA selon leur type de service téléphonique. Document interne, Division des méthodes d'enquêtes sociales, Statistique Canada.

MIAN, I.U.H. (1990). Dual frame estimation of proportions in sample surveys. Document interne, Division des méthodes d'enquêtes sociales, Statistique Canada.

MUIRHEAD, R.C., GOWER, A.R., et NEWTON, F.T. (1975). The telephone experiment in the Canadian Labour Force Survey. *Survey Methodology*, 1, 158-180.

TREWIN, D., et LEE, H. (1988). International comparisons of telephone coverage. *Telephone Survey Methodology*, (éds. R. Groves et coll.). New York: Wiley, 9-24.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 576-579.

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 77–87
Statistique Canada



# Fonctions de vraisemblance marginales et fonctions de vraisemblance conditionnelles approximatives pour l'échantillonnage répété

D.R. BELLHOUSE[1]

RÉSUMÉ

L'auteur définit des fonctions de vraisemblance marginales et des fonctions de vraisemblance conditionnelles approximatives pour les paramètres de corrélation d'un modèle de régression linéaire normal à erreurs corrélées. L'auteur se sert du principe de vraisemblance pour déterminer des fonctions de vraisemblance marginales et des fonctions de vraisemblance conditionnelles approximatives pour les paramètres de corrélation dans un plan d'échantillonnage répété (échantillonnage aléatoire simple et plans plus complexes).

MOTS CLÉS: Inférence fondée sur la vraisemblance; échantillonnage dans le temps; modèles ARMA; modèles d'espace d'états.

## 1. INTRODUCTION

Considérons une population finie de $N$ unités qui peut être sondée $k$ fois. Désignons par $y_{tj}$ la valeur observée pour l'unité de population $j$ au $t$-ième passage de l'enquête, $j = 1, \ldots, N$ et $t = 1, \ldots, k$. Nous supposons que les unités de population sont indépendantes les unes des autres mais qu'il y a corrélation dans le temps pour les valeurs observées pour la même unité. En particulier, nous supposons que pour n'importe quelle unité $j$,

$$(y_{1j}, y_{2j}, \ldots, y_{kj})^T \sim N(\mu, \sigma^2\Omega), \tag{1}$$

où $\Omega$ est une matrice de corrélation $k \times k$ et $\mu$ est le vecteur $1 \times k$ de moyennes fixes $(\mu_1, \mu_2, \ldots, \mu_k)^T$. En vertu de l'hypothèse du modèle explicite en (1), l'auteur utilise dans le présent article une méthode d'estimation fondée sur des modèles. À partir de tous les échantillons prélevés aux $k$ passages, il est utile d'estimer $(\mu_1, \mu_2, \ldots, \mu_k)^T$. La forme d'estimation fondée sur un modèle $(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)^T$, lorsqu'elle est définie par la méthode du maximum de vraisemblance ou par celle des moindres carrés généralisés, par exemple, dépendra de $\sigma^2$ et des paramètres dans $\Omega$. Par conséquent, nous devons obtenir des estimateurs valables pour $\sigma^2$ et les paramètres dans $\Omega$.

Nous reprenons la notation de Bellhouse (1989) pour décrire le plan d'échantillonnage considéré ici, soit un échantillonnage avec renouvellement à un niveau. À chaque passage, on échantillonne $c$ groupes de renouvellement. Le groupe de renouvellement $r$ ($r = 1, 2, \ldots, k + c - 1$), désigné par $G_r$, contient $m_r$ unités. Au passage $t$ ($t = 1, \ldots, k$), l'échantillon comprend les unités des groupes $G_t, G_{t+1}, \ldots, G_{t+c-1}$, de sorte que sa taille $n_t$ est égale à $m_t + m_{t+1} + \ldots + m_{t+c-1}$. Au passage $t + 1$, on laisse tomber $G_t$ de l'échantillon et on ajoute $G_{t+c}$. Chaque groupe de renouvellement est choisi sans remise parmi les unités de population qui n'ont pas été prélevées auparavant. Pour les $k$ passages considérés globalement, la taille de l'échantillon est $m = n_1 + n_2 + \ldots + n_k$. Le nombre maximal de passages où une unité reste dans l'échantillon est $c$.

[1] D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B9.

Lorsque la valeur $c$ est peu élevée, les valeurs estimées pour les paramètres de corrélation dans $\Omega$ peuvent alors être instables, causant une instabilité dans l'estimateur d'intérêt $(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)^T$. Considéré d'une autre manière, le nombre total de paramètres est d'au moins $k + 2$ et augmentera avec le temps, c'est-à-dire à chaque nouveau passage. Étant donné que la dimension de l'espace des paramètres augmentera avec le temps, l'estimateur du maximum de vraisemblance pourrait être biaisé et non convergent. Le problème de la stabilité des estimateurs dans un échantillonnage répété a été étudié, par exemple, par Blight et Scott (1973), qui supposent que les éléments de $(\mu_1, \mu_2, \ldots, \mu_k)^T$ suivent un processus de série chronologique. À partir de cette hypothèse, on fixe la dimension de l'espace des paramètres à un nombre relativement peu élevé de sorte qu'on règle les problèmes d'instabilité, de biais et de non convergence. Dans le présent article, l'auteur adopte une approche différente. Il conserve l'hypothèse des moyennes fixes et établit la fonction de vraisemblance marginale (FM) et la fonction de vraisemblance conditionnelle approximative (FCA) pour les paramètres dans $\Omega$, traitant les moyennes fixes comme des paramètres dérangeants.

On a proposé pour la première fois les FM comme une méthode générale pour éliminer les paramètres dérangeants de la fonction de vraisemblance (Fraser 1967; Kalbfleisch et Sprott 1970). Les FCA ont été définies dans le même but par Cox et Reid (1987). Ceux-ci affirment que la FCA est préférable à la fonction profil de vraisemblance, que l'on obtient en remplaçant les paramètres dérangeants dans la fonction de vraisemblance par l'estimation la plus vraisemblable correspondante lorsque les paramètres d'intérêt sont connus. Bellhouse (1990) a démontré l'équivalence de la FM et de la FCA pour des paramètres de corrélation suivant un modèle normal.

S'inspirant de l'étude de Cox et Reid, Cruddas et coll. (1989) ont établi une FCA pour les paramètres de corrélation dans plusieurs petites séries de processus autorégressifs du premier degré ayant la même variance et les mêmes paramètres d'autocorrélation. Ils ont montré par une étude de simulation que l'estimateur fondé sur la FCA était beaucoup moins biaisé que l'estimateur du maximum de vraisemblance fondé sur la fonction profil et qu'il recouvrait mieux l'intervalle de confiance. La situation décrite par Cruddas et ses coll. (1989) s'applique directement aux enquêtes à passages répétés. Afin de réduire le fardeau de réponse des personnes qui participent à ce genre d'enquêtes, on fait en sorte qu'elles ne fassent pas trop longtemps partie de l'échantillon. On prévoit que l'utilisation des FM et des FCA permettra d'obtenir de meilleures valeurs estimées des paramètres de corrélation et, par conséquent, de meilleures valeurs estimées de la moyenne pour chaque passage.

À l'intérieur d'un groupe de renouvellement, les données sur chaque personne sont normalement modélisées à l'aide d'un processus autorégressif de moyennes mobiles (ARMA), c'est-à-dire que les paramètres dans $\Omega$ sont composés des paramètres de corrélation du processus ARMA. Voir Binder et Hidiroglou (1988) pour une étude de l'application des modèles de séries chronologiques à l'échantillonnage répété. Par conséquent, il est utile de définir des FM et des FCA suivant des modèles ARMA appliqués à un échantillonnage avec renouvellement. Dans la section 2, nous déterminons les FM et les FCA pour les paramètres de corrélation suivant un modèle normal. Nous appliquons ensuite les résultats de cette section aux enquêtes à passages répétés avec plan d'échantillonnage aléatoire simple d'unités dans des groupes de renouvellement. Enfin, dans la section 4, nous présentons plusieurs méthodes qui permettent d'appliquer ces fonctions de vraisemblance à des plans de sondage complexes.

## 2. FONCTION DE VRAISEMBLANCE MARGINALE ET FONCTION DE VRAISEMBLANCE CONDITIONNELLE APPROXIMATIVE POUR DES PARAMÈTRES DE CORRÉLATION SUIVANT UN MODÈLE NORMAL

Supposons que $y$ est un vecteur d'observations de dimensions $m \times 1$ suivant le modèle linéaire

$$y = X\beta + \epsilon \tag{2}$$

où le vecteur d'erreurs $\epsilon \sim N(\mathbf{0},\sigma^2\Phi)$ $\Phi$, étant la matrice de corrélation de dimensions $m \times m$ et $\beta$ le vecteur des coefficients de régression $p \times 1$ de sorte que $X$ est de dimensions $m \times p$. La fonction de vraisemblance logarithmique pour $\beta$, $\sigma^2$ et $\Phi$ est définie

$$L(\beta,\sigma^2,\Phi) = -\left\{m \ln\sigma + (\ln \mid \Phi \mid)/2 + (y - X\beta)^T\Phi^{-1}(y - X\beta)/(2\sigma^2)\right\}. \quad (3)$$

Pour une valeur donnée de $\Phi$,

$$\hat{\beta} = (X^T\Phi^{-1}X)X^T\Phi^{-1}y$$

et

$$s^2 = y^T\Phi^{-1}y - y^T\Phi^{-1}X\,(X^T\Phi^{-1}X)^{-1}\,X^T\Phi^{-1}y \quad (4)$$

sont des estimateurs exhaustifs conjoints de $\beta$ et $\sigma^2$.

On obtient une fonction de vraisemblance marginale pour $\Phi$ par une réduction des données $y$ aux statistiques exhaustives $\hat{\beta}$ et $s^2$ et à la statistique ancillaire

$$a = \Phi^{-1/2}(y - X\,(X^T\Phi^{-1}X)^{-1}\,X^T\Phi^{-1}y)/s,$$

où $\Phi^{-1/2}$ est la matrice de dimensions $m \times m$ telle que $\Phi^{-1} = \Phi^{-1/2}\Phi^{-1/2}$. La fonction de vraisemblance marginale de $\Phi$ correspond à la distribution marginale de la statistique ancillaire $a$ multipliée par le produit des différentielles $da_i$, $i = 1, \ldots, m$. Voir Kalbfleisch et Sprott (1970, équations 6 et 10) pour une analyse globale et une expression générale pour $\Pi da_i$. Bellhouse (1978), suivi quelques années plus tard de Tunnicliffe Wilson (1989), a montré que la fonction de vraisemblance marginale pour $\Phi$ suivant un modèle normal était définie

$$L_M(\Phi) = \left\{\mid \Phi \mid^{1/2} \mid X^T\Phi^{-1}X \mid^{1/2} s^{m-p}\right\}^{-1}. \quad (5)$$

Notons que l'équation (4) est proportionnelle à l'estimateur du maximum de vraisemblance de $\sigma^2$ étant donné $\Phi$ et que $s^2(X^T\Phi^{-1}X)^{-1}$ est proportionnelle à la matrice des variances-covariances estimée de l'estimateur du maximum de vraisemblance de $\beta$ étant donné $\Phi$. Alors, l'équation (5) peut être réécrite

$$L_M(\Phi) = \frac{\mid \text{est var}(\hat{\beta}) \mid^{1/2}}{s^m \mid \Phi \mid^{1/2}}. \quad (6)$$

Pour définir une fonction de vraisemblance conditionnelle approximative (FCA), il faut d'abord transformer les paramètres de manière à obtenir une relation d'orthogonalité entre les paramètres d'intérêt et les paramètres dérangeants, qui peuvent dépendre des premiers. Il y a orthogonalité entre les ensembles de paramètres lorsque la matrice d'information correspondante est une matrice diagonale en blocs, chaque bloc servant lui-même de matrice d'information pour un ensemble de paramètres. La fonction de vraisemblance conditionnelle est liée à la distribution des données $y$, qui dépend elle-même de l'estimateur du maximum de vraisemblance des paramètres dérangeants pour les valeurs déterminées des paramètres d'intérêt. On obtient la fonction de vraisemblance conditionnelle approximative en appliquant deux approximations à cette distribution conditionnelle. Voir Cox et Reid (1987, sous-section 4.1), pour une analyse des calculs. Par exemple, posons $\Theta$ comme le vecteur des paramètres d'intérêt et $\Lambda$, qui dépend vraisemblablement de $\Theta$, comme le vecteur des paramètres dérangeants orthogonal à $\Theta$. La fonction de vraisemblance complète pour les paramètres $\Theta$ et $\Lambda$ est désignée par $L(\Theta,\Lambda)$ tandis que la fonction profil pour $\Theta, L(\Theta,\hat{\Lambda})$ équivaut à la fonction de vraisemblance ordinaire à la différence près que $\Lambda$ est remplacé par l'estimateur du maximum de vraisemblance correspondant. La fonction de vraisemblance conditionnelle approximative pour $\Theta$ est

$$L(\Theta,\hat{\Lambda}) \mid I(\Theta,\hat{\Lambda}) \mid^{1/2},$$

où $I(\Theta,\hat{\Lambda})$ est la matrice d'information observée pour $\Lambda$ à une valeur déterminée de $\Theta$. Voir Cox et Reid (1987, équation 10).

À la suite de Cruddas et coll. (1989), Bellhouse (1990) a proposé pour le modèle (2) la transformation de paramètre $\lambda = \ln\sigma + (\ln \mid \Omega \mid)/(2m)$, $\beta$ demeurant inchangé. Suivant ces nouvelles conditions, la fonction de vraisemblance logarithmique est désignée par $L(\beta,\lambda,\Phi)$ et peut être tirée de l'équation (3). Si les éléments de $\Phi$ sont des fonctions d'un paramètre $\phi$, les paramètres dérangeants $\lambda$ et $\beta$ sont l'un et l'autre orthogonaux à $\Phi$, c'est-à-dire

$$-\frac{1}{m}E\left[\frac{\partial^2 L(\beta,\lambda,\Phi)}{\partial\phi\partial\lambda}\right] = 0$$

et

$$-\frac{1}{m}E\left[\frac{\partial^2 L(\beta,\lambda,\Phi)}{\partial\phi\partial\beta}\right] = 0,$$

lorsque chaque élément de $\Phi$ est une fonction continue et différentielle de $\phi$. En outre dans ce cas, la fonction conditionnelle approximative pour $\Phi$, $L_C(\Phi)$ est identique à la fonction marginale, $L_M(\Phi)$, définie en (5) ou en (6). Voir Bellhouse (1990) pour plus de détails.

Il est possible d'évaluer la FM et la FCA définies en (5) ou en (6) pour n'importe quelle matrice $\Phi$ en se servant de modèles d'espace d'états à la manière de Harvey et Phillips (1979). Une fois que les calculs récursifs visant à estimer $\beta$ et $\sigma^2$ sont terminés, on peut calculer, pour n'importe quelle matrice $\Phi$, la valeur de $s^2$ et de $\mid \Phi \mid^{1/2}$ au moyen des formules proposées par Harvey et Phillips (1979, équations 5.6 et 6.6, et 4.3 respectivement). On n'a alors qu'à calculer $X^T\Phi^{-1}X$ et son déterminant. La dernière étape du processus récursif de Harvey et Phillips (1979, équation 3.4) permet de déterminer la valeur de $X^T\Phi^{-1}X$.

## 3. ÉCHANTILLONNAGE ALÉATOIRE SIMPLE RÉPÉTÉ

### 3.1 Quelques résultats pour l'échantillonnage avec renouvellement

Supposons que le groupe de renouvellement $G_r$ est choisi la première fois au passage $u$ et la dernière fois au passage $v$, $u$ étant égal à 1 ou à $r$ et $v$ correspondant à $r + c - 1$ ou à $k$. Le nombre total de passages où une unité de $G_r$ est incluse dans l'échantillon est $b = v + 1 - u$. Soient $\bar{y}_{u,r}, \ldots, \bar{y}_{v,r}$ les moyennes d'échantillon ou les estimations élémentaires pour $G_r$ aux passages $u, u+1, \ldots, v-1, v$ respectivement. Alors, suivant le modèle (1), la contribution de $G_r$ à la fonction de vraisemblance logarithmique définie en (3) est

$$-\left\{bn_r \ln\sigma + (n_r/2)\ln(\mid \Omega_r \mid) + \left[n_r x_r^T \Omega_r^{-1} x_r + (n_r - 1)\,\mathrm{tr}(\Omega_r^{-1} S_r)\right]/(2\sigma^2)\right\}, \quad (7)$$

où $x_r^T$ est le vecteur $1 \times b$ $(\bar{y}_{u,r} - \mu_u, \bar{y}_{u+1,r} - \mu_{u+1}, \ldots, \bar{y}_{v-1,r} - \mu_{v-1}, \bar{y}_{v,r} - \mu_v)$, $S_r$ est la matrice $b \times b$ des variances et des covariances de l'échantillon pour les observations relatives au groupe de renouvellement et où $\Omega_r$ est la matrice de corrélation $b \times b$ des observations relatives à une unité du groupe de renouvellement. Notons que les paramètres dans $\Omega$ seront aussi les paramètres dans $\Omega_r$. La matrice de corrélation $\Omega$ est fondée sur les valeurs observées à tous les passages 1 à $k$; la matrice de corrélation $\Omega_r$ est tirée du sous-ensemble de données observées aux passages $u$ à $v$. En vertu de l'hypothèse d'indépendance, on obtient la fonction de vraisemblance logarithmique en faisant la sommation de l'expression (7) pour tous les groupes de renouvellement.

Étant donné les paramètres dans $\Omega$, ou bien les paramètres dans $\Omega_1, \ldots, \Omega_{k+c-1}$, il est possible de déterminer des expressions pour les estimateurs du maximum de vraisemblance $\hat{\mu}$ et $\hat{\sigma}^2$, qui servent à estimer $\mu$ et $\sigma^2$ respectivement. De même, il est possible de connaître la matrice des variances-covariances estimée de $\hat{\mu}$, $V(\hat{\mu})$. C'est ce que nous illustrons pour un processus autorégressif du premier degré dans la sous-section 3.2. La fonction de vraisemblance marginale pour $\Omega_1, \ldots, \Omega_{k+c-1}$ est alors définie par l'équation (5), où

$$| \Phi |^{1/2} = \prod_{r=1}^{k+c-1} \Omega_r,$$

$$| X^T \Phi^{-1} X |^{1/2} = V(\hat{\mu}) / s^k, \tag{8}$$

$$s^2 = \sum_{r=1}^{k+c-1} \left\{ n_r \hat{x}_r^T \Omega_r^{-1} \hat{x}_r + (n_r - 1) \operatorname{tr}(\Omega_r^{-1} S_r) \right\},$$

et $p = k$; dans l'équation ci-dessus, $\hat{x}_r$ est $x_r$ à la différence près que $\mu$ est remplacée par l'estimateur du maximum de vraisemblance correspondant.

### 3.2 Processus autorégressifs du premier degré

Lorsqu'on utilise des formes précises des matrices de corrélation $\Omega_1, \ldots, \Omega_{k+c-1}$, il est possible de simplifier quelque peu la forme générale de la fonction de vraisemblance marginale pour les paramètres de corrélation, définie en (5) et en (6). Par exemple, supposons le modèle autorégressif du premier degré

$$y_{tj} = \mu_t + \phi (y_{t-1,j} - \mu_{t-1}) + \epsilon_{tj}, \tag{9}$$

où $\epsilon_{tj} \sim N(0,\sigma^2)$ pour $t = 1, \ldots, k$ et $j = 1, \ldots, N$, et où les $\epsilon$ sont mutuellement indépendants. Le modèle (9), qui correspond essentiellement au modèle de Patterson (1950), est un cas particulier du modèle (1). Comme dans la sous-section 3.1, le vecteur des paramètres de régression $\beta$ est $(\mu_1, \ldots, \mu_k)^T$. Lorsque le vecteur de données $y$ contient pour chaque unité les observations groupées selon les passages où cette unité a été échantillonnée, conformément à la description qui est faite de l'échantillonnage avec renouvellement dans la sous-section 3.1, on peut exprimer la matrice de corrélation $\Phi$, qui est désormais une fonction du paramètre autorégressif $\phi$, comme une somme directe de matrices qui sont toutes les matrices de corrélation d'un processus autorégressif du premier degré.

Nous reprenons la notation utilisée par Patterson (1950) pour désigner des tailles d'échantillon, des moyennes ainsi que des sommes des carrés et des produits (redressées en fonction de la moyenne pertinente) pour le passage $t$:

- $\pi_t$ = la proportion d'unités échantillonnées au passage $t$, qui étaient aussi présentes dans l'échantillon au passage précédent $(t - 1)$;
- $n_t$ = le nombre d'unités échantillonnées au passage $t$;
- $\bar{y}_t'$ = la moyenne pour les unités échantillonnées au passage $t$, qui étaient aussi présentes dans l'échantillon au passage précédent $(t - 1)$;
- $\bar{y}_t''$ = la moyenne pour les unités échantillonnées au passage $t$ et qui n'étaient pas présentes dans l'échantillon au passage précédent $(t - 1)$;
- $\bar{y}_t$ = la moyenne pour toutes les unités échantillonnées au passage $t$;
- $\bar{x}_t'$ = la moyenne pour les unités échantillonnées au passage $t$, qui sont aussi présentes dans l'échantillon au passage suivant $(t + 1)$;

$syy_t'$ = la somme des carrés pour les unités échantillonnées au passage $t$, qui étaient aussi présentes dans l'échantillon au passage précédent $(t - 1)$;

$syy_t''$ = la somme des carrés pour les unités échantillonnées au passage $t$ et qui n'étaient pas présentes dans l'échantillon au passage précédent $(t - 1)$;

$sxx_t'$ = la somme des carrés pour les unités échantillonnées au passage $t$, qui sont aussi présentes dans l'échantillon au passage suivant $(t + 1)$;

$syy_t$ = la somme des carrés pour toutes les unités échantillonnées au passage $t$;

$sxy_t'$ = la somme des produits pour les observations relatives aux unités échantillonnées au passage $t$, qui étaient aussi présentes dans l'échantillon au passage précédent $(t - 1)$.

Suivant le cas particulier du modèle (9), et après de nombreuses transformations algébriques, nous pouvons montrer que, lorsqu'on fait la sommation de l'expression (7) pour tous les groupes de renouvellement $r$, la fonction de vraisemblance logarithmique pour les données se ramène à

$$L(\mu_1, \ldots, \mu_k, \sigma^2, \phi) = -m \ln\sigma + (d/2)\ln(1 - \phi^2) - [A(\mu,\phi) + B(\phi)]/(2\sigma^2), \tag{10}$$

où $d$ est le nombre d'unités distinctes échantillonnées (c'est-à-dire abstraction faite du nombre de fois qu'une unité est échantillonnée) et $m$ est la taille de l'échantillon global $(n_1 + \ldots + n_k)$. De plus, dans l'équation (10),

$$A(\mu,\phi) = (1 - \phi^2)n_1(\bar{y}_1 - \mu_1)^2 + \sum_{t=2}^{k} \left[\pi_t n_t \{\bar{y}_t' - \mu_t - \phi(\bar{x}_{t-1}' - \mu_{t-1})\}^2 + (1 - \pi_\tau)n_t(1 - \phi^2)(\bar{y}_t'' - \mu_t)^2\right] \tag{11}$$

et

$$B(\phi) = (1 - \phi^2)\, syy_1 + \sum_{t=2}^{k} \{\phi^2\, sxx_{t-1}' - 2\phi\, sxy_t' + syy_t' + (1 - \phi^2)\, syy_t''\}. \tag{12}$$

Pour n'importe quelle valeur donnée de $\phi$, les estimateurs du maximum de vraisemblance sont $\hat{\mu} = G^{-1}z$ et $\hat{\sigma}^2 = [A(\hat{\mu},\phi) + B(\phi)]/m$, où $A(\hat{\mu},\phi)$ est défini par l'équation (11), $\mu$ étant remplacé par son estimateur du maximum de vraisemblance, et où $G$ est une matrice bande symétrique $k \times k$ de largeur de bande 3 et $z$ est un vecteur $k \times 1$. Les éléments non nuls de $G$ sont définis

$$g_{tt} = \pi_t n_t + (1 - \pi_t)n_t(1 - \phi^2) + \pi_{t+1}n_{t+1}\phi^2, \quad \text{pour} \quad t = 1, \ldots, k$$

et

$$g_{t,t+1} = -\pi_{t+1}n_{t+1}\phi, \quad \text{pour} \quad t = 1, \ldots, k - 1,$$

où $\pi_1 = \pi_{k+1} = 0$. Les éléments de $z$ sont définis

$$z_t = \pi_t n_t(\bar{y}_t' - \phi\bar{x}_{t-1}') + (1 - \pi_t)n_t\bar{y}_t''(1 - \phi^2) - \pi_{t+1}n_{t+1}(\bar{y}_{t+1}' - \phi\bar{x}_t'),$$

pour $t = 1, \ldots, k$, où $\pi_1 = \pi_{k+1} = 0$ et $\bar{y}_1'' = \bar{y}_1$. Le vecteur des moyennes estimées $\hat{\mu}$ est non biaisé pour $\mu$ selon le modèle (9) et sa matrice des variances-covariances correspondante est $\sigma^2 G^{-1}$. Ainsi, d'après les équations (5) ou (6), la fonction de vraisemblance marginale et la fonction de vraisemblance conditionnelle approximative pour $\phi$ est

$$L_M(\phi) = \frac{(1 - \phi^2)^{d/2}}{\{A(\hat{\mu},\phi) + B(\phi)\}^{(m-k)/2} \mid G \mid^{1/2}}. \tag{13}$$

### 3.3 Exemple

Les données utilisées aux fins du présent exemple sont des données sur la sylviculture tirées de Cunia et Chevrou (1969, p. 220). Elles représentent le volume négociable de bois de construction par tracé observé à trois passages avec remise partielle des unités d'échantillon. Dans un échantillonnage avec renouvellement, on suppose que, une fois qu'on laisse tomber une unité de l'échantillon, celle-ci n'est pas sélectionnée de nouveau. Compte tenu de cette hypothèse, on a rectifié les données dans Cunia et Chevrou. En particulier, on a laissé tomber du présent exemple les valeurs observées pour les unités qui étaient présentes dans l'échantillon aux premier et troisième passages, mais qui n'étaient pas présentes dans l'échantillon au deuxième passage. À partir des données restantes, il est possible de faire les calculs suivants:

$\pi_2 = 86/139$, $\pi_3 = 76/100$, $n_1 = 104$, $n_2 = 139$, $n_3 = 100$, $\bar{y}_2' = 161.5581$, $\bar{y}_3' = 179.9211$, $\bar{y}_1'' = 154.0673$, $\bar{y}_2'' = 167.2075$, $\bar{y}_3'' = 181.125$, $\bar{x}_1' = 147.6512$, $\bar{x}_2' = 163.4342$, $syy_2' = 864129.2$, $syy_3' = 555369.5$, $syy_1'' = 943948.5$, $syy_2'' = 266820.7$, $syy_3'' = 271762.6$, $sxx_1' = 800753.5$, $sxx_2' = 559850.7$, $sxy_2' = 812435.7$, $sxy_3' = 550943.6$, $d = 181$, et $m - k = 340$.

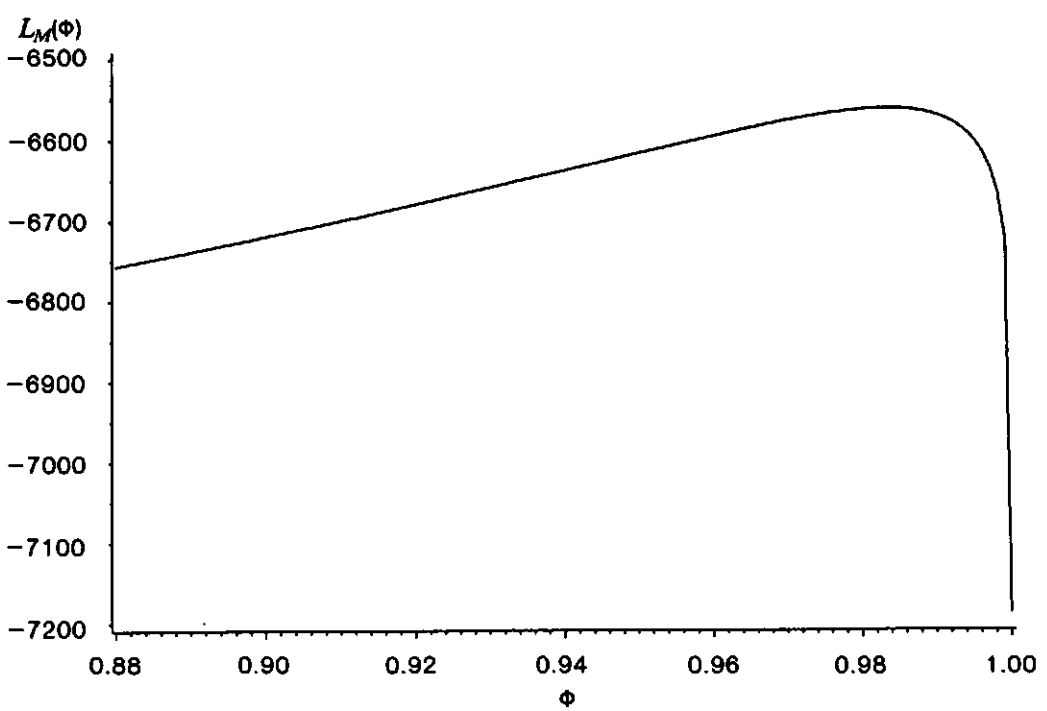



**Figure 1.** Fonction de vraisemblance marginale pour le paramètre AR(1)

En substituant ces données dans l'équation (13), il est possible d'obtenir la fonction de vraisemblance marginale et la fonction de vraisemblance conditionnelle approximative correspondant aux données pour le paramètre autorégressif du premier degré $\phi$. Voir la figure 1.

## 4. ENQUÊTES À PLAN DE SONDAGE COMPLEXE

Il y a plusieurs façons d'analyser des données de séries chronologiques tirées d'enquêtes à plan de sondage complexe. Chaque méthode qu'on peut proposer dépendra des données d'échantillon qui auront pu être recueillies.

Si, par exemple, on dispose de microdonnées, il est possible de calculer pour chaque groupe de renouvellement la matrice, fondée sur le plan de sondage complexe, des variances-covariances des estimations élémentaires. On définit une pseudo fonction de vraisemblance marginale en remplaçant $\mathfrak{L}_r$ et $S_r$ dans les équations (5) et (8) par leurs équivalents pour les enquêtes à plan de sondage complexe. C'est la méthode qu'utilisent par exemple Roberts, Rao et Kumar (1987) dans l'analyse de régression logistique pour plans de sondage complexes: déterminer une fonction de vraisemblance ou un ensemble d'équations de vraisemblance et remplacer les paramètres statistiques habituels par leurs équivalents pour les enquêtes à plan de sondage complexe.

Suivant un plan d'échantillonnage aléatoire simple, la matrice $S_r$ estime la matrice des variances-covariances de la population finie pour les valeurs observées aux passages incluant le groupe de renouvellement $r$. Ainsi, dans un plan de sondage complexe, on remplace la matrice $S_r$ par un estimateur, consistant avec le plan, de la matrice des variances-covariances de la population finie correspondante. Par exemple, Kilpatrick (1981) a examiné un plan de sondage stratifié à deux passages pour évaluer le nombre d'arbres sur pied dans les forêts appartenant à l'État en Irlande du Nord; les strates étaient fondées sur les époques, en commençant par les années vingt, où les forêts ont été plantées. Afin de calculer l'équivalent de $S_r$ dans un plan de sondage stratifié, il est nécessaire d'avoir l'estimateur des moyennes à chaque passage, les moyennes des strates, les variances des strates ainsi que les covariances des strates pour les unités qui étaient présentes et non présentes dans l'échantillon aux deux passages. Pour une population stratifiée, on peut décomposer la variance (ou covariance) de la population finie en des termes comprenant la variation (ou la covariation) entre les strates et la variation (ou la covariation) à l'intérieur des strates; voir, par exemple, Cochran (1977, équation 5.32). On utiliserait les valeurs estimées des moyennes et les moyennes des strates pour obtenir des estimations convergentes de la variation ou de la covariation entre les strates, et les valeurs estimées des variances et des covariances des strates pour obtenir des estimations de la variation et de la covariation à l'intérieur des strates. Malheureusement, seuls certains estimateurs des variances et des covariances des strates s'appliquaient à l'étude de Kilpatrick, de sorte que l'article ne contient pas suffisamment de données pour pouvoir calculer un estimateur du maximum de vraisemblance marginal pour la corrélation entre les volumes de bois aux deux passages.

Or, il arrive rarement que l'on dispose de microdonnées. La méthode d'estimation dépend alors des données disponibles. Nous envisageons ici un scénario mais d'autres sont possibles. Supposons que nous connaissons seulement les estimations élémentaires et les effets du plan correspondant. Soit $\bar{y}_{t,r}$ l'estimation tirée du groupe de renouvellement $G_r$ au passage $t$ et fondée sur un échantillon de taille $m_r$. Soit $\text{deff}_{t,r}$ l'effet du plan qui correspond à $\bar{y}_{t,r}$. Si $\sigma^2/m_r$ est la variance de $\bar{y}_{t,r}$ suivant un échantillonnage aléatoire simple, alors, en vertu du théorème central limite,

$$(\bar{y}_{t,r} - \mu_t)/(\text{deff}_{t,r})^{1/2} \sim N(0,\sigma^2/m_r) \tag{14}$$

approximativement. Nous pouvons établir un modèle en supposant, pour $G_r$, un processus ARMA comme celui ci-dessous

$$(\bar{y}_{t,r} - \mu_t)/(\text{deff}_{t,r})^{1/2} = \phi\,(\bar{y}_{t-1,r} - \mu_{t-1})/(\text{deff}_{t-1,r})^{1/2} + \epsilon_t, \qquad (15)$$

où $\epsilon_t$ a une variance constante. Cela cadre bien avec le modèle (2), où le vecteur d'observations $y$ renferme des données de la forme $\bar{y}_{t,r}/(\text{deff}_{t,r})^{1/2}$, $\beta$ est $(\mu_1, \mu_2, \ldots, \mu_k)^T$ et $X$ renferme des éléments de la forme $1/(\text{deff}_{t-1,r})^{1/2}$. La fonction de vraisemblance marginale, considérée en l'occurrence comme un cas particulier des équations (5) ou (6), peut être évaluée à l'aide du modèle d'espace d'états de Harvey et Phillips (1979) et dont nous avons fait mention dans la section 2. Compte tenu du modèle défini ci- dessus (équ. 20 et 21), il est souhaitable de recourir à l'estimation fondée sur la fonction de vraisemblance marginale et la fonction de vraisemblance conditionnelle approximative. La valeur estimée de $\phi$ repose en l'occurrence sur la variation entre les estimations élémentaires dans chaque groupe de renouvellement, la variance à l'intérieur de ces estimations n'étant pas connue. Comme un groupe de renouvellement passe relativement peu de temps dans l'échantillon, il y a de fortes chances que les estimateurs du maximum de vraisemblance soient biaisés et non convergents.

## 5. ANALYSE

Binder et Dick (1990) ont aussi proposé d'utiliser des techniques d'estimation fondée sur la fonction de vraisemblance marginale pour l'échantillonnage répété. Suivant leur cadre, supposons que nous connaissons les valeurs estimées des moyennes, $\bar{y}_t$, pour chaque passage $t = 1, \ldots, k$. Supposons aussi que la matrice, $S$, des variances-covariances des estimations est connue. Comme dans Binder et Dick (1989, 1990) notamment, les $\bar{y}_t$ peuvent être définies par le modèle

$$\bar{y}_t = \mu_t + e_t, \qquad (16)$$

où $e_t$ est l'erreur d'enquête au passage $t$, la matrice des variances-covariances estimée étant représentée par $S$. Les moyennes pour chaque passage ($\mu_t$ pour le passage $t$) suivent un processus ARMA. Comme le modèle (16) est un cas particulier du modèle de régression avec coefficients aléatoires, la fonction de vraisemblance marginale appropriée est différente de l'équation (5).

On obtient une fonction de vraisemblance marginale ou une fonction de vraisemblance conditionnelle approximative pour les paramètres de corrélation dans un modèle de régression avec coefficients aléatoires de la façon suivante. Supposons, pour ce qui a trait au modèle (2), que $\beta$ est un vecteur aléatoire défini par l'équation $\beta = W\delta + u$, où $W$ est une matrice $p \times q$ de valeurs connues, $\delta$ est un vecteur de paramètres de dimensions $q \times 1$ et $u \sim N(\mathbf{0}, \gamma^2\Gamma)$, sans lien avec $\epsilon$. Selon le modèle composé $y = XW\delta + Xu + \epsilon$, la fonction de vraisemblance logarithmique pour $\delta$, $\Omega$, $\Gamma$, $\gamma^2$, et $\kappa = \sigma^2/\gamma^2$, désignée par $L(\delta,\kappa,\gamma^2,\Gamma,\Omega)$, est définie par l'équation (3), où $\Omega$ est remplacé par l'expression $\kappa\Omega + X\Gamma X^T$ et $X\beta$ par $XW\delta$. De la même façon, la fonction de vraisemblance marginale, désignée par $L_M(\kappa,\Gamma,\Omega)$, est définie par l'équation (5), où $X$ est remplacé par $XW$ et $\Omega$ par $\kappa\Omega + X\Gamma X^T$. Ainsi,

$$L_M(\kappa,\Gamma,\Omega) = \left\{ |\,\kappa\Omega + X\Gamma X^T\,|^{1/2}\;|\,(XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW\,|^{1/2}\; g^{m-q} \right\}^{-1}, \qquad (17)$$

où

$$g = y^T(\kappa\Omega + X\Gamma X^T)^{-1}y$$

$$- y^T(\kappa\Omega + X\Gamma X^T)^{-1}XW((XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW)^{-1}(XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}y.$$

Or, la dimension de $\Omega$ peut être plus grande par rapport à celle de $\Gamma$; c'est une situation qu'on observe parfois dans l'échantillonnage répété. Pour redresser cette situation, on pourrait

prendre la fonction de vraisemblance définie en (3), la multiplier par la distribution de $\beta$, puis intégrer le tout par rapport à $\beta$ pour obtenir la fonction de vraisemblance pour les paramètres du modèle des coefficients aléatoires. On aurait ainsi des matrices de la même dimension que $\Gamma$.

Puisque $S$ est connue, nous pouvons obtenir facilement une estimation d'$\Omega$, la matrice de corrélation des erreurs d'enquête. Nous pouvons aussi obtenir une valeur estimée de $\kappa = \sigma^2/\gamma^2$. Les hypothèses qui sous-tendent la fonction de vraisemblance marginale définie en (17) nous obligent à supposer que $e_t$ dans l'équation (16) est une variable aléatoire stationnaire. Par conséquent, la moyenne des éléments diagonaux de $S$ donne une valeur estimée de $\sigma^2$. Si $\gamma^2$ est la variance des moyennes $\mu$, alors la variation entre $\bar{y}_t$, $t = 1, \ldots, k$ donne une valeur estimée de $\sigma^2 + \gamma^2$. De ces deux valeurs estimées, nous pouvons déduire une valeur estimée pour $\kappa$. Suivant le modèle (16), $X$ dans l'équation (17) est la matrice unité $k \times k$ tandis que $W$ est un vecteur colonne $k \times 1$ formé de uns. La fonction de vraisemblance marginale ainsi obtenue est une pseudo fonction puisque certains des paramètres ont été remplacés par des valeurs estimées. Dans ce cas, nous pouvons déterminer la pseudo fonction de vraisemblance marginale pour les paramètres dans $\Gamma$ (pseudo car $\kappa$ et $\Omega$ ont été remplacés par leurs estimateurs) en prenant l'équation (17) et en effectuant les substitutions appropriées. Les paramètres dans $\Gamma$ sont les paramètres de corrélation du processus ARMA appliqué à $\mu_t$. Si $k$, le nombre de passages, est relativement élevé par rapport au nombre de paramètres dans $\Gamma$, les estimateurs fondés sur la fonction marginale et sur la fonction conditionnelle approximative devraient être semblables à l'estimateur du maximum de vraisemblance. Au point de vue du calcul, il semble que la fonction de vraisemblance intégrale qui utilise les modèles d'espace d'états décrits par Binder et Dick (1989a, section 3) soit la plus simple à appliquer dans les circonstances.

## REMERCIEMENTS

Cette étude a été rendue possible grâce à une subvention du Conseil de recherches en sciences naturelles et en génie du Canada. L'auteur aimerait remercier l'arbitre de ses observations sur une version préliminaire de l'article.

## BIBLIOGRAPHIE

BELLHOUSE, D.R. (1978). Marginal Likelihoods for distributed lag models. *Statistische Hefte*, 19, 2-14.

BELLHOUSE, D.R. (1989). Optimal estimation of linear functions of finite population means in rotation sampling. *Journal of Statistical Planning and Inference*, 21, 69-74.

BELLHOUSE, D.R. (1990). On the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. *Biometrika*, 77, 743-746.

BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics, Volume 6 (Sampling)* (éds. P.R. Krishnaiah et C.R. Rao). Amsterdam: North-Holland, 187-211.

BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées - modelisation et estimation. *Techniques d'enquête*, 15, 31-48.

BINDER, D.A., et DICK, J.P. (1990). Méthode pour l'analyse des modèles ARMMI. *Techniques d'enquête*, 16, 251-265.

BLIGHT, B.J.N., et SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Séries B, 35, 61-68.

COCHRAN, W.G. (1977). *Sampling Techniques*. 3ième édition. New York: Wiley.

COX, D.R., et REID, N. (1987). Parameter orthogonality and approximate conditional inference (avec discussion). *Journal of the Royal Statistical Society*, Séries B, 49, 1-39.

CRUDDAS, A.M., REID, N., et COX, D.R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika*, 76, 231-237.

CUNIA, T., et CHEVROU, R.B. (1969). Sampling with partial replacement on three or more occasions. *Forest Science*, 15, 204-224.

FRASER, D.A.S. (1967). Data transformations and the linear model. *The Annals of Mathematical Statistics*, 38, 1456-1465.

HARVEY, A.C., et PHILLIPS, G.D.A. (1979). Maximum likelihood estimates of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.

KALBFLEISCH, J.D., et SPROTT, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society*, Séries B, 32, 175-208.

KILPATRICK, D.J. (1981). Optimum allocation in stratified sampling of forest populations on successive occasions. *Forest Science*, 27, 730-738.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Séries B, 12, 241-255.

ROBERTS, G., RAO, J.N.K., et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

TUNNICLIFFE WILSON, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society*, Séries B, 51, 15-27.

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 89–100
Statistique Canada



# Renouvellement de l'échantillon et estimation dans l'Enquête sur l'emploi, la rémunération et les heures de travail

IOANA SCHIOPU-KRATINA et K.P. SRINATH[1]

RÉSUMÉ

L'enquête sur l'emploi, la rémunération et les heures de travail que mène actuellement la Division du travail de Statistique Canada est une enquête mensuelle d'envergure permettant de recueillir des données à partir d'un grand échantillon d'établissements d'entreprises. Dans le présent article, l'auteur décrit les techniques d'enquête utilisées. Il donne une brève description des méthodes de stratification, de détermination de la taille de l'échantillon et de répartition, tandis qu'il décrit plus en détail la méthode de renouvellement en raison de sa complexité. Il souligne aussi certaines des simplifications qu'on peut apporter au plan de sondage.

MOTS CLÉS: Établissement; fardeau de réponse; base de sondage.

## 1. INTRODUCTION

### 1.0 Objectifs de l'enquête

L'enquête sur l'emploi, la rémunération et les heures de travail (EERH) est une enquête mensuelle menée auprès des établissements par Statistique Canada.

Les principaux objectifs de l'EERH sont les suivants:

i) établir des estimations mensuelles du nombre total d'employés rémunérés, de la rémunération hebdomadaire moyenne, de la rémunération horaire moyenne, de la moyenne des heures de travail par semaine et d'autres variables connexes pour chaque division d'activités économiques au niveau provincial;

ii) fournir les estimations susmentionnées pour le Canada au niveau des catégories à trois chiffres de la Classification des activités économiques (CAÉ);

iii) déterminer les erreurs types de toutes les estimations produites.

Elle a aussi pour objet de produire annuellement des estimations pour chaque catégorie à trois chiffres de la CAÉ au niveau provincial.

L'enquête englobe toutes les activités économiques, à l'exception des secteurs suivants: agriculture, pêche et trappage, services domestiques privés, organisations religieuses et services militaires.

Dans le présent article, l'auteur décrit la méthode de sélection et de renouvellement de l'échantillon ainsi que la méthode d'estimation adoptée pour l'enquête. Le chapitre 2 présente la méthode de sélection et de renouvellement de l'échantillon en détail, tandis que le chapitre 3 est consacré à la méthode d'estimation. On trouve dans l'annexe certains détails relatifs au chapitre 2. L'annexe présente aussi une autre forme des estimations mensuelles des unités actives.

Pour une description complète de la méthode utilisée pour l'EERH, voir le document de référence [3].

[1] Ioana Schiopu-Kratina et K.P. Srinath, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, K1A 0T6.

### 1.1 Définitions préliminaires

Voici la définition de certains des termes utilisés dans le présent article.

i) Établissement – Plus petite unité étant une entité juridique distincte capable de déclarer tous les éléments des données de base sur les activités économiques. L'établissement est l'unité statistique pour l'EERH. Nous allons utiliser le terme unité pour établissement.

ii) Unité déclarante (pouvant fournir des renseignements sur l'emploi) – Aux fins de la compilation de statistiques géographiques détaillées, l'établissement est souvent subdivisé en unités déclarantes en fonction principalement de l'emplacement et, parfois, d'autres aspects comme, entre autres, la liste de paye.

iii) Classification des activités économiques (CAÉ) – 1970. On attribue à chaque établissement un code CAÉ selon la nature de son activité. Ces codes sont définis dans le Manuel de la CAÉ (document de référence [5]).

Dans l'EERH, le nombre total d'employés rémunérés lié à une unité est la caractéristique choisie pour mesurer la taille de ladite unité.

L'EERH comporte quatre groupes de taille dont les limites sont définies de la façon suivante: 0 à 19 employés (groupe de taille 1), 20 à 49 employés (groupe de taille 2), 50 à 199 employés (groupe de taille 3) et 200 employés et plus (groupe de taille 4).

iv) On définit la superstrate en fonction d'une division d'activités économiques, d'une province et d'un groupe de taille. Puisqu'il y a 16 divisions d'activités économiques, 12 provinces et quatre groupes de taille, il y a donc 768 superstrates.

v) On définit la strate en fonction d'un code CAÉ à trois chiffres, d'une province et d'un groupe de taille. Il s'agit du niveau le plus détaillé pour lequel on établit des estimations.

vi) La partie à tirage complet de la population se compose d'unités qui sont toutes incluses dans l'échantillon avec certitude. Elle contient des unités du groupe de taille 4 et des unités de la population définies à l'avance. La partie à tirage partiel de la population comprend les unités restantes qui sont soumises à l'échantillonnage, comme il est décrit aux sections suivantes.

## 2. SÉLECTION ET RENOUVELLEMENT DE L'ÉCHANTILLON

### 2.0 Détermination de la taille de l'échantillon et méthode de répartition

Dans l'EERH, on détermine la taille de l'échantillon à tirage partiel pour chaque groupe d'activités économiques au niveau d'une province selon un coefficient désigné de variation de l'estimation du nombre total d'employés pour ce groupe d'activités économiques et pour cette province. La taille d'échantillon cherchée et les fractions de sondage sont calculées au niveau de la superstrate à l'aide d'une répartition proportionnelle à la taille, cette dernière étant le nombre total d'employés. Les fractions de sondage sont gardées constantes d'un mois à l'autre. On trouve plus de détails sur la méthode de répartition dans le document de référence [3].

La sélection réelle est faite au niveau de la strate. En raison du minimum d'unités que doit compter l'échantillon à ce niveau, le nombre d'unités échantillonnées est plus élevé que la taille d'échantillon voulue pour chaque groupe d'activités économiques au niveau provincial (voir (2.2)).

### 2.1 Sélection de l'échantillon

Considérons maintenant une strate particulière. Soient $N$ la taille de la partie à tirage partiel de la population et $n$ la taille de l'échantillon à tirage partiel dans cette strate.

Tandis que la répartition de l'échantillon entre les superstrates est proportionnelle à la taille, la répartition au niveau des strates dans chaque superstrate est essentiellement proportionnelle au nombre total d'unités retrouvées dans la partie à tirage partiel de cette strate.

On définit la fraction de sondage dans chaque strate par la formule:

$$f = \max\left(f', \frac{1}{100}\right), \tag{2.1}$$

où $f'$ est calculée au niveau de la superstrate correspondante. Afin de réduire l'instabilité des estimations qui est causée par les valeurs peu élevées de la fraction de sondage, on a décidé de fixer à 1/100 la fraction de sondage minimale pour toutes les strates.

Les calculs détaillés de la taille de l'échantillon au niveau des strates sont donnés à la section 2.3 (voir la façon dont est déterminée la formule (2.8)). Un échantillon est prélevé systématiquement de chaque strate.

### 2.2 Le plan de renouvellement

Le renouvellement de l'échantillon (remplacement périodique partiel de l'échantillon) dans l'EERH sert principalement à réduire le fardeau de réponse. D'après des enquêtes antérieures, il est apparu que le taux de réponse moyen dans les strates où il n'y avait pas de renouvellement était de beaucoup inférieur à celui que l'on constatait dans les strates où il y avait renouvellement. De plus, la présence dans l'échantillon d'un grand nombre d'unités pendant deux mois consécutifs rehausse la fiabilité des estimations de la variation d'un mois à l'autre par rapport aux estimations de la variation qui sont fondées sur des échantillons mensuels indépendants. Le renouvellement de l'échantillon dans chaque strate doit être effectué suivant certaines contraintes comme, par exemple, laisser les unités hors de l'échantillon pendant une certaine période après qu'elles ont été supprimées de celui-ci par renouvellement.

L'échantillon mensuel est composé de 14 groupes numérotés de 0 à 13. Le groupe 0 contient les unités à tirage complet de la strate. Les groupes 1 à 13 sont appelés "groupes de renouvellement". Les numéros 1 à 12 attribués aux groupes de renouvellement indiquent les mois où les unités autres que les créations ont été introduites dans l'échantillon par renouvellement. Par exemple, le groupe de renouvellement 1 renferme surtout des unités qui ont été échantillonnées en janvier et des créations, le groupe de renouvellement 2 contient principalement des unités qui ont été introduites dans l'échantillon en février et des créations, et ainsi de suite. Le groupe de renouvellement 13 se compose d'unités qui ont été présentes dans l'échantillon pendant 12 mois. Ces unités sont les plus anciennes en ce qui a trait au temps passé dans l'échantillon et peuvent être supprimées par renouvellement. Chaque mois, des créations sont échantillonnées et attribuées au hasard aux groupes de renouvellement.

Au moment de la sélection et du renouvellement mensuels, toutes les unités dans le mois de référence sont transférées dans le groupe de renouvellement 13. En février, par exemple, toutes les unités du groupe de renouvellement 2 sont transférées dans le groupe de renouvellement 13. Un groupe de remplacement est choisi à partir d'unités "admissibles à la sélection" et d'unités nouvellement enregistrées (des créations). Les unités du groupe de remplacement (à l'exception de 11/12 des créations) sont ensuite placées dans le groupe de renouvellement 2, et elles ne pourront être supprimées de l'échantillon par renouvellement avant au moins 12 mois. Si l'on dispose de suffisamment d'unités pour un groupe de remplacement, les unités contenues dans le groupe 13 sont éliminées de l'échantillon et ne peuvent pas être ré-échantillonnées avant au moins 12 mois. Sinon, certaines unités du groupe 13 sont conservées dans l'échantillon jusqu'à ce que l'on ne dispose pas d'un nombre suffisant d'unités en-dehors de l'échantillon pour former un groupe de remplacement. Cette mesure sert à maintenir la taille de l'échantillon à un minimum ou à s'assurer de disposer d'un échantillon suffisamment grand pour établir des estimations qui répondent à des normes de fiabilité préétablies. De cette façon, en général au moins 11/12 des unités demeurent dans la partie à tirage partiel de l'échantillon pendant deux mois consécutifs.

Les unités qui sont laissées dans l'échantillon sont attribuées à un groupe d'attente, lequel est divisé en sous-groupes. Un sous-groupe est composé d'unités qui ont toutes été supprimées de l'échantillon le même mois. Le groupe d'attente contient 12 sous-groupes dans chaque strate. Le temps que chaque unité passe en-dehors de l'échantillon est ainsi consigné afin de s'assurer que les unités ne seront pas prélevées de nouveau avant au moins 12 mois. Les unités qui ont passé la période minimale requise dans le groupe des unités "non admissibles à la sélection" sont transférées dans le groupe des unités "admissibles à la sélection", et on leur attribue ainsi une probabilité positive de ré-échantillonnage.

Pour résumer, toute la population à tirage partiel comporte à n'importe quel moment donné quatre groupes d'unités. Il s'agit:

i) des unités qui sont présentes dans l'échantillon pour le mois en question;

ii) des unités admissibles à la sélection (AS);

iii) des unités dans le groupe d'attente qui ont été supprimées de l'échantillon par renouvellement moins de 12 mois auparavant et qui ne sont pas admissibles à la sélection (NAS); et

iv) des créations, c'est-à-dire des unités qui n'ont pas été enregistrées dans la base auparavant.

Le processus mensuel de sélection et de renouvellement comporte un échange d'unités entre ces groupes. Certaines unités quittent le groupe i) pour se retrouver dans le groupe iii) et de nouvelles unités entrent dans le groupe i) en provenance du groupe ii), après que certaines créations échantillonnées du groupe iv) ont été transférées au groupe ii) et que le reste des créations ont été attribuées aux groupes ii) et iii) après sélection. Cette façon de procéder permet de s'assurer que l'échantillon est représentatif de la population à n'importe quel mois donné.

### 2.3 Détermination de la taille de l'échantillon et des poids

#### 2.3.1 Mises à jour mensuelles

La base de sondage contient un grand nombre d'unités qui sont inactives ou hors du champ d'observation, qui ne sont plus en exploitation, *etc.* À part le fardeau de conserver un nombre croissant d'unités inactives dans la base, il est probable que les estimateurs fondés sur des échantillons tirés d'une telle population aient une variance élevée, à cause du fait que l'échantillon renferme une proportion élevée de valeurs observées nulles. L'idéal serait d'éliminer toutes ces unités de la base de sondage avant que l'échantillon mensuel ne soit prélevé. La base est mise à jour chaque mois, entre les opérations mensuelles de sélection et de renouvellement. C'est pour cette raison que les indices que nous utilisons pour désigner les créations et les disparitions dans la base sont d'une unité plus élevés que ceux employés pour désigner la taille de l'échantillon au moment du tirage précédant la mise à jour. Par exemple, après le prélèvement du premier échantillon, supposons qu'il y a $n(0)$ unités dans l'échantillon, desquelles on constate par la suite que $d(1)$ unités sont des unités disparues. Alors $D(1)$ désigne le nombre d'unités disparues dans la partie hors échantillon de la population et $B(1)$ le nombre d'unités enregistrées en tant que créations le mois en question. En calculant la taille d'échantillon voulue pour le mois suivant, $n(1)$, on doit tenir compte de ces mises à jour (voir (2.3)) ainsi que de la taille de la population au moment de la sélection du premier échantillon, $N(0)$.

#### 2.3.2 Détermination de la taille de l'échantillon

La population d'une strate à tirage partiel donnée est une fonction du temps et sera désignée, disons, par $N(t)$, où $t$ est un nombre entier positif qui s'accroît d'une unité d'un mois à l'autre. On détermine la taille d'échantillon voulue pour chaque mois par la formule:

$$n'(t) = [fN(t) + 0.5]. \tag{2.2}$$

Ici, $[a]$ représente le nombre entier le plus grand qui n'est pas supérieur à $a$. La constante 0.5 est utilisée pour obtenir une meilleure approximation au moment de l'arrondissement.

Supposons que $d(t)$ unités sont éliminées de l'échantillon (disparitions à l'intérieur de l'échantillon) et $D(t)$ unités du reste de la population de la strate (disparitions hors échantillon). Supposons aussi que $B(t)$ nouvelles unités (des créations) sont enregistrées durant le même intervalle.

Par conséquent, la taille de la population de la cellule au moment de la $t$-ième sélection est calculée à l'aide de la formule:

$$N(t) = N(t-1) - d(t) - D(t) + B(t). \tag{2.3}$$

Puisque les mises à jour ne sont pas exhaustives, on s'attend à ce que la population compte des unités inactives (disparues) non décelées.

Soit $n_t(t)$ le nombre d'unités actives laissées dans l'échantillon du mois précédent (après les mises à jour), c'est-à-dire:

$$n_t(t) = n(t-1) - d(t). \tag{2.4}$$

Nous supposons qu'il n'y a pas d'unités disparues non décelées dans l'échantillon à ce point-ci. Nous pouvons croire que la population de la strate est répartie dans deux domaines: le domaine des unités "actives" et celui des unités "disparues". La taille du domaine des unités "disparues" n'est pas connue, mais on peut calculer un estimateur $\hat{U}_d(t)$ d'après les renseignements fournis par l'échantillon et les mises à jour (voir l'annexe). Soit $\hat{U}_d(t)$ un estimateur du nombre d'unités "disparues" non décelées dans la population au moment de la $t$-ième sélection mensuelle. Alors:

$$N(t) = \hat{U}_t(t) + \hat{U}_d(t), \tag{2.5}$$

où $\hat{U}_t(t)$ est l'estimateur du nombre d'unités actives.

La probabilité de choisir une unité disparue lorsqu'on prélève une unité au hasard parmi les unités hors échantillon est définie à l'aide de la formule:

$$\hat{P}_d(t) = \min\left\{\frac{\hat{U}_d(t)}{N(t) - n_t(t)},\ 1\right\}. \tag{2.6}$$

On détermine le nombre voulu d'unités actives dans l'échantillon par la formule:

$$n_t'(t) = f\hat{U}_t(t). \tag{2.7}$$

On calcule la taille de l'échantillon de remplacement de façon à s'assurer que le nombre prévu d'unités actives dans l'échantillon après la sélection est $n_t'(t)$.

Supposons maintenant qu'au moment des $t$-ièmes sélection et renouvellement des $n_t(t)$ unités actives dans l'échantillon, $n_o(t)$ unités peuvent être supprimées de l'échantillon par renouvellement.

Étant donné qu'il reste $n_t(t) - n_o(t)$ unités actives dans l'échantillon, on a besoin de $n_t'(t) - n_t(t) + n_o(t)$ plus d'unités actives dans l'échantillon pour le $t$-ième mois.

Afin de représenter adéquatement les créations dans l'échantillon, $b(t) = fB(t)$ créations doivent être prélevées au hasard et incluses dans l'échantillon.

Par conséquent:

$$\ell(t) = \max(n_t'(t) - n_t(t) + n_o(t) - b(t), 0),$$

unités actives doivent être tirées du groupe des unités admissibles à la sélection et être ajoutées à l'échantillon avec les créations prélevées. Compte tenu de l'existence d'un nombre inconnu d'unités inactives dans la population et de l'arrondissement au nombre entier, il faut que:

$$n_i(t) = \min\left( \left| \frac{\ell(t)}{1 - \hat{P}_d(t)} + 0.5 \right|, n''(t) \right),$$

plus d'unités soient introduites dans le même échantillon par renouvellement, $\hat{P}_d(t)$ étant définie par la formule (2.6) et $n''(t)$ étant égale à $N(t) - n_o(t)$.

En calculant $n_i(t)$, nous avons supposé qu'il n'y a pas d'unités inactives parmi les créations, de sorte que le facteur d'élargissement $[1 - \hat{P}_d(t)]^{-1}$ ne s'applique qu'aux unités "plus anciennes" dans le groupe des unités AS.

On détermine la taille de l'échantillon, $n(t)$, pour le $t$-ième mois à l'aide de la formule:

$$n(t) = \max\{n_t(t) - n_o(t) + n_i(t) + b(t), m\}, \tag{2.8}$$

où $m$ représente la taille minimale d'échantillon cherchée pour une strate, taille qui est présentement posée à 3. Cette exigence additionnelle fait accroître la taille de l'échantillon de 3,000 unités dans toutes les strates, dont 1,800 qui devraient être présentes dans l'échantillon pendant un laps de temps considérable.

Des $n_i(t)$ unités qui sont introduites dans l'échantillon par renouvellement, on s'attend à trouver des unités inactives (dont le nombre est calculé à l'aide de la formule $\hat{n}_d(t) = \hat{P}_d(t)n_i(t)$) et des unités actives (dont le nombre est déterminé par la formule $\hat{n}_t(t) = n_i(t) - \hat{n}_d(t)$). Ainsi, des $n(t)$ unités présentes dans l'échantillon après les $t$-ièmes prélèvement et renouvellement mensuels, on prévoit qu'un certain nombre d'entre elles (calculé en utilisant la formule $\hat{n}_t(t) = n_t(t) - n_o(t) + \hat{n}_t(t) + b(t)$) seront "actives", et elles représentent les $\hat{U}_t(t)$ unités du domaine des unités "actives" au taux approprié $f$ (voir (2.7) et (2.8)), lorsque $n(t) > m$ dans (2.8).

### 2.3.3 Détermination des poids

Le poids $w(t)$ utilisé aux fins de l'estimation pour le $t$-ième mois est exprimé en fonction de la taille de la population et de l'échantillon au passage $t$. Toutefois, l'utilisation de $N(t)/n(t)$ comme poids aux fins de l'estimation pourrait amener une surestimation des unités actives de la population. En effet, $n(t)$ dans la formule (2.8) a été choisi pour que le nombre prévu d'unités actives présentes dans l'échantillon égale le nombre cherché. Il se peut que le nombre d'unités disparues présentes dans l'échantillon prélevé comme il est décrit ci-dessus ne représente pas la taille du domaine des unités disparues au taux approprié. Dans la formule (2.8), $n_i$ est tiré de la population en général et on s'attend alors à ce qu'il préserve la proportion entre le domaine des unités "disparues" et celui des unités "actives". Aucune disparition ne devrait être trouvée parmi les créations et, par conséquent, $b(t)$ représente bien le sous-groupe des créations. Il reste toutefois dans l'échantillon $n_t(t) - n_o(t)$ unités d'une sélection précédente, après le renouvellement et les mises à jour de l'échantillon. Il est probable que la proportion des disparitions parmi ces unités soit de beaucoup inférieure à la proportion correspondante dans la population en général, malgré le fait que les mises à jour sont fondées sur des renseignements provenant de sources autres que l'enquête. La valeur de $N(t)/n(t)$ doit alors être corrigée en fonction de la sous-représentation des unités "disparues" dans l'échantillon. On obtient ainsi, lorsque $n(t) > m$,

$$\frac{N(t)}{n(t) + \hat{u}(t)} = \frac{1}{f}, \tag{2.9}$$

où $\hat{u}(t)$ sera déterminé ultérieurement (voir (2.10)). La valeur d'$\hat{u}(t)$ représente les "disparitions non réelles" qui ont été ajoutées à l'échantillon pour représenter correctement les unités "disparues" dans la population. Notons que, lorsque le premier échantillon est prélevé ou qu'on procède à un nouveau prélèvement, un tel redressement n'est pas nécessaire, c'est-à-dire que $\hat{u}(0) = 0$.

Afin de trouver une formule pour $\hat{u}(t)$, nous utilisons la formule (2.5) dans le numérateur de (2.9) et la formule (2.8) dans le dénominateur. D'après (2.7) et (2.8), nous devons aussi avoir:

$$\frac{\hat{U}_d'(t)}{\hat{n}_d(t) + \hat{u}(t)} = \frac{1}{f}.$$

Étant donné que $\hat{n}_d(t)$ est égal à $\hat{P}_d(t)n_i(t)$, nous obtenons, à partir de l'équation ci-dessus:

$$\hat{U}_d'(t) = \frac{1}{f}[\hat{n}_d(t) + \hat{u}(t)] \quad \text{ou} \quad \hat{u}(t) = f\hat{U}_d'(t) - \hat{n}_d(t). \tag{2.10}$$

On peut faire la correction en fonction des disparitions à l'aide de la formule:

$$v(t) = \begin{cases} \hat{u}(t) & \text{si} \quad \hat{u}(t) \geq -\hat{n}_d(t) \\ 0 & \text{si} \quad \hat{u}(t) < -\hat{n}_d(t) \end{cases} \tag{2.11}$$

et on calcule le poids utilisé dans l'estimation suivant la formule

$$w(t) = \frac{N(t)}{n(t) + \hat{v}(t)}. \tag{2.12}$$

Notons que le poids dans (2.12) est défini à l'aide d'un estimateur et qu'il s'agit donc d'une variable aléatoire.

L'utilisation du poids défini par la formule (2.12) suppose que l'estimateur du nombre d'unités actives dans la population, laquelle est établie par l'équation $\hat{U}_t(t) = w(t)\hat{n}_t(t)$, ne dépasse pas $N(t)$, la taille de la population au moment du $t$-ième échantillonnage.

Définissons:

$$\hat{U}_d(t) = w(t)\,[\hat{v}(t) + \hat{n}_d(t)].$$

D'après (2.10) et (2.11), il s'ensuit que $\hat{U}_d(t) \geq 0$ et que sa valeur minimale est 0 lorsque $\hat{v}(t) - \hat{n}_d(t) = 0$. D'après (2.5), la valeur maximale de $\hat{U}_t(t)$ est alors $N(t)$.

On traitera dans la section 3.1 du fait que la restriction selon laquelle l'estimateur des unités actives soit tronqué à $N(t)$ a des répercussions sur l'estimation. L'estimateur $\hat{U}_d(t)$ est calculé récursivement (voir l'annexe) à l'aide des formules (2.10) et (2.11), et $\hat{v}(0) = \hat{u}(0) = 0$.

Il faut noter que la formule (2.11) est légèrement différente de celle qui permet de faire la correction en fonction des disparitions dans l'EERH. Premièrement, pour simplifier, nous n'avons pas considéré ici les cas où l'on doit utiliser la taille minimale de l'échantillon, $m$. Dans de tels cas, il n'est pas approprié de se servir de la fraction de sondage $f$ dans (2.10). Dans l'EERH, on utilise dans tous les cas le poids du mois précédent (2.10) au lieu de $f$. Deuxièmement, la correction en fonction des disparitions dans l'EERH est toujours considérée positive. La formule (2.11) montre qu'elle peut être négative, tant qu'elle est supérieure à $-\hat{n}_d(t)$. Il est très rare que cette situation se présente ou, de façon plus générale, lorsque $\hat{u} \leq 0$.

L'annexe présente une formule qui permet de calculer l'estimateur des unités actives et qui ne nécessite pas la correction en fonction des disparitions.

### 2.4 Échantillonnage des créations

Comme nous l'avons mentionné précédemment, de nouvelles unités sont enregistrées dans la base chaque mois. Étant donné qu'on estime que ces nouvelles unités (créations) peuvent différer des "anciennes", nous avons conçu une stratégie particulière pour représenter adéquatement les créations dans l'échantillon.

L'idéal serait que, si $B$ créations sont enregistrées durant le mois courant et si $f$ désigne la fraction de sondage de la strate, $b = fB$ créations devraient être présentes dans l'échantillon durant le mois en question. Les créations échantillonnées sont attribuées au hasard aux groupes de renouvellement décrits dans la section 2.2. De cette façon, on s'assure que la probabilité de suppression de l'échantillon par renouvellement est la même pour les nouvelles unités et pour les "anciennes", de sorte que la répartition par âge des unités à l'intérieur de l'échantillon est la même que celle des unités hors échantillon.

En utilisant la notation de la section précédente, soient $n_i$ le nombre d'unités qui doivent être introduites dans l'échantillon par renouvellement au moment de la sélection mensuelle de l'échantillon, à l'exclusion des créations, et $N'$ le nombre d'unités dans le groupe des unités AS (groupe ii) de la section 2.2). La stratégie relative aux créations consiste en une sélection en deux phases. On prélève des créations ou des unités "plus anciennes" pour former échantillon préliminaire, duquel un échantillon est ensuite tiré. Cette façon de procéder est nécessaire parce qu'un groupe, normalement celui des créations, est petit. Ensuite, $n'$ unités sont pré-échantillonnées du groupe des unités AS et placées dans le groupe des créations (voir (2.15)). Si:

$$\frac{n_i}{N'} > f, \tag{2.13}$$

un certain nombre de créations, $b'$, sont pré-échantillonnées du groupe des créations et placées dans le groupe des unités AS.

La valeur de $b'$ est déterminée par la formule:

$$b' = \frac{b\,N'}{n_i}. \tag{2.14}$$

L'inégalité dans (2.13) garantit que $b' \leq B$, ce qui signifie que le groupe des créations est suffisamment grand et que le pré-échantillonnage peut avoir lieu. À partir de l'échantillon préliminaire de $N' + b'$ unités, $n_i + b$ unités sont ensuite échantillonnées.

Le choix de $b'$, tel qu'il est déterminé par la formule (2.14), garantit que le nombre prévu de créations présentes dans l'échantillon est celui qui est souhaité, puisque la probabilité de prélever une création de l'échantillon préliminaire est $b'/(b' + N')$; par conséquent, le nombre prévu de créations lorsque $n_i + b$ unités sont tirées sans remise de cet échantillon préliminaire est $(n_i + b)\,b'/(b' + N')$, qui, d'après (2.14), est égal à $b$.

Dans la situation complémentaire, lorsque la formule (2.13) n'est pas valide, nous avons que:

$$n' = n_i/f \leq N'. \tag{2.15}$$

Alors, $n'$ unités "plus anciennes" sont prélevées du groupe des unités AS et placées dans le groupe des créations. Notons que, dans cette situation, $b'$, tel que défini par la formule (2.14), est supérieur à $B$ et que, par conséquent, on ne peut pas appliquer la première méthode.

Nous calculons maintenant le nombre prévu d'unités "anciennes" dans l'échantillon. Étant donné que la probabilité de prélever une unité "ancienne" est désormais $n'/(n' + B)$ et que $n_i + b$ unités sont tirées de l'échantillon préliminaire de $n' + B$ unités et placées dans l'échantillon, le nombre prévu d'unités "anciennes" est $(n_i + b)\,n'/(n' + B)$, qui est égal à $(n_i + b)$.

Il faut noter qu'une sous-représentation des créations peut se produire dans certains cas. Par exemple, lorsque, dans une strate quelconque, aucune unité ne doit être introduite dans l'échantillon par renouvellement, alors, pour le mois en question, les créations ne seront pas représentées dans l'échantillon prélevé de cette strate. Toutefois, cette situation survient d'habitude lorsque la taille de la population est peu élevée et que, à la longue, on peut s'attendre à ce que, en moyenne, les créations soient représentées correctement.

Les $b$ créations réellement prélevées sont attribuées au hasard aux groupes de renouvellement 1 à 12 dans l'échantillon, ce qui se traduit, en moyenne, à $b/12$ de créations qui sont attribuées à chacun de ces groupes. Ainsi, on s'assure que la probabilité de voir une création être supprimée de l'échantillon par renouvellement est la même que pour n'importe quelle autre unité.

Afin de garder constante la répartition par âge des unités des groupes i) à iii) (voir la section 2.1), les créations non échantillonnées seront attribuées au hasard au groupe des unités AS et à celui des unités NAS. Cela signifie que, si $N'$ est le nombre d'unités dans le groupe des unités AS et $N''$ le nombre d'unités dans le groupe des unités NAS, $N'(B - b)/(N' + N'')$ créations non échantillonnées seront alors attribuées au groupe des unités AS et $N''(B - b)/(N' + N'')$ créations non échantillonnées au groupe des unités NAS.

## 3. ESTIMATION

### 3.0 Introduction

Dans le présent chapitre, nous décrivons la méthode d'estimation du "nombre total d'employés rémunérés".

Comme nous l'avons indiqué à la section 2.0, seule la fiabilité des estimations du nombre total d'employés au niveau d'agrégation des groupes d'activité économiques (échelle provinciale) est définie à l'avance. Les estimations relatives à des caractéristiques autres que le nombre total d'employés présentent différents degrés de fiabilité. Par exemple, on s'attend à ce que les estimations de la rémunération hebdomadaire moyenne soient plus fiables que celles du nombre total d'employés.

Le niveau d'agrégation le plus bas auquel les estimations sont publiées est le niveau de la CAÉ et de la province, mais les unités de base pour la production des estimations sont les strates.

Une valeur aberrante dans l'EERH est une valeur observée dans la partie à tirage partiel de l'échantillon, qui est supérieure à une valeur préétablie.

Le poids de chaque strate est d'abord calculé comme dans la section 2.3, puis il est corrigé en fonction des valeurs aberrantes dans la strate en question. Des estimations du nombre total d'employés sont établies pour chaque strate à l'aide des poids corrigés. Il n'y a pas de correction en fonction de la non-réponse, puisque les valeurs correspondant aux non-répondants font l'objet d'une imputation. On obtient l'estimation du nombre total d'employés pour chaque strate en additionnant les totaux suivants:

i) le nombre total d'employés dans les unités à tirage complet;

ii) le nombre total d'employés dans les unités à valeurs aberrantes;

iii) la somme des valeurs pondérées du nombre d'employés dans les unités à tirage partiel, à l'exclusion des unités à valeurs aberrantes.

Étant donné que le poids qui est attribué à chaque unité à valeurs aberrantes est de un, les unités à valeurs aberrantes sont traitées comme des unités à tirage complet aux fins de l'estimation et ne sont donc pas utilisées dans l'estimation de la variance.

### 3.1 Estimation du total pour une caractéristique

Considérons une strate particulière pour un mois donné dans l'enquête. Soient $N$ la taille de la partie à tirage partiel de la population de la strate et $n$ le nombre d'unités à tirage partiel présentes dans l'échantillon pour le mois en question. Si $\hat{v}$ correspond à la correction en fonction

des disparitions (voir (2.11)), on calcule le poids initial qui est attribué à chaque unité présente dans l'échantillon aux fins de l'estimation à l'aide de la formule (voir (2.12)):

$$w = \frac{N}{n + \hat{v}}. \tag{3.1}$$

Toutefois, si $t$ unités à valeurs aberrantes sont présentes dans l'échantillon ($t \geq 1$), on modifie alors ce poids en accordant à chaque unité à valeur aberrante un poids de 1 et en attribuant aux unités restantes dans la partie à tirage partiel de l'échantillon le poids calculé suivant la formule

$$w' = \frac{N - t}{n + \hat{v} - t}. \tag{3.2}$$

Supposons que $S$ représente l'ensemble des unités présentes dans l'échantillon. Si $Y(u)$ désigne la valeur du nombre d'employés correspondant à l'unité $u$ dans l'échantillon, on détermine alors l'estimateur du nombre total d'employés dans la strate par la formule

$$\hat{Y} = \sum_{u \in S} w(u) Y(u), \tag{3.3}$$

où $w(u) = 1$ si $u$ est une unité à valeur aberrante ou une unité à tirage complet et où $w(u) = w'$ dans le cas de toutes les autres unités présentes dans l'échantillon (voir (3.2)).

Des estimations des totaux à n'importe quel niveau d'agrégation supérieur à celui de la strate sont obtenues en additionnant les estimations des totaux pour toutes les strates dans le niveau d'agrégation en question.

Nous allons analyser certaines des propriétés de l'estimateur défini par le dernier terme du côté droit de la formule (3.3).

Étant donné que, aux fins de l'estimation, les unités à valeurs aberrantes peuvent être considérées comme des unités à tirage complet, nous pouvons remplacer, pour simplifier, $w'$ par $w$, ce dernier étant défini par la formule (3.1). Il faut alors modifier $N$ et $n$ en conséquence (voir (3.2)).

Soient $N_\ell$ la taille du domaine des unités "actives" qu'on retrouve dans la population de la strate et $n_\ell$ le nombre d'unités "actives" présentes dans l'échantillon. Supposons aussi que $\bar{Y}_\ell$ représente le nombre d'employés moyen pour les unités "actives" présentes dans l'échantillon.

Étant donné que seules les unités "actives" contribuent à l'estimation du nombre total d'employés, on obtient une estimation du total pour la cellule grâce à la formule:

$$\hat{Y}_\ell = \hat{U}'_\ell \bar{Y}_\ell. \tag{3.4}$$

Nous considérons le cas $fN > m$.

Dans la formule (3.4), $\hat{U}'_\ell$ est une valeur estimée de $N_\ell$, le nombre d'unités actives dans la population, et est définie par:

$$\hat{U}'_\ell = \frac{1}{f} n_\ell. \tag{3.5}$$

Dans la formule (3.5), $f$ désigne la fraction de sondage de la strate qui est gardée fixe. L'estimateur fondé sur les valeurs de $\hat{U}'_\ell$ est non biaisé, c'est-à-dire:

$$E(\hat{U}'_\ell) = N_\ell. \tag{3.6}$$

Étant donné que l'estimateur fondé sur (3.5) est non biaisé, $\hat{U}'_\ell$ peut excéder $N$ dans certains cas, puisque $\hat{U}'_\ell$ peut avoir des valeurs peu élevées pour compenser.

Dans l'EERH, la valeur estimée des unités "actives" $\hat{U}_t$ est définie par:

$$\hat{U}_t = wn_t, \tag{3.7}$$

le poids étant déterminé à l'aide de la formule (3.1).

La définition du poids dans l'EERH (voir la fin de la section 2.3) suppose que:

$$\hat{U}_t = \begin{cases} \hat{U}'_t & \text{si} \quad n_t \leq fN \\ N & \text{si} \quad n_t > fN. \end{cases} \tag{3.8}$$

L'estimation du nombre total d'employés dans l'EERH est définie par:

$$\hat{Y}_t = \hat{U}_t \bar{Y}_t. \tag{3.9}$$

Un estimation du nombre total d'employés fondée sur (3.5) est obtenue par la formule:

$$\hat{Y}'_t = \hat{U}'_t \bar{Y}_t. \tag{3.10}$$

L'estimateur fondé sur la formule (3.8) est biaisé et, par conséquent, l'estimateur du nombre total d'employés dans l'EERH l'est aussi.

Toutefois, l'erreur quadratique moyenne de l'estimateur fondé sur (3.9), conditionnellement à la valeur de $n_t$, est inférieure à celle de l'estimateur fondé sur (3.10), conditionnellement à la valeur de $n_t$. Nous allons maintenant tenter de prouver cette affirmation.

On s'aperçoit facilement que, pour chaque résultat particulier, le biais $B_t$ de l'estimateur fondé sur (3.9) et conditionnellement de la valeur de $n_t$ est déterminé par:

$$B_t = (\hat{U}_t - N_t)\bar{Y}_t. \tag{3.11}$$

De même, nous obtenons pour l'estimateur fondé sur (3.10) l'équation suivante:

$$B'_t = (\hat{U}'_t - N_t)\bar{Y}_t. \tag{3.12}$$

Nous montrons que l'erreur quadratique moyenne conditionnelle de l'estimateur fondé sur (3.9) est inférieure à celle de l'estimateur fondé sur (3.10). Le même résultat est aussi valide pour les erreurs quadratiques moyennes inconditionnelles. Nous prenons comme condition la taille d'échantillon réelle des unités "actives" $n_t$. À partir des formules (3.8) à (3.10), $\text{Var}[\hat{U}'_t \bar{Y}_t \mid n_t] - \text{Var}[\hat{U}_t \bar{Y}_t \mid n_t] = [n_t^2 f^{-2} - N^2]\, 1\{n_t > fN\} \text{Var}[\bar{Y}_t \mid n_t]$. Notons que $n_t^2 f^{-2} - N^2 > 0$ dans l'ensemble $\{n_t > fN\}$. Nous comparons maintenant $[B'_t]^2$ et $B_t^2$:

$$[B'_t]^2 - B_t^2 = 1\{n_t f^{-1} > N\}\{[\hat{U}'_t - N_t]^2 \bar{Y}_t^2 - (N - N_t)^2 \bar{Y}_t^2\}.$$

Cependant, $\hat{U}'_t - N_t = f^{-1} n_t - N_t > N - N_t$ si $n_t f^{-1} > N$.

Par conséquent, $[B'_t]^2 - B_t^2 \geq 0$. Étant donné que l'EQM $[\hat{U}'_t \bar{Y}_t \mid n_t] = \text{Var}[\hat{U}'_t \bar{Y}_t \mid n_t] + [B'_t]^2$ et que l'EQM $[\hat{U}_t \bar{Y}_t \mid n_t] = \text{Var}[\hat{U}_t \bar{Y}_t \mid n_t] + [B_t]^2$, la comparaison terme par terme nous amène à conclure que:

$$\text{MSE}[\hat{U}'_t \bar{Y}_t \mid n_t] \geq \text{MSE}[\hat{U}_t \bar{Y}_t \mid n_t].$$

Cette importante propriété motive le choix de l'estimateur fondé sur (3.9) plutôt que de l'estimateur fondé sur (3.10) pour évaluer le nombre total d'employés dans l'EERH.

## ANNEXE

Dans la présente annexe, nous utilisons la notation de la sous-section 2.3.2.

Nous déterminons d'abord la formule pour $\hat{U}_d(t)$, la taille du domaine des unités "disparues" utilisé dans l'EERH pour la $t$-ième sélection.

Rappelons-nous de l'expression $\hat{U}_d(t) = w(t)[\hat{v}(t) + \hat{n}_d(t)]$. Au moment de la $t$-ième mise à jour, on trouve $d(t+1)$ unités "disparues" dans l'échantillon. Nous pouvons donc remplacer $\hat{n}_d(t)$, le nombre estimé d'unités disparues présentes dans l'échantillon, par $d(t+1)$, afin d'obtenir une mise à jour d'$\hat{U}_d(t)$, c'est-à-dire $\hat{N}_d(t+1)$:

$$\hat{N}_d(t+1) = w(t)[\hat{v}(t) + d(t+1)]. \tag{1.1}$$

La formule (I.1) utilise la correction en fonction des disparitions du mois précédent. La valeur initiale de $\hat{v}$ est $\hat{v}(0) = 0$ (voir la remarque après la formule (2.9) et la définition d'$\hat{u}$) et $w(0) = f^{-1}$. La taille du domaine des unités "disparues" pour la $(t+1)$-ième sélection mensuelle est désormais estimée par:

$$\hat{U}_d(t+1) = \max(\hat{N}_d(t+1) - D(t+1) - d(t+1), 0). \tag{1.2}$$

Notons que $\hat{U}_t(t+1)$ peut être calculée à partir de (2.5) lorsque $\hat{U}_d(t+1)$ est connue, et réciproquement. On obtient une autre forme pour $\hat{U}_t(t+1)$ par des calculs récursifs de la façon suivante. Supposons que $\hat{U}_t(t)$ est connue avant la $t+1$-ième sélection du $t$-ième échantillon (rappelons-nous que l'expression $t = 0$ est utilisée pour désigner la première sélection). Alors, $\hat{U}_d(t)$ est aussi connue et peut être utilisée pour calculer $\hat{P}_d(t)$, la probabilité de prélever une unité disparue parmi les unités hors échantillon (voir la formule (2.6)). Cette probabilité est ensuite utilisée pour calculer le nombre voulu d'unités qui devraient être introduites dans l'échantillon par renouvellement comme il est décrit à la sous-section 2.3.2, ainsi que le nombre prévu d'unités actives présentes dans l'échantillon au moment de la $(t+1)$-ième sélection, $\tilde{n}_t(t)$.

Le poids utilisé dans l'estimation pour la sélection suivante est alors déterminé par $\hat{U}_t(t)/\tilde{n}_t$. Après la sélection, la $(t+1)$-ième mise à jour a lieu et on constate que le nombre réel d'unités actives présentes dans l'échantillon est $n_t(t+1)$. On peut estimer la taille du domaine des unités actives pour la prochaine sélection à l'aide de la formule suivante:

$$\hat{U}_t(t+1) = \min\left\{\frac{\hat{U}_t(t)}{\tilde{n}_t(t)} n_t(t+1) + B(t+1), N(t+1)\right\}$$

et ainsi de suite. Pour amorcer le processus, notons que le poids utilisé dans la première estimation est déterminé par l'équation $w(0) = f^{-1}$ et que, après la première mise à jour,

$$\hat{U}_t(1) = \min[w(0) \times n_t(1) + B(1), N(1)\}.$$

## BIBLIOGRAPHIE

COTTRELL-BOYD, T.M., DUNN, M.R., HUNTER, G.E., et SRINATH, K.P. (1980). Development of the redesign of the Canadian establishment based employment surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 8-15.

SCHIOPU-KRATINA, I., et SRINATH, K.P. (1986). The methodology of the Survey of Employment, Payroll and Hours. Working Paper No. BSMD-86-010E, Statistique Canada.

STATISTIQUE CANADA (1970). *Classification des activités économiques*. N° 12-501 au catalogue, Ottawa: Statistique Canada.

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 101–109
Statistique Canada



# Estimation d'un système d'équations linéaires à l'aide de données d'enquête

PHILLIP S. KOTT[1]

RÉSUMÉ

Dans cet article, nous élaborons un modèle permettant d'estimer un système d'équations linéaires à l'aide de données d'enquête. La théorie classique des sondages, axée sur des plans, nous est peu utile ici, bien que quelques-unes des techniques élaborées en vertu de cette théorie puissent être incorporées à des méthodes d'estimation robuste fondées sur un modèle. Les estimateurs de variance qui ont la forme de l'estimateur de "linéarisation" à équation unique sont quasi-non biaisés suivant de nombreuses structures d'erreur complexes. En outre, l'absence possible de variables explicatives peut être compensée par l'introduction de poids d'échantillonnage dans l'estimation par régression. Dans certains cas toutefois, l'absence de variables explicatives peut rendre incertaine l'estimation d'un système d'équations.

MOTS CLÉS: Poids d'échantillonnage; variable explicative présumée manquante; robuste; quasi-non biaisé.

## 1. INTRODUCTION

Kott (1991) a montré que les techniques à base de plans conçues pour estimer les équations de régression linéaire simples pouvaient servir dans des analyses fondées sur un modèle. Il a montré notamment que l'utilisation de la régression pondérée (pour échantillon) pouvait compenser l'absence possible de variables explicatives et que l'estimateur de variance dit "de linéarisation" pouvait produire des estimateurs quasi-non biaisés de l'erreur quadratique moyenne pour de nombreuses structures de variance complexes.

Dans le présent article, nous étendons les résultats de cette analyse à l'estimation d'un système ou "groupe" d'équations linéaires, sujet de très grand intérêt pour les économètres (voir, par exemple, Johnston (1972), p. 238-241). Deux exemples élémentaires seront peut-être éclairants pour le lecteur qui n'est pas habitué aux méthodes économétriques ou à leur équivalent.

Supposons que nous avons un échantillon d'exploitants agricoles et que nous voulons déterminer la relation entre la superficie ensemencée au soja et la taille de l'exploitation agricole. Zellner (1962) a montré, en fait, que même si une équation quadratique simple avec erreurs indépendantes et identiquement distribuées décrivait correctement la population, il existait peut-être un meilleur estimateur que celui obtenu par la méthode des moindres carrés ordinaires (MCO). On pouvait déterminer cet estimateur en considérant d'autres relations linéaires – par exemple, la relation entre la superficie ensemencée au *maïs* et la taille de l'exploitation agricole – dont le terme d'erreur était corrélé avec celui de la relation originale. L'estimation globale d'un tel groupe d'équations a été désignée par Zellner comme la "régression sans rélation apparente". Étrangement, pour que l'estimateur par les moindres carrés généralisés (MCG) que propose Zellner donne des résultats différents de ceux de l'estimateur par les MCO, certaines équations doivent contenir des variables explicatives que l'on ne retrouve pas dans d'autres équations. On peut aussi imaginer que toutes les équations renferment les mêmes variables explicatives mais que dans certains cas, des coefficients sont nuls.

[1] Phillip S. Kott, Special Assistant for Economic Survey Methods, U.S. Bureau of the Census, Room 3061-3, Washington, DC, 20233, É.-U.

Pour notre second exemple, supposons que nous avons un échantillon d'entreprises qui fabriquent un produit donné, $y$, à partir de deux intrants, $x_1$ et $x_2$, dont le prix unitaire est $p_1$ et $p_2$ respectivement. Les économistes vont souvent supposer que les entreprises se situent toutes au même niveau technologique (à quelque différence près). Étant donné $p_1$, $p_2$ et $y$, chaque entreprise choisira $x_1$ et $x_2$ de manière à minimiser le coût total, $c = p_1x_1 + p_2x_2$. Supposons que nous puissions exprimer la relation entre $p_1$, $p_2$ et $y$ et le coût total $c$ au moyen de l'équation suivante (en moyenne):

$$\log(c) = b_0 + b_1\log(p_1) + b_2\log(p_2) + b_3\log(y). \tag{1}$$

La théorie économique nous dit que devant une équation de coût implicite comme l'équation (1), une entreprise avisée déterminera le niveau de $x_1$ de manière que

$$x_1p_1/c = b_1. \tag{2}$$

Naturellement, pour que nous puissions estimer les équations (1) et (2), nous devons prévoir une structure stochastique. Pour des raisons de simplicité, nous supposerons que les deux équations décrivent fidèlement le comportement de toutes les entreprises, moyennant des erreurs aléatoires indépendantes (d'une entreprise à l'autre) et identiquement distribuées. Outre la forte possibilité d'une corrélation entre les termes d'erreur des deux équations pour une entreprise en particulier, notons qu'il y a un coefficient ($b_1$) commun aux deux équations.

Lorsque nous sommes en présence d'un système d'équations linéaires dont les coefficients sont assujettis à des conditions, il est peu utile de recourir aux méthodes de régression linéaire fondées sur un plan dont fait état Kott (1990a). C'est pourquoi les méthodes exposées dans cet article seront envisagées uniquement dans l'optique de l'approche fondée sur un modèle de Kott (1991), même si bon nombre d'entre elles s'inspirent d'approches fondées sur un plan.

Dans la section 2, nous exposons le modèle théorique pour l'estimation d'un système d'équations linéaires à l'aide de données d'une population. Dans la section suivante, nous présentons l'équivalent pondéré des estimateurs par les MCO et les MCG d'une population pour un système d'équations linéaires. La section 4 traite de l'estimation robuste de l'erreur quadratique moyenne de ces versions pondérées par une simple généralisation de l'estimateur de variance "de linéarisation" (voir, par exemple, Shah, Holt et Folsom 1977). Dans la section 5, nous examinons une méthode générale permettant de construire des critères statistiques qui peuvent servir à déterminer, entre autres choses, si les versions pondérées des estimateurs par les MCO et les MCG estiment toutes deux le même paramètre. Nous présentons un exemple simple dans la section 6 tandis que dans la section suivante, nous entreprenons d'étendre la méthodologie exposée précédemment à ce que les économètres appellent les "systèmes simultanés". En ce qui a trait à la version stochastique de l'équation 1 par exemple, de nombreux économistes croient que l'on devrait considérer $\log(y)$ comme une variable aléatoire et que l'on peut supposer $\log(c)$ fixe. Ces conditions ont pour effet de créer un biais de simultanéité si l'on ne recourt pas à des méthodes comme les doubles ou les triples moindres carrés (voir Johnston 1972, p. 341-420). Enfin, la section 8 renferme une brève analyse.

## 2. ESTIMATION POUR UNE POPULATION

### 2.1 Système non assujetti à des conditions

Supposons que nous avons une population avec $M$ données simples. Chacune de ces données ($i$) se rapporte à $G + \bar{K}$ variables observées qui satisfont le modèle suivant:

$$Y = \bar{X}\tilde{\beta} + U + V, \tag{3}$$

où $Y$ est la matrice $M \times G$ des variables dépendantes observées (la $i$-ième ligne de $Y$ renferme les variables dépendantes qui se rapportent à la $i$-ième donnée simple),

$\tilde{X}$ est la matrice $M \times \tilde{K}$ des variables indépendantes (ou explicatives) observées (la $i$-ième ligne de $\tilde{X}$ renferme les variables indépendantes qui se rapportent à la $i$-ième donnée simple),

$\tilde{\beta}$ est la matrice $\tilde{K} \times G$ des paramètres,

$U$ est une matrice $M \times G$ qui satisfait l'équation $\lim_{M\to\infty} \tilde{X}'U/M = 0_{\tilde{K}\times G}$ (un $G$-vecteur de valeurs nulles) – cela suppose l'existence d'un processus générateur de données simples qui pourrait, en principe, générer de telles données *à l'infini* (voir Kott 1991),

$V$ est une matrice $M \times G$ de variables aléatoires telle que $E(V) = 0_{M\times G}$ (une matrice de valeurs nulles) et $E(v_{is}v_{it}) = \sigma_{st(i)}$.

Nous savons tous que si $U \equiv 0_{M\times G}$, $E(v_{is}v_{jt}) = 0$ pour $i \neq j$, et $\sigma_{st(i)} = \sigma_{st}$ pour tous $i$, alors

$$\tilde{B}_{\text{MCO}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y \tag{4}$$

est le meilleur estimateur linéaire sans biais de $\tilde{\beta}$ (voir, par exemple, Johnston 1972, p. 240). Cela signifie que la $g$-ième colonne de $\tilde{B}_{\text{MCO}}$ (appelons-la $B_{.g}$) est le meilleur estimateur linéaire sans biais de $\beta_{.g}$, où

$$y_{.g} = \tilde{X}\beta_{.g} + u_{.g} + v_{.g}, \tag{5}$$

et $y_{.g}$, $u_{.g}$ et $v_{.g}$ sont les $g$-ième colonne de $Y$, $U$ et $V$ respectivement. L'équation (5) peut être considérée comme la $g$-ième équation du système représenté par l'équation (3).

Désignons la matrice $U$ de l'équation (3) comme la matrice des *variables explicatives présumées manquantes*. Dans les analyses de régression classiques (c.-à-d. fondées sur un modèle), on suppose généralement que la portion des variables dépendantes qui ne peut être décrite par une combinaison linéaire des variables indépendantes est purement aléatoire. Toutefois, dans la présente analyse, nous reprenons le raisonnement de Kott (1991) et reconnaissons la possibilité de variables explicatives manquantes non aléatoires. Notons que même lorsque $U \neq 0_{M\times G}$, $\tilde{B}_{\text{MCO}}$ est quasi (c.-à-d. asymptotiquement) sans biais.

### 2.2 Système (vraisemblablement) assujetti à des conditions

Il est plus difficile de réaliser une estimation efficace lorsque des contraintes s'appliquent à certains éléments de $\tilde{\beta}$, par exemple lorsqu'on sait que $\tilde{\beta}_{kg}$ est nul ou que $\tilde{\beta}_{kf}$ égale $\tilde{\beta}_{hj}$.

Dans cet article, nous nous intéressons à un système d'équations (vraisemblablement) assujetti à des conditions qui peut être représenté par l'équation suivante:

$$y = X\beta + u + v, \tag{6}$$

où $y = (y_{.1}', y_{.2}', \ldots, y_{.G}')'$, $u$ et $v$ sont définis de la même manière, $X$ est une matrice $MG \times K$, $\beta$ est un vecteur $K \times 1$ et $K \leq G\tilde{K}$. Par définition, $\lim_{M\to\infty} X'u/M = 0_K$.

Lorsque la matrice $\tilde{\beta}$ de l'équation (3) n'est pas assujettie à des contraintes, $K = G\tilde{K}$ et

$$X = \begin{bmatrix} \tilde{X} & & & \\ & \tilde{X} & & \\ & & \ddots & \\ & & & \tilde{X} \end{bmatrix}.$$

Toutefois, lorsque c'est l'inverse, $K < G\tilde{K}$. Par exemple, si on sait qu'un des éléments de $\tilde{\beta}$ est nul, on peut supprimer cet élément du vecteur $\beta$ de l'équation (6) en même temps que la colonne correspondante de la matrice $X$. Si deux éléments d'une même ligne de $\tilde{\beta}$ sont égaux, on peut extraire le second du vecteur $\beta$ de l'équation (6) et corriger $X$ en conséquence (il ne s'agira plus alors d'une matrice diagonale par blocs).

Lorsque $u \equiv 0_{MG}$ et $\text{Var}(v) = \Sigma \otimes I_M$ (où $\Sigma = \{\sigma_{st}\}$), alors $b_{\text{MCO}} = (X'X)^{-1}X'y$ est un estimateur sans biais de $\beta$ mais $b_{\text{MCG}} = (X'[\Sigma^{-1} \otimes I_M]X)^{-1}X'[\Sigma^{-1} \otimes I_M]y$, où $I_M$ est la matrice unité $M \times M$, est le meilleur estimateur linéaire sans biais. Dans la pratique, les éléments de $\Sigma$ doivent être estimés à partir de l'échantillon, par exemple au moyen de la formule $\hat{o}_{gf} = r_{.g}'r_{.f}/M$, où $r_{.g} = y_{.g} - X_{.g}b_{\text{MCO}}$.

Il est notoire que $b_{\text{MCO}}$ et $b_{\text{MCG}}$ sont égaux lorsque la matrice des paramètres en (3) n'est assujettie à aucune condition (voir de nouveau Johnston 1972, p. 240). Dans le cas contraire, si $u \neq 0_{\text{MG}}$, $b_{\text{MCO}}$ et $b_{\text{MCG}}$ sont tous deux des estimateurs quasi-non biaisés de $\beta$ lorsque l'équation $\lim_{M-\infty}\tilde{X}'U/M = 0_{\tilde{K}\times G}$ se vérifie, comme nous l'avons supposé à l'origine. Seulement, $b_{\text{MCG}}$ peut ne pas être quasi-non biaisé suivant l'hypothèse moins rigoureuse que $\lim_{M-\infty}X'u/M = 0_K$, ce qui est plus en rapport avec le modèle élargi de Kott (1991) lorsque (6) représente une équation unique et non un système d'équations.

Pour comprendre le phénomène, définissons $X_{.g}$ comme la matrice $M \times K$ formée des lignes $\{(g-1)M+1\}$ à $\{gM\}$ de $X$ et $\Sigma^{-1} = \{\sigma^{fg}\}$; alors,

$$E(b_{\text{MCG}} - \beta) \propto X'[\Sigma^{-1} \otimes I_M]u/M =$$

$$\sum_g X_{.g}'\left(\sum_f \sigma^{fg}u_{.f}\right)\Big/M = \sum_g \sum_f \sigma^{fg}X_{.g}'u_{.f}\Big/M,$$

expression dont la valeur tend vers zéro lorsque $M$ augmente, suivant l'hypothèse plus robuste mais pas nécessairement suivant l'hypothèse moins robuste.

## 3. ESTIMATION À L'AIDE DE DONNÉES D'ENQUÊTE

Supposons maintenant que nous n'observons des valeurs de variables que pour un échantillon aléatoire de la population. Soit $P = \text{diag}\{p_i\}$, où $p_i$ est la probabilité de sélection de la donnée simple $i$. Soit $S = \text{diag}\{s_i\}$, où $s_i = 1$ si la donnée simple $i$ est incluse dans l'échantillon et 0 dans le cas contraire. Enfin, posons $W = (m/M)P^{-1}S = \text{diag}\{w_i\}$ comme la matrice des poids d'échantillonnage, où $m$ représente la taille de l'échantillon. Notons que lorsque $p_i = m/M$ pour tous les $i$, $W = S$.

Il est facile de montrer que pour de nombreuses populations et de nombreux plans de sondage (voir Kott 1990b, 1991), l'*estimateur par les MCO pondéré*

$$\hat{\beta}_{W\cdot\text{MCO}} = (X'[I_G \otimes W]X)^{-1}X'[I_G \otimes W]y \tag{7}$$

est un estimateur convergent selon le plan de $b_{\text{MCO}}$, ce qui implique que $\text{plim}_{m-\infty}(\hat{\beta}_{W\cdot\text{MCO}} - b_{\text{MCO}}) = 0_K$. Dans les mêmes conditions, l'*estimateur par les MCG pondéré*

$$\begin{aligned}\hat{\beta}_{W\cdot\text{MCG}} &= (X'[I_G \otimes W][\hat{\Sigma}^{-1} \otimes I_M]X)^{-1}X'[I_G \otimes W][\hat{\Sigma}^{-1} \otimes I_M]y \\ &= (X'[\hat{\Sigma}^{-1} \otimes W]X)^{-1}X'[\hat{\Sigma}^{-1} \otimes W]y,\end{aligned} \tag{8}$$

où

$$\hat{o}_{gf} = r_{.g}'Wr_{.f}\Big/\sum_{i=1}^{M} w_i, \quad \text{et} \quad r = y - X\hat{\beta}_{W\cdot\text{MCO}},$$

est un estimateur convergent selon le plan de $b_{MCG}$. Comme $b_{MCO}$ et $b_{MCG}$ (et pour les mêmes raisons), $\hat{\beta}_{W\cdot MCO}$ et $\hat{\beta}_{W\cdot MCG}$ sont égaux lorsque le système d'équations n'est assujetti à aucune condition.

Si $\hat{\beta}_{W\cdot MCO}$ et $\hat{\beta}_{W\cdot MCG}$ sont convergents selon le plan, ils sont aussi des estimateurs quasi-non biaisés de $\beta$ lorsque $\lim_{M\to\infty} \bar{X}'U/M = 0_{R\times G}$ car $b_{MCO}$ et $b_{MCG}$ le sont; toutefois, $\hat{\beta}_{W\cdot MCG}$ - comme $b_{MCG}$ - peut ne pas être quasi-non biaisé suivant l'hypothèse moins rigoureuse que $\lim_{M\to\infty} X'u/M = 0_K$. (Dans cet article, la propriété d'être sans biais est toujours définie par rapport au modèle de l'équation (6)).

## 4. ESTIMATION DE L'ERREUR QUADRATIQUE MOYENNE

Supposons que nous avons un plan d'échantillonnage qui prévoit $H$ strates, $n_h$ u.p.é. tirées de la strate $h$, et $m_{hj}$ données simples tirées de l'u.p.é. $hj$. Les estimateurs $\hat{\beta}_{W\cdot MCO}$ et $\hat{\beta}_{W\cdot MCG}$ sont tous deux de la forme $\hat{\beta} = Cy$. Sans perte de généralité, nous pouvons réexprimer ces estimateurs sous la forme $\hat{\beta} = C^*y^*$, où $y^* = (y_{11}', \ldots, y_{Hn_H}')$ renferme uniquement les éléments qui correspondent aux données simples échantillonnées, et $y_{hj}$ est le vecteur des valeurs-$y$ $G \times m_{hj}$ relatives aux données simples de l'u.p.é. $hj$. Définissons $r^*$ et $r_{hj}$ de la même manière que $y^*$ et $y_{hj}$.

Soit $D_{hj}$ une matrice diagonale de zéros et de uns telle que $D_{hj}y^* = (0', \ldots, y_{hj}', \ldots, 0')$ et posons $g_{hj} = C^*D_{hj}r^*$. Par une extension directe de l'estimateur "de linéarisation", fondé sur un plan, nous pouvons exprimer l'estimateur de l'erreur quadratique moyenne de $\hat{\beta} = C^*y^*$ sous la forme:

$$\text{eqm} = \sum_{h=1}^{H} \frac{n_h}{n_h - 1}\left[\sum_{j=1}^{n_h} g_{hj}g_{hj}' - \frac{1}{n_h}\left(\sum_{j=1}^{n_h} g_{hj}\right)\left(\sum_{j=1}^{n_h} g_{hj}'\right)\right]. \tag{9}$$

Suivant des contraintes modérées pour le plan d'échantillonnage, eqm est quasi-non biaisé lorsque $U$ (d'après l'équation (3)) $\equiv 0_{M\times G}$ et $V$ répond à la propriété suivante:

$$|E(v_{sg}v_{tf})| \begin{cases} = 0 & \text{lorsque } s \text{ et } t \text{ ne viennent pas de la même u.p.é.} \\ < Q & \text{dans le cas contraire.} \end{cases}$$

Voir Kott (1991) pour la démonstration dans le cas où $G = 1$; la démonstration pour $G > 1$ est triviale. L'erreur quadratique moyenne de l'estimateur demeure à un niveau raisonnable lorsque $U \not\equiv 0_{M\times G}$ (voir Kott 1990a).

## 5. CRITÈRES STATISTIQUES

Soit $\hat{\beta}_{I\cdot MCO}$ et $\hat{\beta}_{I\cdot MCG}$ les équivalents non pondérés de $\hat{\beta}_{W\cdot MCO}$ et $\hat{\beta}_{W\cdot MCG}$, que l'on obtient en remplaçant $W$ par $S$ dans les équations (7) et (8). On cherche souvent à savoir si l'utilisation de poids d'échantillonnage a un effet quelconque. Cela nous amène à vérifier s'il existe une différence significative entre $\hat{\beta}_{I\cdot MCO}$ et $\hat{\beta}_{W\cdot MCO}$, autrement dit si les deux estimateurs servent à estimer la même chose.

Une fois que l'importance de l'utilisation de poids *a été établie*, on veut aussi savoir s'il existe une différence significative entre $\hat{\beta}_{W\cdot MCO}$ et $\hat{\beta}_{W\cdot MCG}$, autrement dit si la relation $\lim_{M\to\infty} \bar{X}'U/M = 0_{R\times G}$ se vérifie de telle sorte que les deux estimateurs en question servent à estimer la même chose.

Un critère général permettant de vérifier si

$$\hat{\beta}_{(1)} = \sum_h \sum_j \{C_{(1)}{}^* D_{hj} y^*\} \quad \text{et} \quad \hat{\beta}_{(2)} = \sum_h \sum_j \{C_{(2)}{}^* D_{hj} y^*\}$$

sont égaux est

$$T^2 = [\hat{\beta}_{(1)} - \hat{\beta}_{(2)}]' A^{-1} [\hat{\beta}_{(1)} - \hat{\beta}_{(2)}], \tag{10}$$

où

$$A = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left[ \sum_{j=1}^{n_h} d_{hj} d_{hj}' - \frac{1}{n_h} \left( \sum_{j=1}^{n_h} d_{hj} \right) \left( \sum_{j=1}^{n_h} d_{hj}' \right) \right],$$

$d_{hj} = C_{(1)}{}^* D_{hj} r_{hj(1)} - C_{(2)}{}^* D_{hj} r_{hj(2)}$, et $r_{hj(f)} = y_{hj} - X_{hj} \hat{\beta}_{(f)}$.

Suivant l'hypothèse nulle, le critère $T^2$ est une variable aléatoire distribuée asymptotiquement selon une loi de chi carré avec $K$ degrés de liberté. Compte tenu de l'importance que nous accordons à la robustesse, il semble prudent de s'interroger sur la validité de l'hypothèse nulle lorsque prob$(\chi^2_{(K)} > T^2)$ est largement inférieure au niveau normal de 0.1 ou de 0.05 mais non lorsque $T^2$ est inférieur à son espérance, $K$.

## 6. EXEMPLE

Considérons l'exemple ci-dessous, construit à l'aide de données de l'enquête agricole de juin 1989 du National Agricultural Statistics Service. La série de données pertinente, analysée antérieurement dans Kott (1990a), est décrite brièvement ci-après.

On a tout d'abord prélevé 17 unités primaires d'échantillonnage à même 4 strates. Ces unités ont ensuite fait l'objet d'un sous-échantillonnage qui a permis d'obtenir un échantillon global de 252 exploitations agricoles. Bien qu'il s'agît d'un échantillon aléatoire, la probabilité de sélection n'était pas la même pour toutes les exploitations.

Supposons que nous voulons estimer les paramètres $\beta_1$ et $\beta_2$ de l'équation suivante:

$$y_{1i} = x_{1i}\beta_1 + x_{2i}\beta_2 + u_{1i} + v_{1i}, \tag{11}$$

où $i$ désigne une exploitation agricole;

$y_{1i}$ est le rapport de la superficie ensemencée au soja à la superficie de terre labourable pour l'exploitation $i$ si cette exploitation a des terres labourables; dans le cas contraire, ce rapport est nul;

$x_{1i}$ est égal à 1 si l'exploitation $i$ a des terres labourables et égal à 0 dans le cas contraire;

$x_{2i}$ est le quotient de la superficie de terre labourable de l'exploitation $i$ par 10,000.

(Nota: le fait d'éliminer de l'équation de régression toutes les exploitations échantillonnées qui n'ont pas de terre labourable n'influe aucunement sur l'estimation des paramètres mais peut avoir des effets sur l'estimation de l'erreur quadratique moyenne.)

Si nous posons $\hat{\beta}_{(1)}$ dans l'équation (10) comme l'estimateur par les MCO non pondéré pour le vecteur $(\beta_1, \beta_2)'$ et $\hat{\beta}_{(2)}$ comme l'estimateur pondéré, nous obtenons une valeur $T^2$ de 4.58. Suivant l'hypothèse nulle que l'estimateur par les MCO et son équivalent pondéré permettent d'arriver au même résultat (ce pour quoi $u_{1i} \equiv 0$ est une condition suffisante mais

**Tableau 1**

Estimateurs pour les paramètres de l'équation (11)

| | |
|---|---|
| MCO | $y_{1i} = \underset{(.044)}{0.268x_{1i}} - \underset{(3.95)}{0.92x_{2i}} + u_{1i} + v_{1i}$ |
| Pondéré | $y_{1i} = \underset{(.075)}{0.191x_{1i}} + \underset{(9.95)}{12.15x_{2i}} + u_{1i} + v_{1i}$ |
| MCG - pondéré | $y_{1i} = \underset{(0.71)}{0.197x_{1i}} + \underset{(6.97)}{10.26x_{2i}} + u_{1i} + v_{1i}$ |

Les chiffres entre parenthèses représentent la racine carrée de l'erreur quadratique moyenne.

non nécessaire), $T^2$ est distribué asymptotiquement selon la loi chi carré avec 2 degrés de liberté. Nous ne pouvons rejeter l'hypothèse nulle à un seuil de 0.1. Cependant, comme la valeur de $T^2$ est beaucoup plus élevée que 2, il est fortement permis de croire qu'il manque une variable explicative. Par conséquent, il conviendrait d'utiliser l'estimateur par régression pondéré plutôt que l'estimateur par les MCO.

Le tableau 1 donne les estimations des coefficients de l'équation (11) calculées à l'aide de l'estimateur par les MCO et de son équivalent pondéré. Bien que l'écart entre l'estimateur pondéré de $\beta_2$ et la valeur nulle ne soit pas statistiquement significatif à un seuil de 0.1, nous conservons cet estimateur dans le modèle car sa valeur est supérieure à la valeur estimée de la racine carrée de l'erreur quadratique moyenne correspondante. Cela rejoint notre argument en faveur de l'estimateur par régression pondéré.

Il faut toutefois noter la perte d'efficacité qui découle de l'utilisation d'un estimateur pondéré. En effet, la valeur estimée de la racine carrée de l'erreur quadratique moyenne de l'estimateur de $\beta_2$ fait plus que doubler lorsqu'on passe de l'estimateur par les MCO à son équivalent pondéré (nota: dans les deux cas, la racine carrée de l'erreur quadratique moyenne a été calculée au moyen de l'équation (9)).

Il est possible d'accroître l'efficacité de l'estimateur pondéré en introduisant une seconde équation pour l'exploitation agricole et en la jumelant à l'équation (11) pour obtenir un système à estimer. Soit

$$y_{2i} = x_{1i}\beta_3 + u_{2i} + v_{2i}, \tag{12}$$

où $y_{2i}$ est le rapport de la superficie ensemencée au maïs à la superficie de terre labourable pour l'exploitation $i$ si cette exploitation a des terres labourables; dans le cas contraire, ce rapport est nul.

L'existence d'un système d'équations ne change rien aux estimations calculées au moyen de l'estimateur par les MCO pondéré ni à la valeur estimée de la racine carrée de l'erreur quadratique moyenne correspondante (tableau 1). Cependant, nous pouvons aussi calculer désormais la valeur estimée de $\beta_1$ et $\beta_2$ au moyen de l'estimateur par les MCG pondéré; les valeurs pertinentes figurent aussi dans le tableau 1. Il convient de noter, à cet égard, la réduction d'environ 30% de la valeur estimée de la racine carrée de l'erreur quadratique moyenne pour $\beta_2$ sans qu'il y ait perte de robustesse, en supposant que les versions pondérées de l'estimateur par les MCO et de l'estimateur par les MCG servent à estimer la même chose.

La valeur $T^2$ pour un test permettant de comparer les versions pondérées de l'estimateur par les MCO et de l'estimateur par les MCG pour le vecteur $(\beta_1, \beta_2)'$ est 0.97. Comme cette valeur est sensiblement inférieure à 2, il est permis de croire que les deux estimateurs donnent le même résultat. Autrement dit, l'une des équations ne renferme aucune variable explicative

additionnelle qui ait un lien avec la variable explicative présumée manquante de l'autre équation (ce qui n'est pas étonnant étant donné que lorsqu'on a introduit un terme $x_{2i}$ dans le membre de droite de l'équation (12), le coefficient estimé correspondant était inférieur à la valeur estimée de la racine carrée de l'erreur quadratique moyenne).

## 7. SYSTÈME SIMULTANÉ

Dans un système simultané, on retrouve quelques-unes des colonnes de la matrice des variables dépendantes, $Y$, (voir équ. (3)) dans le membre de droite de la $g$-ième équation (voir (5)). Formellement, nous pouvons écrire

$$y = Y_{(\cdot)}\alpha + X\beta + u + v \quad \text{ou} \quad y = Z\delta + u + v,$$

où

$$Y_{(\cdot)} = \begin{bmatrix} Y_{(1)} \\ Y_{(2)} \\ \cdot \\ \cdot \\ Y_{(G)} \end{bmatrix},$$

$$Z = (Y_{(\cdot)}X), \quad \text{et} \quad \delta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

La plupart des colonnes de $Y_{(g)}$ sont des vecteurs nuls. Les autres (pas plus que G-1 colonnes) proviennent de la matrice $Y$ de l'équation (3).

Définissons $\hat{Y}_{(g)}$ comme $\tilde{X}(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'WY_{(g)}$. Remplaçons ensuite $X$ dans l'équation (5) par $\hat{Z} = (\hat{Y}_{(\cdot)}X)$ et refaisons les opérations précédentes. L'équation (7) donne $\hat{\delta}_{W\text{-MCO}}$, ce qui tient des doubles moindres carrés, tandis que l'équation (8) donne $\hat{\delta}_{W\text{-MCG}}$, qui tient des triples moindres carrés. En ce qui concerne l'estimation de l'erreur quadratique moyenne, on applique le raisonnement qui a permis d'obtenir l'équation (9).

## 8. ANALYSE

Par cet article, nous visions à montrer comment les méthodes décrites dans les ouvrages portant sur la théorie des sondages à base de plans - en particulier, la régression pondérée (pour échantillon) et l'estimateur de variance "de linéarisation" - pouvaient être appliquées à l'estimation d'un système d'équations linéaires.

Notre analyse nous a amenés à faire une constatation quelque peu surprenante: lorsqu'on estime les paramètres d'un système assujetti à des conditions, il se peut que les équivalents pondérés de l'estimateur par les MCO et de l'estimateur par les MCG ne servent pas à estimer la même chose. À bien y penser toutefois, cette constatation n'est pas si étonnante qu'on le croit. S'il manque des variables explicatives dans notre modèle de travail, peut-être nous ne connaissons pas toujours assez le vrai modèle pour pouvoir appliquer des conditions aux paramètres.

Il importe de préciser que l'équation (9) peut servir à estimer l'erreur quadratique moyenne d'estimateurs de paramètres même lorsqu'il ne manque aucune variable explicative. L'avantage de cette formule par rapport aux formules classiques est qu'elle est conçue en fonction de l'hétéroscédasticité et de corrélations complexes entre les données simples (mais à l'intérieur des u.p.é.).

Néanmoins, lorsqu'il ne manque aucune variable explicative, l'estimateur ci-dessous présente tous les avantages de l'estimateur de l'équation (9) et est généralement plus efficace:

$$\text{eqm}' = \sum_{h=1}^{H} \sum_{j=1}^{n} \frac{n}{n-1} \{g_{hj} g_{hj}' - gg'/n\}, \tag{13}$$

où $n = \sum n_h$ et $g = \sum \sum g_{hj}$ (nota: si $Xq = (1, \ldots, 1)'$ pour un vecteur quelconque $q$ de dimension $K$, alors $g = 0$).

Lorsqu'il *manque* des variables explicatives, les éléments diagonaux de eqm' risquent d'être affectés d'un biais par excès. Le raisonnement que nous venons d'exposer est conforme à celui appliqué par Wolter (1985) pour les estimateurs de la variance pour strates regroupées selon la théorie des sondages à base de plans.

## BIBLIOGRAPHIE

JOHNSTON, J. (1972). *Econometric Methods*, (deuxième édition). New York: McGraw Hill.

KOTT, P.S. (1991). A model-based look at linear regression with survey data. *American Statistician*, à paraître.

KOTT, P.S. (1990a). What does performing a linear regression on survey data mean? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, à paraître.

KOTT, P.S. (1990b). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, 24, 287-296.

SHAH, B.V., HOLT, M.M., et FOLSOM, R.E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, 47, 43-57.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57, 348-368.

Techniques d'enquête, juin 1991
Vol. 17, n° 1, pp. 111–131
Statistique Canada



# L'évaluation des erreurs dans les données de comptabilité nationale: les estimations provisoires et revisées

LUIGI BIGGERI et UGO TRIVELLATO[1]

RÉSUMÉ

Cet article présente et discute de façon critique les récents développements dans l'évaluation de la fiabilité des estimations provisoires de comptabilité nationale. On introduit d'abord un schéma théorique du processus de révision des estimations d'un agrégat, et on considère les implications qu'il a sur le dessin des analyses des erreurs dans les données provisoires. Une attention particulière est ensuite portée au choix des mesures élémentaires des erreurs et d'indices synthétiques de précision convenables, et à l'impact des révisions sur les agrégats à prix constants et les déflateurs implicites. Enfin, les résultats de quelques analyses empiriques sur les discordances entre des estimations provisoires et révisées, menées comparativement sur des données de comptabilité nationale du Canada, des États-Unis et de l'Italie, sont synthétiquement présentés.

MOTS CLÉS: Révision des données; comptes nationaux; indices de précision; agrégats à prix constants.

## 1. INTRODUCTION

L'évaluation de la fiabilité des estimations de comptabilité nationale est importante à cause de l'influence de ces estimations sur les analyses et les décisions de politique économique. Elle est en même temps difficile à effectuer, en raison des très nombreuses et complexes sources d'erreur dont elles sont affectées.

Il est donc problématique d'arriver à un unique critère de fiabilité, à la fois convaincant au niveau conceptuel et praticable. Il apparaît en revanche raisonnable et possible d'établir de nombreux critères partiels, convenables pour mesurer les aspects principaux de la fiabilité. Pour une revue des différents critères nous renvoyons le lecteur à Novak (1975) et Trivellato (1987).

L'objectif de cet article est de présenter et de discuter de façon critique certains développements récents dans l'analyse des erreurs dans les estimations provisoires de comptabilité nationale.

Pour les agrégats de comptabilité nationale, ainsi que pour beaucoup d'autres agrégats et indicateurs économiques, des estimations provisoires sont initialement publiées, et sont ensuite soumises à différentes révisions.

Le processus de révision est déterminé avant tout par l'exigence d'assurer des informations rapides, et, en même temps, par le temps requis pour le recueil et l'élaboration de l'ensemble des données utilisées couramment pour l'estimation des agrégats ("actualité contre précision", selon l'efficace polarisation de Wilton et Smith 1974). Typiquement ceci donne lieu à un processus de révision caractérisé par une estimation préliminaire et par une suite de révisions de routine à périodicité rapprochée. De temps à autre il peut ensuite arriver que des modifications soient apportées aux classifications et schémas de comptabilité et/ou dans les enquêtes statistiques de base et dans les méthodes d'estimation, pour améliorer la précision ou la pertinence des données. Ceci ouvre la voie à de nouvelles évaluations, et nous parlons dans ce cas de révisions occasionnelles ou extraordinaires.

La particularité des erreurs dans les estimations provisoires réside dans le fait qu'elles sont présentes dans ces estimations et, par définition, éliminées dans les estimations révisées. Elles peuvent donc être mesurées en comparant les deux estimations. Certes, à strictement parler,

[1] Luigi Biggeri, Dipartimento Statistico, Università di Firenze, Italia; Ugo Trivellato, Dipartimento di Scienze Statistiche, Università di Padova, Italia.

la différence entre une estimation provisoire et l'estimation correspondante révisée est une pure mesure de la différence dans l'erreur entre les deux estimations successives. Toutefois, quand le processus de révision se base sur des informations de plus en plus complètes et sur des méthodes d'évaluation de plus en plus affinées, comme il arrive justement dans les pays développés, il est raisonnable de considérer l'estimation finale comme la plus proche de la vraie valeur inconnue, et dans des buts de comparaison, de la traiter comme valeur de référence.

La connaissance des caractéristiques et du comportement des discordances entre estimations provisoires et révisées est d'une utilité certaine. De fortes et fréquentes différences constituent, en effet, un symptôme préoccupant quant à la qualité des données de base et/ou des méthodes d'estimation employées dans les évaluations préliminaires et en suggèrent, tout au moins, un usage prudent. (Ceci ne signifie naturellement pas que de petites discordances soient nécessairement une garantie de qualité des données. Le fait qu'une évaluation préliminaire ne soit pas révisée peut simplement vouloir dire que nous n'avons pas d'éléments suffisants pour l'améliorer, abstraction faite de sa fiabilité.) Si en plus les différences présentent des traits de systématicité, le fait peut représenter, en même temps, un utile avertissement tant pour l'utilisateur, qui éventuellement adoptera des correctifs, que pour le producteur, qui pourra en déduire des suggestions pour améliorer les estimations provisoires.

Un compte-rendu des nombreuses analyses statistiques des discordances entre estimations provisoires et révisées qui se sont succédées, est donné par Trivellato (1986a). D'autres récentes contributions sont de Zarnowitz (1982), Mankiw, Runkle et Shapiro (1984), Mankiw et Shapiro (1986), McNees (1986), Mork (1987) et Lefrançois (1988).

Sur ce thème la littérature est donc considérable. Mais certaines questions, tant méthodologiques que substantielles, concernant la manière de mener les analyses, ont été partiellement négligées ou résolues de façon quelquefois peu satisfaisante. En particulier, ceci vaut pour:

(a) l'identification d'un dessin des analyses de la précision des estimations provisoires, qui soit cohérent avec les caractéristiques du processus de révision;

(b) la définition des mesures élémentaires appropriées des discordances entre estimations provisoires et révisées; le choix d'indices synthétiques convenables, qui soient capables de nous informer de la précision globale d'une série d'estimations provisoires, et, si le processus de révision se déroule en plusieurs étapes, qui puissent être décomposés de façon à pouvoir vérifier la convergence des estimations provisoires vers la valeur finale;

(c) l'examen des implications des erreurs dans les estimations provisoires sur les séries dérivées, notamment sur les agrégats à prix constants et les déflateurs implicites.

Les trois sections suivantes sont consacrées, dans l'ordre, à ces problématiques. La section 5, enfin, rapporte de façon sommaire les résultats de quelques analyses empiriques sur les discordances entre des estimations provisoires et révisées, menées comparativement sur des données de comptabilité nationale du Canada, des États-Unis et de l'Italie.

## 2. UN SCHÉMA THÉORIQUE DU PROCESSUS DE RÉVISION

Un schéma suffisamment général du processus de révision des agrégats économiques est représenté dans le tableau 1, qui permet une visualisation immédiate du lien entre la période de référence et la période de publication des différentes estimations.

De l'examen du schéma émergent clairement trois types d'estimations et de révisions:

(a) Celles qui sont placées sur la diagonale principale et sur les diagonales inférieures, qui décrivent un processus de révision courant se déroulant en $m$ pas, allant de l'estimation préliminaire ${}_1p_t$ à l'estimation définitive $r_t$ ("définitive" par rapport au processus de révision courant).

(b) Celles placées horizontalement $({}_br_t)$, qui incorporent les ajustements aux jalons, où la reconstruction de la série se fait en référence à une période qui va d'un étalonnage au suivant.

**Tableau 1**

Schéma de publication des estimations successives d'un aggrégat[a]

| Période de publication | Période de référence | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $t-2h+1$ | . | . | $t-h$ | $t-h+1$ | . | $t-m$ | . | . | $t-1$ | $t$ | $t+1$ | . | . | $T-m$ | . | . | $T-1$ | $T$ | $T+1$ |
| $t-2h+2$ | ${}_1p_{t-2h+1}$ | | | | | | | | | | | | | | | | | | | |
| . | ${}_2p_{t-2h+1}$ | | | | | | | | | | | | | | | | | | | |
| . | . | . | . | | | | | | | | | | | | | | | | | |
| . | . | . | . | ${}_1p_{t-h}$ | | | | | | | | | | | | | | | | |
| $t-2h+m+2$ | ${}_mp_{t-2h+1}$ | . | . | ${}_2p_{t-h}$ | ${}_1p_{t-h+1}$ | | | | | | | | | | | | | | | |
| $t-2h+m+3$ | $r_{t-2h+1}$ | . | . | . | ${}_2p_{t-h+1}$ | . | | | | | | | | | | | | | | |
| . | | . | . | . | . | . | ${}_1p_{t-m}$ | | | | | | | | | | | | | |
| . | | | . | ${}_mp_{t-h}$ | . | . | ${}_2p_{t-m}$ | ${}_1p_.$ | | | | | | | | | | | | |
| . | | | | $r_{t-h}$ | ${}_mp_{t-h+1}$ | . | . | ${}_2p_.$ | ${}_1p_.$ | | | | | | | | | | | |
| . | | | | | $r_{t-h+1}$ | . | . | . | ${}_2p_.$ | ${}_1p_{t-1}$ | | | | | | | | | | |
| . | ${}_br_{t-2h+1}$ | . | . | ${}_br_{t-h}$ | | . | ${}_mp_{t-m}$ | . | . | ${}_2p_{t-1}$ | ${}_1p_t$ | | | | | | | | | |
| $t+2$ | | | | | | | $r_{t-m}$ | ${}_mp_.$ | . | . | ${}_2p_t$ | ${}_1p_{t+1}$ | | | | | | | | |
| . | | | | | | | | $r_.$ | ${}_mp_.$ | . | . | ${}_2p_{t+1}$ | . | | | | | | | |
| . | | | | | | | | | $r_.$ | ${}_mp_{t-1}$ | . | . | . | . | | | | | | |
| . | | | | | | | | | | $r_{t-1}$ | ${}_mp_t$ | . | . | . | ${}_1p_{T-m}$ | | | | | |
| $t+m+2$ | | | | | | | | | | | $r_t$ | ${}_mp_{t+1}$ | . | . | ${}_2p_{T-m}$ | ${}_1p_.$ | | | | |
| $t+m+3$ | | | | | | | | | | | | $r_{t+1}$ | . | . | . | ${}_2p_.$ | ${}_1p_.$ | | | |
| . | | | | | ${}_br_{t-h+1}$ | | ${}_br_{t-m}$ | . | . | ${}_br_{t-1}$ | ${}_br_t$ | | . | . | . | . | ${}_2p_.$ | ${}_1p_{T-1}$ | | |
| . | | | | | | | | | | | | | . | . | ${}_mp_{T-m}$ | . | . | ${}_2p_{T-1}$ | ${}_1p_T$ | |
| $T+2$ | | | | | | | | | | | | | | | $r_{T-m}$ | ${}_mp_.$ | . | . | ${}_2p_T$ | ${}_1p_{T+1}$ |
| $T+3$ | ${}_sr_{t-2h+1}$ | . | . | ${}_sr_{t-h}$ | ${}_sr_{t-h+1}$ | . | ${}_sr_{t-m}$ | . | . | ${}_sr_{t-1}$ | ${}_sr_t$ | ${}_sr_{t+1}$ | . | . | ${}_sr_{T-m}$ | $r^*$ | ${}_mp^*$ | . | . | ${}_2p^*_{T+1}$ |

[a] Le processus de révision prévoit $m$ révisions successives de type courant $({}_2p, \ldots, {}_mp, r)$, une révision aux jalons due aux contrôles effectués toutes les $h$ périodes $({}_br)$ et une révision extraordinaire à l'occasion de laquelle on reconstruit une série rétrospective à partir du temps $T+2$ (Source: Biggeri 1984, p.24).

(c) Celles qui sont la conséquence d'une révision occasionnelle ($_s r_t$), elles aussi placées horizontalement, qui concernent la reconstruction rétrospective de la série pour une période généralement assez longue.

Ce schéma met aussi en évidence que la présence de révisions aux jalons et occasionnelles, qui se superposent au processus de révision courant, résulte en des révisions mixtes. La révision que l'on effectue ainsi n'est homogène ni avec les précédentes, ni avec les suivantes dans la séquence des estimations courantes.

En disposant d'une série chronologique d'estimations provisoires d'un agrégat pour les temps de 1 à $n$ et de celle des estimations révisées correspondantes, le problème de l'évaluation de la précision des premières par rapport aux deuxièmes est reconduisible, sur le plan formel, à celui de l'évaluation de la validité des prévisions, sur lequel existe une ample littérature. Toutefois le mécanisme de révision des agrégats de comptabilité nationale, schématisé dans le tableau 1, présente certaines particularités qu'il faut avoir présent à l'esprit.

Tout d'abord, il est opportun que les analyses du processus de révision tiennent compte explicitement de l'existence des trois types de révision mentionnés plus haut. De plus, pour saisir correctement les caractéristiques des révisions en question, toutes les comparaisons qui entraînent des révisions mixtes doivent être exclues de l'analyse.

Une deuxième considération particulièrement importante concerne les méthodes et critères pour l'analyse de la précision des estimations provisoires. La plupart des travaux dans la littérature statistique et économique portant sur les révisions des comptes nationaux, visent à établir la précision des estimations provisoires par le moyen de mesures statistiques, généralement descriptives. Bien que différentes approches, fort intéressantes, furent récemment proposées (par exemple, par Mankiw et Shapiro 1986, et par Lefrançois 1988), elles ne sont peut-être pas entièrement satisfaisantes pour une analyse des propriétés des données provisoires. En effet, elles constituent des intégrations convenables mais non des alternatives à l'analyse de base des propriétés des données provisoires, pour laquelle il convient d'adopter des critères et des mesures de précision essentiellement descriptifs.

Enfin, en effectuant l'analyse de la précision des estimations provisoires, il semble approprié de traiter de façon non symétrique les estimations provisoires et les définitives et de considérer la série révisée $r_t$ comme série de référence. Ce choix est motivé précisément du fait que nous nous proposons d'évaluer la précision des estimations provisoires, étant données les estimations définitives (ou de toutes façons assimilables à celles-ci dans la comparaison en question).

## 3. MESURES ÉLÉMENTAIRES ET INDICATEURS SYNTHÉTIQUES DES ERREURS DANS LES ESTIMATIONS PROVISOIRES

### 3.1 Sur le choix entre erreurs et erreurs relatives

L'erreur dans l'estimation provisoire ($p_t$) du niveau d'un agrégat économique est donnée par

$$v_t = p_t - r_t. \tag{1}$$

Celle-ci peut être inappropriée pour comparer la précision de plusieurs agrégats, parce que les résultats dépendent de l'unité de mesure et de l'ordre de grandeur de l'agrégat considéré. Il peut alors être préférable d'utiliser l'erreur relative, définie comme

$$e_t = (p_t - r_t)/r_t = p_t/r_t - 1. \tag{2}$$

Le choix entre l'analyse des erreurs ou des erreurs relatives est sous de nombreux aspects crucial, et mérite d'être approfondi. En effet, comme on le verra dans la section 3.2, les indices

synthétiques pertinents de précision des estimations provisoires sont des moyennes simples respectivement des $v_t$ et des $e_t$ (ou de leurs transformations convenables). Or, l'emploi de moyennes simples est raisonnable si les séries proviennent d'un processus purement aléatoire, et en particulier si elles ne contiennent pas une composante de tendance. Dans le cas contraire, de l'information est perdue et l'analyse peut être peu éclairante et, à la limite, fourvoyante. Pour cette raison des vérifications préalables s'imposent, qui peuvent être menées avec plusieurs tests (voir, par exemple, Malinvaud 1969, p. 473-481; Kendall 1973, p. 22-28; Box et Jenkins 1970, p. 34-36 et 287-298).

Par rapport au problème spécifique du choix entre erreurs et erreurs relatives, un critère particulièrement utile est offert par l'analyse des paramètres d'un modèle convenable entre les données provisoires et les données définitives. Dans sa formulation la plus simple, celui-ci peut être spécifié comme:

$$p_t = \alpha + \beta r_t + \epsilon_t, \tag{3}$$

où $\epsilon_t$ est l'erreur aléatoire. Du modèle (3) il est immédiat de déduire:

$$v_t = \alpha + (\beta - 1) r_t + \epsilon_t, \tag{4}$$

$$e_t = (\beta - 1) + \alpha \frac{1}{r_t} + \frac{\epsilon_t}{r_t}. \tag{5}$$

Les relations (4) et (5) mettent en évidence que, en général, aussi bien l'erreur que l'erreur relative de l'estimation provisoire dépendent du niveau de l'estimation définitive correspondante. Nous pouvons aussi en déduire les conditions qui doivent être satisfaites pour que l'emploi d'indices synthétiques basés respectivement sur les erreurs ou bien sur les erreurs relatives soit justifié:

(a) $v_t$ ne dépend pas de $r_t$ si $\beta$ est égal à un et les $\epsilon_t$ sont homoscédastiques (et non-corrélés temporellement);

(b) $e_t$ ne dépend pas de $r_t$ si $\alpha$ est égal à zéro, et si la variance des $\epsilon_t$ est proportionnelle au carré du niveau de la série (les $\epsilon_t$ étant non-corrélés temporellement).

On observe immédiatement qu'estimer l'équation (4) avec les moindres carrés ordinaires équivaut à estimer l'équation (5) avec les moindres carrés généralisés (avec $E(\epsilon\epsilon') = \sigma_\epsilon^2 \Omega$, où $\Omega$ est une matrice diagonale avec $w_{tt} = r_t^2$). Il est aussi immédiat de noter que les équations (3) et (4) ne diffèrent que de 1 dans le coefficient angulaire. La vérification de l'homoscédasticité des $\epsilon_t$ dans l'équation (3) devient donc cruciale pour le choix entre l'analyse des erreurs ou des erreurs relatives. Dans ce but, Trivellato, Di Fonzo et Rettore (1986) ont élaboré un simple test non-paramétrique, basé sur l'ordre des résidus estimés, qui n'engage pas à spécifier une structure stochastique particulière pour l'équation (3), ce qui est en accord avec le peu de connaissance que nous avons a priori sur la relation entre les estimations provisoires et définitives.

Ce test peut aussi servir à l'examen de l'hypothèse de stabilité des paramètres de l'équation (3). Il est en effet évident que les conditions (a) et (b) considérées plus haut ne sont pas exhaustives, et qu'une raison plausible pour qu'elles ne soient pas satisfaites est la présence d'instabilités dans le processus de révision. L'intérêt de vérifications de ce type est, entre autres, d'étudier la superposition d'une révision occasionnelle au processus de révision courant. Si le résultat du test favorise l'hypothèse de stabilité, il est approprié d'analyser le processus de révision courant dans son entier. Dans le cas contraire, l'analyse doit être menée séparément dans les deux périodes précédente et successive à la révision occasionnelle.

Il faut noter, enfin, que la spécification (3) est intentionellement stylisée, et se prête essentiellement à l'exploration du processus de révision d'une série annuelle quand les résidus $\epsilon_t$ ne présentent pas d'autocorrélation. Il est cependant facile de la généraliser de diverses façons: (i) en introduisant un vecteur de variables explicatives non aléatoires, pour tenir compte d'éventuels facteurs influant le processus de révision (révisions aux jalons, facteurs saisonniers, *etc.*); (ii) en admettant une autocorrélation dans les résidus; (iii) en modélisant conjointement le processus de révision de plusieurs séries, au moyen d'un système de régressions apparemment indépendantes. Pour ces développements, voir Trivellato et Rettore (1986) et Bordignon et Trivellato (1989).

### 3.2 Indices synthétiques de la précision des estimations provisoires et "cohérence faible"

Pour caractériser les indices synthétiques appropriés de la précision globale d'un vecteur d'estimations provisoires, il convient de se référer à la propriété de "cohérence faible" définie dans Trivellato (1986b). En substance, par référence à deux séries d'estimations provisoires correspondant à la même série d'estimations définitives, celle-ci requiert que lorsque la première série montre des erreurs en valeur absolue inférieures ou égales à celles de la seconde, avec inégalité stricte dans au moins un cas, l'indice de précision de la première série doit être plus petit que l'indice de la seconde série, signalant ainsi que la première série d'estimations provisoires est plus proche des données définitives que la seconde.

Comme on le verra, la référence centrale aux indices synthétiques faiblement cohérents n'innove pas dans les mesures habituellement utilisées (nous utiliserons essentiellement l'erreur absolue moyenne et l'erreur quadratique moyenne). Ce concept, uni au choix d'analyser les erreurs ou bien les erreurs relatives et à l'éventuel repérage de sous-périodes homogènes du processus de révision, permet toutefois de mener de manière appropriée les analyses empiriques, tout en évitant les fréquentes imprécisions de la littérature.

Il est aussi utile de considérer les décompositions des indices synthétiques faiblement cohérents. Dans une perspective légèrement différente, l'existence et le poids de composantes spécifiques peuvent éventuellement être mis en lumière en examinant les paramètres estimés de l'équation (3) (Mincer et Zarnowitz 1969; Hatanaka 1974; Hempenius 1980; Trivellato 1986a).

Les indicateurs sur lesquels il convient de se pencher sont les suivants:

(I) **Erreurs**

Si le critère de choix entre les erreurs et les erreurs relatives (voir section 3.1) porte à mener l'analyse sur les erreurs, deux indices synthétiques faiblement cohérents sont l'erreur absolue moyenne et la racine carrée de l'erreur quadratique moyenne:

$$\bar{v}' = \sum | v_t | /n, \tag{6}$$

$$d_v = \sqrt{(\sum v_t^2/n)}, \tag{7}$$

où la sommation est étendue aux $n$ termes de la série.

La décomposition suivante, analogue à la décomposition asymétrique de Theil (1966) mais développée en traitant $r_t$ comme série de référence, permet de mettre en évidence les insuffisances dans la performance des estimations provisoires:

$$d_v^2 = (\bar{p} - \bar{r})^2 + (s_r - \hat{\rho} s_p)^2 + (1 - \hat{\rho}^2)\, s_p^2, \tag{8}$$

où $s_r$ et $s_p$ sont les écarts-type respectifs de $r_t$ et $p_t$, et $\hat{\rho}$ est le coefficient de corrélation entre $p_t$ et $r_t$. De la relation (8) on tire la décomposition relative suivante:

$$1 = U_v^M + U_v^R + U_v^D. \tag{9}$$

Les trois termes de l'équation (9) représentent des fractions de l'erreur quadratique moyenne interprétables comme suit: composante de biais entre les moyennes des deux séries, $U_v^M$; composante de régression, attribuable à l'écartement de 1 du coefficient de la régression de $p_t$ sur $r_t$, $U_v^R$; composante d'erreur aléatoire, attribuable à la variance des erreurs de régression, $U_v^D$.

En jugeant de la qualité des estimations, il est aussi utile de considérer la moyenne et l'écart quadratique moyen des erreurs, c'est-à-dire respectivement $\bar{v} = \sum v_t/n = \bar{p} - \bar{r}$ et $s_v = \sqrt{(\sum (v_t - \bar{v})^2/n)}$. Bien entendu, $\bar{v}$ n'est pas un indice de précision des estimations provisoires. Celui-ci nous informe seulement de la direction et de la dimension de l'erreur moyenne de niveau. Cette statistique est, cependant, remarquable pour au moins deux raisons: (i) si $\bar{v} \simeq 0$, il n'y a pas de composante de biais ($U_v^M = 0$); (ii) si $\bar{v} \neq 0$, il est alors instructif d'examiner conjointement $\bar{v}$ et $\bar{v}'$, puisque $|\bar{v}| \simeq \bar{v}'$ si, et seulement si, les erreurs sont presque toujours dans la même direction. La comparaison des deux indices peut donc mettre en évidence une éventuelle composante systématique dans les erreurs du niveau des estimations provisoires.

(II) **Erreurs relatives**

Quand le critère de choix de la section 3.1 nous porte à l'analyse des erreurs relatives, deux indices synthétiques appropriés sont l'erreur relative absolue moyenne et la racine carrée de l'erreur relative quadratique moyenne, respectivement:

$$\bar{e}' = \sum |e_t|/n, \tag{10}$$

$$d_e = \sqrt{(\sum e_t^2/n)}. \tag{11}$$

Les deux indices sont définis si $r_t \neq 0$ pour chaque $t$, condition qui n'est pas restrictive pour la plupart des agrégats économiques.

Une décomposition raisonnable de $d_e$ est la suivante:

$$d_e^2 = [(\overline{p/r}) - 1]^2 + \frac{1}{n} \sum [(p_t/r_t) - (\overline{p/r})]^2, \tag{12}$$

dont découle la décomposition relative

$$1 = U_e^b + U_e^c, \tag{13}$$

$U_e^b$ étant la fraction de l'erreur relative quadratique moyenne due au biais et $U_e^c$ étant la fraction due à la composante aléatoire.

Deux autres composantes spécifiques de manque de précision, la moyenne et l'écart quadratique moyen des erreurs relatives (c'est-à-dire respectivement $\bar{e} = \sum e_t/n = (\overline{p/r}) - 1$ et $s_e = \sqrt{(\sum (e_t - \bar{e})^2/n)}$) sont aussi d'intérêt. *Mutatis mutandis*, ce qui a été dit pour $\bar{v}$ est valable pour $\bar{e}$, autant pour son interprétation que pour la comparaison avec $\bar{e}'$.

### 3.3 Décomposition du processus de révision de l'estimation préliminaire à la définitive

Les indicateurs de précision qui ont été présentés jusqu'ici concernent la comparaison entre deux vecteurs $p$ et $r$. Nous avons déjà observé, cependant, que le processus de révision des agrégats économiques ne se termine usuellement pas après un seul stade. Comment, donc, estimer synthétiquement la convergence de la succession des estimations provisoires à l'estimation définitive?

Nous considérons ici la décomposition d'un processus de révision qui s'articule en deux stades. (Une procédure pour la situation plus générale de $m - 1$ étapes se trouve dans Wilton et Smith 1974.) Cette décomposition sera évaluée, dans l'ordre, pour les erreurs et pour les erreurs relatives.

Quand nous analysons les erreurs, la décomposition de l'erreur quadratique moyenne de ${}_1p$ par rapport à $r$ dans les deux phases de ${}_1p$ à ${}_2p$ et de ${}_2p$ à $r$ est simple. Soit ${}_1v = {}_1p - {}_2p$, ${}_2v = {}_2p - r$, ${}_Tv = {}_1p - r$ et l'on indique par $d^2_{v1}$, $d^2_{v2}$, $d^2_{vT}$ l'erreur quadratique moyenne associée respectivement aux vecteurs ${}_1v$, ${}_2v$ et ${}_Tv$. Il en résulte:

$$d^2_{vT} = d^2_{v1} + d^2_{v2} + \frac{2}{n} \sum {}_1v_t \, {}_2v_t , \tag{14}$$

dont on tire la décomposition relative:

$$1 = D_I + D_{II} + D_{I,II}, \tag{15}$$

où les deux premières composantes représentent la fraction de $d^2_{vT}$ attribuable à l'erreur quadratique moyenne de la première et de la seconde révision respectivement, tandis que $D_{I,II}$ représente l'intéraction entre ${}_1v$ et ${}_2v$.

En opérant sur les erreurs relatives, une décomposition synthétique du processus de révision dans les deux stades peut, au contraire, être problématique. D'utiles indications peuvent être tirées de l'identité $({}_1p_t - r_t)/r_t = ({}_1p_t - {}_2p_t)/r_t + ({}_2p_t - r_t)/r_t$, dont en divisant les deux membres par $({}_1p_t - r_t)/r_t$ et en considérant les valeurs moyennes, nous obtenons:

$$1 = \frac{1}{n} \sum ({}_1p_t - {}_2p_t)/({}_1p_t - r_t) + \frac{1}{n} \sum ({}_2p_t - r_t)/({}_1p_t - r_t). \tag{16}$$

Sous l'hypothèse d'ordre stable des estimations, du type ${}_1p_t < {}_2p_t < r_t$ pour chaque $t$, les deux termes de l'identité (16) sont interprétables comme les fractions moyennes de la discordance entre ${}_1p$ et $r$ éliminées par la première et par la seconde révision respectivement. Cette interprétation est toutefois discutable lorsqu'il y a violation de l'ordre, surtout lorsqu'elle apparaît conjointement avec des petites différences ${}_1p_t - r_t$. En général, il est plus raisonnable d'écarter l'hypothèse que l'ordre entre les estimations soit respecté durant toute la période considérée. Dans cette situation, une appréciation qualitative du degré de stabilité de la relation d'ordre entre les estimations, et de l'importance des deux stades du processus de révision, peut être fournie par l'inspection des valeurs de $\bar{e}$ et $\bar{e}'$ associées aux comparaisons entre ${}_1p$ et ${}_2p$, entre ${}_2p$ et $r$ et entre ${}_1p$ et $r$. Evidemment, la relation $| \bar{e} | = \bar{e}'$ n'est valable que si l'ordre entre les deux séries d'estimations faisant l'objet de la comparaison est toujours respecté. Nous pouvons cependant retenir que l'ordre est "le plus souvent" respecté quand $| \bar{e} | \simeq \bar{e}'$, en entendant par là:

$$(\bar{e}' - | \bar{e} |)/\bar{e}' < 2f, \tag{17}$$

où $f$ est une fraction de violations de l'ordre fixée à l'avance, considérée comme étant acceptable.

L'équation (17) se justifie comme suit. Soient $e_t (t = 1, 2, \ldots, n)$ les erreurs relatives, $n_1$ étant strictement positives, $n_2$ strictement négatives et $n_3$ égales à zéro $(n_1 + n_2 + n_3 = n)$. Il est facile de vérifier la relation suivante:

$$\bar{e}' - | \bar{e} | = \begin{cases} 2n^{-1} \sum_{i=1}^{n1} e_i & \text{si} \quad \bar{e} < 0 \\ - 2n^{-1} \sum_{j=1}^{n2} e_j & \text{si} \quad \bar{e} > 0, \end{cases}$$

où les quantités $\sum_{i=1}^{n1} e_i$ et $- \sum_{j=1}^{n2} e_j$ correspondent à la valeur absolue de la somme des violations au signe de $\bar{e}$.

La quantité

$$(\bar{e}' - |\bar{e}|)/\bar{e}' = \begin{cases} 2\sum_{i=1}^{n1} e_i / \left(\sum_{i=1}^{n1} e_i - \sum_{j=1}^{n2} e_j\right) & \text{si} \quad \bar{e} < 0 \\ -2\sum_{j=1}^{n2} e_j / \left(\sum_{i=1}^{n1} e_i - \sum_{j=1}^{n2} e_j\right) & \text{si} \quad \bar{e} < 0 \end{cases}$$

est donc un indice de l'importance des violations du signe de $\bar{e}$. Cet indice est borné, $0 \leq (\bar{e}' - |\bar{e}|)/\bar{e}' \leq 1$, la borne inférieure étant atteinte lorsque $|\bar{e}| = \bar{e}'$, c'est-à-dire lorsque $n_1 = 0$ ou $n_2 = 0$ (si $n_1 = n_2 = 0$ il n'existe pas de problèmes de violations de l'ordre), et la borne supérieure étant atteinte lorsque $\bar{e} = 0$, c'est-à-dire lorsque $\sum_{i=1}^{n1} e_i = \sum_{j=1}^{n2} e_j$.

La relation $|\bar{e}| = \bar{e}'$ est définie en comparant la valeur assumée par l'indice avec une valeur critique calculée sous l'hypothèse de l'égalité, en valeur absolue, de toutes les erreurs relatives non-nulles. Ceci donne le résultat suivant:

$$(\bar{e}' - |\bar{e}|)/\bar{e}' = \begin{cases} 2n_1/(n_1 + n_2) & \text{si} \quad \bar{e} < 0 \\ 2n_2/(n_1 + n_2) & \text{si} \quad \bar{e} > 0. \end{cases}$$

Sous cette hypothèse particulière, l'indice est égal au double de la fraction des violations strictes au signe de $\bar{e}$.

Il est donc possible de: (i) ordonner les estimations ${}_1p$ et ${}_2p$ sur la base d'un critère moyen, c'est-à-dire des valeurs de $\bar{e}$ associées aux différentes comparaisons; (ii) examiner le degré de stabilité de l'ordre sur la base de la relation (17); (iii) lorsq'on constate un degré élevé de stabilité, utiliser la décomposition (16) ou de quelque manière estimer qualitativement l'importance des deux stades en observant les valeurs de $\bar{e}$ associées aux différentes comparaisons.

## 4. IMPLICATIONS DES ERREURS DANS LES ESTIMATIONS PROVISOIRES SUR LES AGRÉGATS À PRIX CONSTANTS

Le processus de révision a évidemment des implications sur les séries dérivées, qui constituent souvent l'information représentant le plus grand intérêt pour analystes et décideurs. Ceci vaut en particulier pour l'estimation du taux de variation, les séries désaisonnalisées et les agrégats à prix constants et les déflateurs implicites. L'impact des révisions sur la mesure de la variation a été analysé d'une façon descriptive par Trivellato, Di Fonzo et Rettore (1986), et par Rao, Srinath et Quenneville (1989) dans la perspective de déterminer le meilleur estimateur de la variation. Les effets des révisions sur les procédures de désaisonnalisation sont traités dans les articles de Pierce (1980), Wallis (1982) et Maravall et Pierce (1983). Les implications du processus de révision sur les agrégats à prix constants et les déflateurs implicites sont examinées dans cette section.

Les critères et les indicateurs présentés dans la section 3.2 peuvent évidemment être utilisés pour l'analyse de la précision des estimations provisoires à prix constants. Il est toutefois intéressant d'illustrer les caractéristiques formelles des erreurs de ces estimations, et mettre en évidence les relations entre les erreurs d'estimation dans les agrégats à prix courants et ceux à prix constants ainsi que les déflateurs implicites. L'attention est circonscrite aux agrégats des comptes économiques nationaux constitués par des flux de marchandises, et aux agrégats obtenus par solde comptable de ceux-ci.

Considérons le cas simple d'une série d'estimations provisoires et de la série d'estimations définitives correspondante. En observant un quelconque agrégat au temps $t$, composé de $k$ biens et services élémentaires, les estimations définitives s'obtiennent comme suit (la sommation se faisant sur les $k$ biens et services élémentaires): $A_t = \sum p_t q_t$ est l'agrégat à prix courants; $C_t = \sum p_o q_t$ l'agrégat à prix constants (base 0); $P_t$ l'indice de prix, de type Paasche (c'est-à-dire, $P_t = \sum p_t q_t / \sum p_o q_t$); $Q_t$ l'indice de quantité, de type Laspeyres (c'est-à-dire, $Q_t = \sum p_o q_t / \sum p_o q_o$); $D_t = A_t / C_t$ le déflateur implicite, dont la structure est de type Paasche. Les estimations provisoires correspondantes sont identifiées avec l'indice $p$ (par exemple, ${}_pA_t$ et $A_t$ sont respectivement l'estimation provisoire et définitive de l'agrégat à prix courants).

Pour expliciter les caractéristiques des erreurs dans les estimations provisoires des agrégats à prix constants, il est utile de rappeler que l'évaluation de l'agrégat à prix constants au temps $t$ peut être obtenue par trois voies: (i) si nous disposons de la série des quantités pour tous les biens et services qui composent l'agrégat et des prix correspondants pour l'année de base, nous appliquons la relation directe: $C_t = \sum p_o q_t$; (ii) si nous disposons d'un indice de quantité de type Laspeyres, nous multiplions la valeur de l'agrégat dans l'année de base par cet indice: $C_t = (\sum p_o q_o)\, Q_t$; (iii) si nous disposons d'un indice de prix de type Paasche, nous divisons l'agrégat à prix courants par cet indice: $C_t = (\sum p_t q_t)/P_t$.

En relation avec les différents critères d'évaluation utilisables, l'erreur dans l'estimation provisoire du niveau (ainsi que l'erreur relative) d'un agrégat à prix constants est donnée par:

$$ {}_pC_t - C_t \equiv \sum p_o\, {}_pq_t - \sum p_o q_t = \sum p_o q_o \left( \frac{\sum p_o\, {}_pq_t}{\sum p_o q_o} - \frac{\sum p_o q_t}{\sum p_o q_o} \right) = A_o({}_pQ_t - Q_t), \quad (18.1) $$

$$ \equiv (\sum p_o q_o)\, {}_pQ_t - (\sum p_o q_o)\, Q_t = A_o({}_pQ_t - Q_t), \quad (18.2) $$

$$ \equiv {}_pA_t/{}_pP_t - A_t/P_t = (\sum {}_pp_t\, {}_pq_t)/{}_pP_t - (\sum p_t q_t)/P_t . \quad (18.3) $$

Les équations (18) montrent que si nous adoptons le critère d'évaluation directe (18.1) ou celui d'extrapolation au moyen d'un indice de quantité (18.2), l'erreur d'estimation coïncide avec l'erreur d'estimation dans l'indice de quantité, multiplié par la constante $A_o$. Nous n'avons pas en revanche un tel résultat en utilisant le critère indirect de déflation au moyen d'un indice de prix, parce qu'en général l'équation (18.3) ne se simplifie pas d'une façon utile. En d'autres mots, dans ce troisième cas l'erreur dans l'estimation du niveau de l'agrégat à prix constants dépend, en général, aussi bien des erreurs dans les estimations provisoires des quantités que de celles dans les prix.

L'implication pour l'interprétation des erreurs dans les estimations des agrégats à prix constants est vite tirée. Même si théoriquement les trois critères d'évaluation sont identiques, cela est loin d'être le cas en pratique, en raison de la disponibilité et de la qualité des données. Il faut se rappeler que l'évaluation à prix constants est généralement effectuée à un niveau très désagrégé, et que les agrégats sont obtenus ensuite par sommation. L'évaluation de ces agrégats se faisant donc partiellement avec le critère de la déflation, les erreurs dans les estimations à prix constants ne sont pas uniquement des erreurs dans les estimations du volume, mais incorporent aussi les erreurs dans les estimations des prix.

Si nous considérons maintenant l'erreur relative d'estimation d'un agrégat à prix courants et si nous en explicitons la relation avec l'erreur relative de l'agrégat à prix constants correspondant, nous avons l'identité suivante:

$$ \frac{{}_pA_t - A_t}{A_t} = \frac{{}_pC_t - C_t}{C_t} + \frac{{}_pD_t - D_t}{D_t} + \frac{{}_pC_t - C_t}{C_t} \times \frac{{}_pD_t - D_t}{D_t} . $$

Cette identité est évidemment valable pour l'erreur relative moyenne d'une série de $n$ estimations, et nous obtenons la décomposition relative suivante:

$$1 = \bar{e}(C)/\bar{e}(A) + \bar{e}(D)/\bar{e}(A) + \bar{e}(C;D)/\bar{e}(A), \tag{19}$$

où

$$\bar{e}(A) = \frac{1}{n} \Sigma \left( \frac{{}_pA_t - A_t}{A_t} \right)$$

($\bar{e}(C)$ et $\bar{e}(D)$ sont définis de manière analogue) et

$$\bar{e}(C;D) = \frac{1}{n} \Sigma \left( \frac{{}_pC_t - C_t}{C_t} \right) \left( \frac{{}_pD_t - D_t}{D_t} \right).$$

En laissant de côté la composante d'interaction, nous avons donc une décomposition approximativement additive: l'erreur relative moyenne de l'agrégat à prix courants est égale à la somme des erreurs relatives moyennes de l'agrégat à prix constants et du déflateur implicite. (Ou mieux, en ayant présent à l'esprit le processus d'estimation: l'erreur relative moyenne du déflateur implicite est approximativemem égale à la différence entre les erreurs relatives moyennes de l'agrégat à prix courants et de celui à prix constants.)

La décomposition, notons-le bien, est entre $\bar{e}(C)$ et $\bar{e}(D)$, et non pas entre "erreur dans les quantités" et "erreur dans les prix". Nous avons déjà vu que, en général, ${}_pC_t - C_t$ et $\bar{e}(C)$ reflètent aussi les erreurs dans les estimations provisoires des prix. D'autre part, par définition, ${}_pD_t - D_t$ et $\bar{e}(D)$ sont fonction des erreurs dans les estimations provisoires tant des prix que des quantités. En définitive, donc, les deux composantes "erreur d'estimation à prix constants" et "erreur d'estimation du déflateur implicite" incorporent toutes les deux des erreurs d'estimation de prix et quantités, et la possibilité d'interpréter ${}_pC_t - C_t$ et ${}_pD_t - D_t$ comme "erreur dans les quantités" et "erreur dans les prix" ne peut se faire qu'en des cas extrêmes.

Nous pouvons identifier les cas extrêmes suivants:

(a) ${}_pC_t - C_t$ est interprétable comme "erreur dans les quantités" seulement lorsque l'évaluation de l'agrégat à prix constants se fait avec le critère direct ou par extrapolation au moyen d'un indice de quantité.

(b) ${}_pD_t - D_t$ est interprétable comme "erreur dans les prix" seulement en l'absence de révision dans les quantités. Dans ce cas l'erreur d'estimation à prix constants est évidemment nulle, et l'erreur relative à prix courants peut s'exprimer comme une combinaison linéaire des erreurs d'estimation dans les prix des $k$ biens et services élémentaires:

$$({}_pA_t - A_t)/A_t = ({}_pD_t - D_t)/D_t = \Sigma q_t({}_pp_t - p_t)/\Sigma p_t q_t = \Sigma a_t({}_pp_t - p_t),$$

où

$$a_t = q_t/\Sigma p_t q_t.$$

(c) S'il n'y a aucune révision dans les prix, on a:

$$({}_pA_t - A_t)/A_t = \Sigma p_t({}_pq_t - q_t)/\Sigma p_t q_t = ({}_pC_t/C_t)({}_pD_t/D_t) - 1.$$

L'absence de révision dans les prix n'est donc pas suffisante pour distinguer entre "erreur dans les quantités" et "erreur dans les prix", puisque la révision dans les quantités modifie le déflateur implicite.

## 5. QUELQUES ANALYSES EMPIRIQUES SUR LES RÉVISIONS DES DONNÉES DE COMPTABILITÉ NATIONALE DU CANADA, DES ÉTATS-UNIS ET DE L'ITALIE

### 5.1 Les processus de révision des données de comptabilité nationale dans les trois pays et les analyses effectuées

Dans cette dernière section, nous présentons sommairement quelques résultats de l'analyse empirique des révisions des comptes nationaux du Canada, des États-Unis et de l'Italie.

La comparaison des caractéristiques du processus de révision des trois pays considérés comporte un certain nombre de simplifications inévitables dans la description et dans l'analyse des révisions. Pour des études récentes et approfondies relatives à chacun des pays en question, il est utile de se référer à Lefrançois (1988) pour le Canada, à Parker (1986) et Mork (1987) pour les États-Unis et à Di Fonzo, Rettore et Trivellato (1986) pour l'Italie.

Une schématisation des processus de révision des données de comptabilité nationale annuelle et trimestrielle dans les trois pays est reportée dans les tableaux 2, 3 et 4.

Les données sur lesquelles portent les analyses empiriques sont les suivantes:

(a) Canada: séries des estimations trimestrielles non désaisonnalisées et des estimations annuelles, à prix courants, pour la période 1953-1982. Il s'agit de la base des données habituellement utilisée par Statistique Canada pour l'analyse des révisions.

(b) États-Unis: séries des estimations trimestrielles désaisonnalisées, pour la période 1968-1983. Dans ce cas aussi il s'agit de la base des données habituellement utilisée par le *Bureau of Economic Analysis* (BEA) pour l'analyse des révisions, mais contrairement aux données canadiennes, elles ne coïncident pas avec les données publiées, puisque le BEA y apporte des ajustements et corrections pour éliminer l'effet de changements de définition.

(c) Italie: séries des estimations annuelles, pour la période 1961-1985. (Nous n'avons pas considéré les estimations trimestrielles, car leur publication, commencée en 1976, s'est interrompue en 1982.) Il s'agit des données publiées par l'*Istituto Centrale di Statistica* (ISTAT).

En tenant compte des données disponibles, il est donc possible de comparer le processus trimestriel du Canada avec celui des États-Unis et le processus annuel du Canada avec celui de l'Italie.

Afin d'effectuer des analyses comparatives entre les processus de révision des estimations trimestrielles du Canada et des États-Unis, il faut en premier lieu établir les séries des États-Unis pouvant être utilisées en parallèle avec les séries canadiennes. En tenant compte du décalage temporel entre les estimations successives, des données disponibles et des caractéristiques du processus de révision, nous avons choisi les séries du tableau 3. De cette façon, dans les deux pays on dispose, pour chaque trimestre de référence, de trois estimations auxquelles nous donnons le même sigle: $P$ est la série des premières estimations trimestrielles, $RT$ est la série des révisions trimestrielles choisie et $A1$ est la série des premières révisions annuelles. À ces trois estimations nous ajoutons l'estimation finale, qui coïncide avec la dernière estimation publiée, que nous désignons par $F$.

Il apparaît alors convenable d'effectuer les comparaisons suivantes:

($P$,$RT$), qui rend compte de l'effet de la révision de type trimestriel (infra-annuel). Il faut pourtant noter qu'au Canada, seules les données préliminaires des trois premiers trimestres d'une année sont sujettes à des révisions strictement trimestrielles, puisque la révision apportée aux données d'un quatrième trimestre (désignée par ${}_2p_{t,iv}$ dans le tableau 2) est composée d'une révision trimestrielle et d'une révision annuelle. Ceci se traduit, entre autres, par une différence de comportement des révisions qui est nettement perceptible (voir Lefrançois 1988). On a tenu compte de cette particularité d'une façon simple, en limitant l'analyse aux seuls trois premiers trimestres de chaque année.

($RT$,$A1$), qui informe de la contribution apportée par l'étalonnage annuel au processus de révision des données trimestrielles.

$(A1,F)$, qui nous permet d'évaluer l'effet de l'estimation aux jalons quinquennaux aux États-Unis et de le confronter à l'effet du processus de révision extraordinaire au Canada. À ce propos, il faut d'ailleurs observer que l'évaluation comparée doit être effectuée avec une extrême prudence, puisque les révisions historiques canadiennes sont en majeure partie causées par des changements de définitions, et ne sont donc pas strictement comparables à l'ajustement quinquennal des données aux États-Unis.

En ce qui concerne la comparaison entre les révisions courantes sur les données annuelles en Italie et au Canada, des problèmes mineurs de comparabilité subsistent, car dans les deux pays le processus de révision se déroule en deux stades. Nous désignons les trois séries d'estimations qui en résultent par ${}_1p$, ${}_2p$ et $r$. Les comparaisons d'intérêt sont alors: $({}_1p,r)$ pour évaluer l'effet global du processus de révision courant; $({}_2p,r)$ pour vérifier dans quelle mesure la révision s'accomplit au second stade (et indirectement tirer des indications sur le premier stade); la comparaison simultanée $({}_1p,{}_2p,r)$ pour évaluer la convergence de la succession d'estimations.

**Tableau 2**

Schéma de publication des estimations trimestrielles et annuelles des agrégats de comptabilité nationale au Canada (révisions courantes seulement)[a]

| Période de publication | | Période de référence | | | | |
|---|---|---|---|---|---|---|
| | | Données trimestrielles | | | | Donnée annuelle |
| Année | Mois et trimestre | I | II | III | IV | |
| $t$ | 1 | | | | | |
| | 2 | | | | | |
| | 3 | | | | | |
| | 4 | | | | | |
| | 5 | ${}_1p_{t.I}$ | | | | |
| | 6 | | | | | |
| | 7 | | | | | |
| | 8 | ${}_2p_{t.I}$ | ${}_1p_{t.II}$ | | | |
| | 9 | | | | | |
| | 10 | | | | | |
| | 11 | ${}_3p_{t.I}$ | ${}_2p_{t.II}$ | ${}_1p_{t.III}$ | | |
| | 12 | | | | | |
| $t+1$ | 1 | | | | | |
| | 2 | ${}_4p_{t.I}$ | ${}_3p_{t.II}$ | ${}_2p_{t.III}$ | ${}_1p_{t.IV}$ | ${}_1p_t$ |
| | 3 | | | | | |
| | II | ${}_5p_{t.I}$ | ${}_4p_{t.II}$ | ${}_3p_{t.III}$ | ${}_2p_{t.IV}$ | ${}_2p_t$ |
| | III | | | | | |
| | IV | | | | | |
| $t+2$ | I | | | | | |
| | II | $r_{t.I}$ | $r_{t.II}$ | $r_{t.III}$ | $r_{t.IV}$ | $r_t$ |
| | III | | | | | |
| | IV | | | | | |

[a] Les séries mises en évidence sont celles que nous avons utilisées pour les analyses empiriques des révisions dans les données trimestrielles.

**Tableau 3**

Schéma de publication des estimations trimestrielles et annuelles des agrégats de comptabilité nationale aux États-Unis[a]

| Période de publication | | Période de référence | | | | |
|---|---|---|---|---|---|---|
| | | Données trimestrielles | | | | Donnée annuelle |
| Année | Mois et trimestre | I | II | III | IV | |
| $t$ | 1 | | | | | |
| | 2 | | | | | |
| | 3 | $_0p_{t.I}$ | | | | |
| | 4 | $_1p_{t.I}$ | | | | |
| | 5 | $_2p_{t.I}$ | | | | |
| | 6 | $_3p_{t.I}$ | $_0p_{t.II}$ | | | |
| | 7 | | $_1p_{t.II}$ | | | |
| | 8 | | $_2p_{t.II}$ | | | |
| | 9 | | $_3p_{t.II}$ | $_0p_{t.III}$ | | |
| | 10 | | | $_1p_{t.III}$ | | |
| | 11 | | | $_2p_{t.III}$ | | |
| | 12 | | | $_3p_{t.III}$ | $_0p_{t.IV}$ | |
| $t+1$ | 1 | | | | $_1p_{t.IV}$ | |
| | 2 | | | | $_2p_{t.IV}$ | |
| | 3 | | | | $_3p_{t.IV}$ | |
| | II | | | | | |
| | III | $_4p_{t.I}$ | $_4p_{t.II}$ | $_4p_{t.III}$ | $_4p_{t.IV}$ | $_1p_t$ |
| | IV | | | | | |
| $t+2$ | I | | | | | |
| | II | | | | | |
| | III | $_5p_{t.I}$ | $_5p_{t.II}$ | $_5p_{t.III}$ | $_5p_{t.IV}$ | $_2p_t$ |
| | IV | | | | | |
| $t+3$ | I | | | | | |
| | II | | | | | |
| | III | $r_{t.I}$ | $r_{t.II}$ | $r_{t.III}$ | $r_{t.IV}$ | $r_t$ |
| | IV | | | | | |
| $t+5$ | IV | $_br'_{t.I}$ | $_br'_{t.II}$ | $_br'_{t.III}$ | $_br'_{t.IV}$ | $_br'_t$ |
| $t+10$ | IV | $_br''_{t.I}$ | $_br''_{t.II}$ | $_br''_{t.III}$ | $_br''_{t.IV}$ | $_br''_t$ |

[a] Les séries mises en évidence sont celles que nous avons utilisées pour les analyses empiriques des révisions dans les données trimestrielles.

Les analyses ont été menées sur les quatre agrégats suivants:

*PL* = Produit intérieur brut pour l'Italie; Produit national brut pour le Canada et les États-Unis.

*CF* = Consommation finale intérieure des ménages et des administrations privées pour l'Italie; Dépenses personnelles en biens et services pour le Canada et les États-Unis.

**Tableau 4**

Schéma de publication des estimations trimestrielles et annuelles des agrégats de comptabilité nationale en Italie (révisions courantes seulement)

| Période de publication | | Période de référence | | | | |
|---|---|---|---|---|---|---|
| | | Données trimestrielles | | | | Donnée annuelle |
| Année | Mois et trimestre | I | II | III | IV | |
| $t$ | 1 | | | | | |
| | 2 | | | | | |
| | 3 | | | | | |
| | 4 | | | | | |
| | 5 | | | | | |
| | 6 | ${}_1p_{t.\mathrm{I}}$ | | | | |
| | 7 | | | | | |
| | 8 | | | | | |
| | 9 | | ${}_1p_{t.\mathrm{II}}$ | | | |
| | 10 | | | | | |
| | 11 | | | | | |
| | 12 | | | ${}_1p_{t.\mathrm{III}}$ | | |
| $t+1$ | 1 | | | | | |
| | 2 | | | | | |
| | 3 | ${}_2p_{t.\mathrm{I}}$ | ${}_2p_{t.\mathrm{II}}$ | ${}_2p_{t.\mathrm{III}}$ | ${}_1p_{t.\mathrm{IV}}$ | ${}_1p_t$ |
| | II | | | | | |
| | III | | | | | |
| | IV | | | | | |
| $t+2$ | I | ${}_3p_{t.\mathrm{I}}$ | ${}_3p_{t.\mathrm{II}}$ | ${}_3p_{t.\mathrm{III}}$ | ${}_2p_{t.\mathrm{IV}}$ | ${}_2p_t$ |
| | II | | | | | |
| | III | | | | | |
| | IV | | | | | |
| $t+3$ | I | $r_{t.\mathrm{I}}$ | $r_{t.\mathrm{II}}$ | $r_{t.\mathrm{III}}$ | $r_{t.\mathrm{IV}}$ | $r_t$ |
| | II | | | | | |
| | III | | | | | |
| | IV | | | | | |

$CP$ = Consommation collective des administrations publiques pour l'Italie; Dépenses publiques en biens et services pour le Canada et les États-Unis.

$IN$ = Formation brute de capital fixe.

Il faut se rappeler que les différents schémas comptables de référence (SNA pour le Canada et les États-Unis, SEC pour l'Italie) font que la configuration des agrégats n'est pas complètement homogène dans les trois pays.

Les principaux résultats portant sur l'analyse des erreurs dans les estimations du niveau des agrégats sont reportés dans les tableaux 5 à 8. Des résultats plus complets sont toutefois disponibles, et nous y ferons référence si l'occasion se présente.

### 5.2 Une tentative de synthèse comparative des résultats des analyses

La comparaison des trois schémas des tableaux 2, 3 et 4, nous amène les réflexions suivantes.

Le système de production et de révision des estimations de comptabilité nationale des États-Unis apparaît beaucoup plus structuré et fortement orienté en faveur de la comptabilité trimestrielle, au sujet de laquelle il faut remarquer l'extrême rapidité dans la publication des estimations préliminaires. Lorsque le Canada produit (et l'Italie produisait) la première estimation trimestrielle, aux États-Unis celle-ci a déjà été révisée deux ou trois fois, en utilisant une importante quantité d'informations directes additionelles (Parker 1986). Moins d'importance est donnée, au contraire, à l'estimation annuelle, qui sort avec un retard de neuf mois.

Par contre, la rapidité semble être un trait distinctif des estimations annuelles italiennes, puisqu'elles sortent au premier trimestre de l'année suivante, tandis qu'au Canada la première estimation annuelle n'est disponible qu'au second trimestre (l'estimation préliminaire publiée dans le premier trimestre est une simple agrégation des données trimestrielles). Cette caractéristique n'est cependant pas nécessairement une qualité, et il est raisonnable de penser que les temps rapides de production des premières estimations annuelles dépendent aussi des lacunes du système italien de comptabilité trimestrielle.

Enfin, il faut souligner que seuls les États-Unis effectuent avec régularité non seulement des révisions aux jalons annuels pour les estimations trimestrielles, mais aussi des révisions aux jalons de plus longue périodicité, fondées sur les tableaux entrée-sortie et sur les résultats de relevés plus amples, des recensements par exemple. Ceci assure une plus grande cohérence, non seulement comptable, aux estimations des différents agrégats. Au Canada et en Italie aussi il y a des révisions analogues, effectuées cependant plus occasionnellement et presque toujours accompagnées de modifications dans les critères de classification ou d'estimation.

**Tableau 5**

Indicateurs de la précision des estimations provisoires
du niveau des agrégats à prix courants
Données trimestrielles – Canada

| Agrégats | $\bar{e}$ | $\bar{e}'$ | $s_e$ | $d_e$ | $U_e^b$ |
|---|---|---|---|---|---|
| | Comparaisons entre $P$ et $RT$ (période 1953-83; $T = 89$) | | | | |
| *PL* | −.0018 | .0034 | .0042 | .0046 | .1567 |
| *CF* | −.0009 | .0025 | .0039 | .0040 | .0560 |
| *CP* | −.0041 | .0106 | .0193 | .0197 | .0425 |
| *IN* | −.0047 | .0133 | .0206 | .0211 | .0498 |
| | Comparaisons entre $RT$ et $A1$ (période 1953-83; $T = 69$) | | | | |
| *PL* | −.0046 | .0064 | .0065 | .0080 | .3372 |
| *CF* | −.0054 | .0069 | .0086 | .0102 | .2877 |
| *CP* | −.0008 | .0176 | .0262 | .0262 | .0009 |
| *IN* | −.0075 | .0205 | .0270 | .0280 | .0721 |
| | Comparaisons entre $A1$ et $F$ (période 1953-70; $T = 72$) | | | | |
| *PL* | −.0556 | .0556 | .0197 | .0590 | .8882 |
| *CF* | −.0617 | .0617 | .0192 | .0646 | .9110 |
| *CP* | −.0673 | .0727 | .0585 | .0892 | .5699 |
| *IN* | −.0228 | .0311 | .0326 | .0398 | .0328 |

**Tableau 6**

Indicateurs de la précision des estimations provisoires du niveau des agrégats à prix courants
Données annuelles - Canada

| Agrégats | $\bar{e}$ | $\bar{e}'$ | $s_e$ | $d_e$ | $U_e^b$ |
|---|---|---|---|---|---|
| | Comparaisons entre ${}_1p$ et $r$ (période 1953-83; $n = 20$) | | | | |
| *PL* | −.0084 | .0093 | .0058 | .0102 | .6720 |
| *CF* | −.0080 | .0086 | .0077 | .0112 | .5191 |
| *CP* | −.0069 | .0118 | .0136 | .0153 | .2052 |
| *IN* | −.0173 | .0183 | .0114 | .0208 | .6966 |
| | Comparaisons entre ${}_2p$ et $r$ (période 1953-83; $n = 24$) | | | | |
| *PL* | −.0032 | .0041 | .0044 | .0054 | .3444 |
| *CF* | −.0029 | .0030 | .0048 | .0056 | .2619 |
| *CP* | −.0038 | .0063 | .0087 | .0096 | .1627 |
| *IN* | −.0040 | .0042 | .0075 | .0085 | .2274 |
| | Comparaisons entre $r$ et $F$ (période 1953-70; $n = 18$) | | | | |
| *PL* | −.0464 | .0464 | .0188 | .0500 | .8588 |
| *CF* | −.0544 | .0544 | .0200 | .0580 | .8809 |
| *CP* | −.0523 | .0523 | .0348 | .0629 | .6928 |
| *IN* | −.0130 | .0147 | .0140 | .0192 | .4637 |

**Tableau 7**

Indicateurs de la précision des estimations provisoires du niveau des agrégats à prix courants
Données trimestrielles - États-Unis

| Agrégats | $\bar{e}$ | $\bar{e}'$ | $s_e$ | $d_e$ | $U_e^b$ |
|---|---|---|---|---|---|
| | Comparaisons entre $P$ et $RT$ (période 1968-83; $T = 64$) | | | | |
| *PL* | −.0017 | .0031 | .0040 | .0044 | .1459 |
| *CF* | −.0014 | .0041 | .0054 | .0055 | .0607 |
| *CP* | −.0013 | .0050 | .0064 | .0065 | .0433 |
| *IN* | −.0047 | .0104 | .0138 | .0146 | .1052 |
| | Comparaisons entre $RT$ et $A1$ (période 1968-83; $T = 52$) | | | | |
| *PL* | −.0054 | .0064 | .0071 | .0089 | .3706 |
| *CF* | −.0049 | .0073 | .0078 | .0092 | .2854 |
| *CP* | .0001 | .0088 | .0109 | .0109 | .0002 |
| *IN* | −.0074 | .0176 | .0237 | .0249 | .0878 |
| | Comparaisons entre $A1$ et $F$ (période 1968-83; $T = 52$) | | | | |
| *PL* | −.0093 | .0097 | .0070 | .0117 | .6411 |
| *CF* | −.0040 | .0070 | .0088 | .0096 | .1662 |
| *CP* | .0039 | .0070 | .0088 | .0096 | .1662 |
| *IN* | −.0408 | .0409 | .0253 | .0480 | .7222 |

**Tableau 8**

Indicateurs de la précision des estimations provisoires du niveau des agrégats à prix courants
Données annuelles - Italie

| Agrégats | $\bar{e}$ | $\bar{e}'$ | $s_e$ | $d_e$ | $U_e^b$ |
|---|---|---|---|---|---|
| | Comparaisons entre ${}_1p$ et $r$ (période 1963-85; $n = 13$) | | | | |
| *PL* | −.0095 | .0095 | .0093 | .0133 | .5091 |
| *CF* | −.0067 | .0068 | .0057 | .0088 | .5781 |
| *CP* | −.0111 | .0122 | .0129 | .0170 | .4249 |
| *IN* | −.0062 | .0078 | .0091 | .0110 | .3141 |
| GR.1 | −.0032 | .0113 | .0140 | .0144 | .0489 |
| 2 | −.0253 | .0282 | .0283 | .0380 | .4443 |
| 3 | −.0107 | .0379 | .0462 | .0475 | .0508 |
| 4 | −.0133 | .0208 | .0271 | .0302 | .1937 |
| 5 | −.0056 | .0101 | .0113 | .0126 | .1994 |
| 6 | −.0044 | .0172 | .0206 | .0210 | .0429 |
| 7 | −.0143 | .0146 | .0150 | .0207 | .4750 |
| 8 | −.0061 | .0132 | .0170 | .0181 | .1121 |
| Total | −.0106 | .0106 | .0106 | .0150 | .4987 |
| | Comparaisons entre ${}_2p$ et $r$ (période 1963-85; $n = 17$) | | | | |
| *PL* | −.0030 | .0037 | .0054 | .0062 | .2299 |
| *CF* | −.0015 | .0017 | .0024 | .0029 | .2900 |
| *CP* | −.0013 | .0049 | .0063 | .0064 | .0426 |
| *IN* | −.0021 | .0026 | .0052 | .0056 | .1445 |
| GR.1 | .0002 | .0026 | .0038 | .0038 | .0018 |
| 2 | −.0054 | .0096 | .0172 | .0180 | .0884 |
| 3 | −.0015 | .0096 | .0177 | .0177 | .0073 |
| 4 | −.0058 | .0060 | .0105 | .0120 | .2304 |
| 5 | .0008 | .0016 | .0026 | .0027 | .0954 |
| 6 | .0011 | .0055 | .0085 | .0085 | .0167 |
| 7 | −.0039 | .0069 | .0093 | .0101 | .1459 |
| 8 | −.0008 | .0008 | .0030 | .0031 | .0673 |
| Total | −.0026 | .0038 | .0058 | .0063 | .1648 |

Gr.1: Agriculture; Produits alimentaires et boissons; Tabac.
Gr.2: Textiles; Vêtement et chaussures; Peaux et cuir.
Gr.3: Bois et meubles en bois.
Gr.4: Minéraux non métal.; Produits chimiques; Industrie extractive; Énergie et eau.
Gr.5: Bâtiment et génie civil.
Gr.6: Transports et communications.
Gr.7: Autres branches.
Gr.8: Location de bâtiments.

Nous pouvons faire les observations suivantes à partir des tableaux 5 à 8:

(a) Dans la comparaison Canada-États-Unis, le profil des révisions courantes des estimations trimestrielles est substantiellement homogène, du moins en ce qui concerne les agrégats considérés et les limites du processus de révision décrit par les trois séries d'estimations disponibles. Les révisions infra-annuelles ont en moyenne des dimensions modestes (dans les deux cas l'agrégat *IN* fait exception) et un comportement assez irrégulier, bien que la tendance soit à réviser vers le haut. L'effet de la révision aux jalons annuels est relativement plus marqué et plus systématique, surtout en ce qui concerne *PL*, et tend aussi à rectifier vers le haut l'évaluation précédente; dans ce second stade du processus de révision c'est aussi l'agrégat *IN* qui montre des erreurs de dimension plus élevée et de nature plus irrégulière.

(b) Dans la comparaison Italie-Canada des estimations annuelles, nous remarquons aussi un comportement très uniforme. L'estimation préliminaire de *PL* sous-évalue en moyenne d'environ .9% la donnée finale courante; cette tendance se présente avec une systématicité marquée et avec un ordre suffisamment stable selon lequel ${}_1p \leq {}_2p \leq r$. Ceci entraîne une sous-estimation des taux de variation quand ils sont calculés, comme on le fait généralement, sur la base des comparaisons "horizontales". Il faut ajouter que le poids de la seconde révision est relativement plus marqué pour le Canada.

Dans l'ensemble, les résultats empiriques montrent que les estimations provisoires sont valides, surtout dans le cas des évaluations trimestrielles des États-Unis et du Canada. Au contraire, la systématicité de la sous-évaluation de la part des estimations préliminaires annuelles, aussi bien au Canada qu'en Italie, laisse perplexe, bien que la dimension de l'erreur ne soit certainement pas préoccupante.

Ce phénomène est répandu dans plusieurs pays (voir Glejser et Schavey 1974). Il y a toutefois certains aspects du cas italien qui méritent plus amples réflexions. Le premier est le comportement extrêmement différencié des erreurs au niveau désagrégé par branche d'activité économique, qui signale des situations où l'information est certainement faible: dans certaines branches on enregistre une sous-évaluation systématique de 6 à 7% en moyenne (Marliani 1986).

L'autre aspect est lié à l'effet marqué de réévaluation dû aux révisions extraordinaires, affectant généralement tous les agrégats (voir Di Fonzo, Rettore et Trivellato 1986). La tendance à réviser vers le haut les évaluations antérieures, à l'occasion de modifications apportées aux enquêtes statistiques de base et/ou aux méthodes d'estimation, n'est certainement pas particulière à l'Italie. Les données canadiennes elle-mêmes (et indirectement celles des États-Unis par l'effet de l'ajustement quinquennal aux jalons) montrent une tendance analogue.

Il s'agit de signaux qui mettent l'accent sur des aspects de faiblesse des systèmes des comptes nationaux et induisent à considérer avec moins d'assurance la petite taille moyenne des erreurs des estimations provisoires générées par le processus de révision courant.

## REMERCIEMENTS

Les résultats présentés dans cet article furent obtenus dans le cadre d'un projet de recherche financé par le Ministère de l'Education Nationale Italien. Les données ont été aimablement fournies par la Division de l'Analyse Economique Courante de Statistique Canada, pour le Canada, par Robert P. Parker, Directeur Adjoint pour les Comptes Nationaux auprès du BEA, pour les États-Unis, par l'ISTAT pour l'Italie. Nous tenons à remercier Silvano Bordignon, Tommaso Di Fonzo, Gianni Marliani et Enrico Rettore (qui faisaient partie du groupe de recherche), ainsi que les deux critiques qui, par leurs commentaires et leurs conseils, ont grandement amélioré cette étude. Nous sommes reconnaissants à Cristina Pozzato et Claudio Palmieri pour l'assistance dans les analyses empiriques, et surtout à Gianni Marliani qui nous a permis d'utiliser quelques-uns des résultats d'un de ses récents travaux (Marliani 1987). Une première version de cette étude a été présentée aux Journées de Statistique de l'ASU, qui se sont tenues à Lausanne du 18 au 20 mai 1987.

## BIBLIOGRAPHIE

BIGGERI, L. (1984). Caratteristiche e analisi dei processi di revisione nelle valutazioni di aggregati ed indici economici: alcuni confronti internazionali. *Note Economiche*, 4, 18-63.

BORDIGNON, S., et TRIVELLATO, U. (1989). On the optimal use of provisional data in forecasting with dynamic models. *Journal of Business & Economic Statistics*, 7, 275-286.

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

DI FONZO, T., RETTORE, E., et TRIVELLATO, U. (1986). L'accuratezza delle stime provvisorie degli aggregati di contabilità nazionale annuale. *Errori nei dati preliminari, previsioni e politiche economiche* (éd. U. Trivellato). Padova: Cleup, 113-145.

GLEJSER, H., et SCHAVEY, P. (1974). An analysis of revisions on national accounts data for 40 countries. *The Review of Income and Wealth*, 20, 317-332.

HATANAKA, M. (1974). A simple suggestion to improve the Mincer-Zarnowitz criterion for the evaluation of forecasts. *Annals of Economic and Social Measurement*, 3, 521-524.

HEMPENIUS, A.L. (1980). *Forecast accuracy analysis applied to forecasts of the Dutch Central Planning Bureau, 1964-1978*. Katholieke Hogeschool Tilburg, Subfaculteit der Econometrie.

KENDALL, M.G. (1973). *Time Series*. London: Griffin.

LEFRANÇOIS, B. (1988). Application des séries chronologiques à l'étude des révisions. *La Revue Canadienne de Statistique*, 16, 83-96.

MALINVAUD, E. (1969). *Méthodes statistiques de l'économétrie*. Paris: Dunod.

MANKIW, N.G., RUNKLE, D.E., et SHAPIRO, M.D. (1984). Are preliminary announcements of the money stock rational forecasts? *Journal of Monetary Economics*, 14, 15-27.

MANKIW, N.G., et SHAPIRO, M.D. (1986). News or noise? An analysis of GNP revisions. *Survey of Current Business*, 66, 20-25.

MARAVALL, A., et PIERCE, D.A. (1983). Preliminary data error and monetary aggregate targeting. *Journal of Business & Economic Statistics*, 1, 179-186.

MARLIANI, G. (1986). L'accuratezza delle stime provvisorie del valore aggiunto per settore di attività economica. *Errori nei dati preliminari, previsioni e politiche economiche* (éd. U. Trivellato). Padova: Cleup, 113-146.

MARLIANI, G. (1987). Stime provvisorie e revisioni dei dati di contabilità nazionale: il caso italiano ed alcuni confronti internazionali. *Attendibilità e tempestività delle stime di contabilità nazionale* (éd. U. Trivellato). Padova: Cleup, 35-78.

McNEES, S.K. (1986). Estimating GNP: The trade-off between timeliness and accuracy. *New England Economic Review*, January/February, 3-10.

MINCER, J., et ZARNOWITZ, V. (1969). The evaluation of economic forecasts. Dans *Economic Forecasts and Expectations* (éd. J. Mincer). New York: NBER, Columbia University Press, 3-46.

MORK, K.A. (1987). Ain't behavin': Forecast errors and measurement errors in early GNP estimates. *Journal of Business & Economic Statistics*, 5, 165-175.

NOVAK, G.J. (1975). Reliability criteria for national accounts. *The Review of Income and Wealth*, 21, 323-344.

PARKER, P.R. (1986). Revisioni nelle stime iniziali del prodotto nazionale lordo trimestrale degli Stati Uniti. *Errori nei dati preliminari, previsioni e politiche economiche* (éd. U. Trivellato). Padova: Cleup, 181-217.

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

RAO, J.N.K., SRINATH, K.P., et QUENNEVILLE, B. (1989). Estimation of level and change using current preliminary data. *Panel Surveys* (éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: Wiley, 457-479.

THEIL, H. (1966). *Applied Economic Forecasting*, Amsterdam: North-Holland.

TRIVELLATO, U. (1986a) (éd.). *Errori nei dati preliminari, previsioni e politiche economiche*. Padova: Cleup.

TRIVELLATO, U. (1986b). Sulla valutazione dell'accuratezza di stime provvisorie di aggregati economici. *Studi in onore di Silvio Vianelli*. Palermo: Università degli Studi, 1587-1620.

TRIVELLATO, U. (1987). Problemi e metodi di valutazione dell'attendibilità delle stime di contabilità nazionale. *Statistica*, 47, 365-388.

TRIVELLATO, U., DI FONZO, T., et RETTORE, E. (1986). L'accuratezza delle stime provvisorie di aggregati e di indici economici: orientamenti metodologici. *Errori nei Dati Preliminari, Previsioni e Politiche Economiche* (éd. U. Trivellato). Padova: Cleup, 61-86.

TRIVELLATO, U., et RETTORE, E. (1986). Preliminary data errors and their impact on the forecast error of simultaneous-equation models. *Journal of Business & Economic Statistics*, 4, 445-453.

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

WILTON, D.A., et SMITH, P.M. (1974). The efficiency of the G.N.P. revision process: An historical analysis. Division des analyses de conjoncture, Statistique Canada, Ottawa (document interne).

ZARNOWITZ, V. (1982). On functions, quality, and timeliness of economic information. *Journal of Business*, 55, 87-119.

# DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

## 1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

## 2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

## 3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, $\omega$; o, O, 0; 1, l).

3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

## 4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

## 5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.